Financial asset pricing theory

Claus Munk

this version: September 26, 2007

Contents

Preface

INCOMPLETE!

Preliminary and incomplete lecture notes intended for use at an advanced master's level or an introductory Ph.D. level. I appreciate comments and corrections from Simon Lysbjerg Hansen and students exposed to earlier versions of these notes. Additional comments and suggestions are very welcome!

Alternative textbooks covering much of the same material include Ingersoll (1987), Huang and Litzenberger (1988), Dothan (1990), Cochrane (2001), and Duffie (2001).

Claus Munk
Department of Business and Economics
University of Southern Denmark
Campusvej 55, DK-5230 Odense M, Denmark
E-mail: cmu@sam.sdu.dk
Internet homepage: http://www.sam.sdu.dk/∼cmu

Chapter 1

Introduction and overview

1.1 What is modern asset pricing?

Asset pricing models are models for the pricing of financial assets. It is interesting in itself to be able to model and understand the pricing mechanisms in the seemingly complex financial markets, but it is also important for a number of financial problems faced by individuals and corporations such as

- asset allocation: how individual and institutional investors combine various financial assets into portfolios;
- the measurement and management of financial risks, e.g. in banks and other financial institutions;
- capital budgeting decision in firms;
- capital structure decisions in firms;
- the identification and possible resolution of potential conflicts of interest between the stakeholders of a firm, e.g. shareholders vs. creditors, shareholders vs. managers.

To the extent that central banks and governments want to control or at least influence financial markets, e.g. setting interest rates or limiting stock market volatility, they also need a deep understanding of the asset pricing mechanisms and the link between financial markets and macroeconomics. Finally, there is a trend in accounting regulation towards more market-oriented valuation of assets and liabilities instead of the traditional valuation at historical costs.

Undoubtedly, the Capital Asset Pricing Model (CAPM) developed by Sharpe (1964), Lintner (1965), and Mossin (1966) is the best known asset pricing model. The key message of the model is that the expected excess return on a risky financial asset is given by the product of the market-beta of the asset and the expected excess return on the market portfolio. Here the "excess return" of an asset or a portfolio is the return less the risk-free return and the "market-beta" of an asset is the covariance between the return on this asset and the return on the market portfolio, divided by the variance of the return on the market portfolio. Only the risk correlated with the market will give a risk premium in terms of a higher expected return (assuming the market-beta is positive).

The remaining risk can be diversified away and is therefore not priced in equilibrium. In principle, the market portfolio includes all assets, not only traded financial assets but also non-traded assets like the human capital (value of labor income) of all individuals. However, the market portfolio is typically approximated by a broad stock index, although this approximation is not necessarily very precise.

The CAPM has been very successful as a pedagogical tool for presenting and quantifying the tradeoff between risk and (expected) return, and it has also been widely used in practical applications. It captures some important characteristics of the pricing in financial markets in a rather simple way. However, the CAPM is insufficient in many aspects and it is built on a number of unrealistic assumptions. Here is a partial list of problems with the CAPM:

1. The original CAPM is formulated and derived in a one-period world where assets and investors are only modeled over one common period. In applications, it is implicitly assumed that the CAPM repeats itself period by period which intuitively demands some sort of independence between the pricing mechanisms in different periods, which again requires the unrealistic assumption that the demand and supply of agents living for several periods are the same in all periods.

2. The CAPM is not designed for capturing variations in asset prices over time and cannot do so.

3. Typical derivations of the CAPM assume that all asset returns over the fixed period are normally distributed. For assets with limited liability you cannot loose more than you have invested so the rate of return cannot be lower than -100%, which is inconsistent with the normal distribution that associates a positive probability to any return between $-\infty$ and $+\infty$. Empirical studies show that for many assets the normal distribution is not even a good approximation of the return distribution.

4. The true market portfolio contains many unobservable assets so how should you find the expected return and variance on the market portfolio and its covariances with all individual assets?

5. The CAPM is really quite unsuccessful in explaining empirical asset returns. Differences in market-betas cannot explain observed differences in average returns of stocks.

6. The CAPM is not a full asset pricing model in the sense that it does not say anything about what the return on the risk-free asset or the expected return on the market portfolio should be. And it does not offer any insight into the links between financial markets and macroeconomic variables like consumption, production, and inflation.

The purpose of this book is to develop a deeper understanding of asset pricing than the CAPM can offer.

When an investor purchases a given asset, she obtains the right to receive the future payments of the asset. For many assets the size of these future payments is uncertain at the time of purchase since it may depend on the overall state of the economy and/or the state of the issuer of the asset at the payment dates. Risk-averse investors will value a payment of a given size more highly if they receive it in a "bad" state than in a "good" state. This is captured by the term "state price"

introduced by Arrow (1953). A state price for a given state at a given future point in time indicates how much investors are willing to sacrifice today in return for an extra payment of one unit in that future state. Presumably investors will value a given payment in a given state the same no matter which asset the payment comes from. Therefore state prices are valid for all assets. The value of any specific asset is determined by the general state prices in the market and the state-contingent future payments of the asset. Modern asset pricing theory is based on models of the possible states and the associated state prices.

The well-being of individuals will depend on their consumption of goods throughout their lives. By trading financial assets they can move consumption opportunities from one point in time to another and from one state of the world to another. The preferences for consumption of individuals determine their demand for various assets and thereby the equilibrium prices of these assets. Hence, the state price for any given state must be closely related to the individuals' (marginal) utility of consumption in that state. Many modern asset pricing theories and models are based on this link between asset prices and consumption.

1.2 Elements of asset pricing models

1.2.1 Assets

For potential investors the important characteristics of a financial asset or any other investment opportunity is its current price and its future payments which the investor will be entitled to if she buys the asset. Stocks deliver dividends to owners. The dividends will surely depend on the well-being of the company. Bonds deliver coupon payments and repayments of the outstanding debt, usually according to some predetermined schedule. For bonds issued by most governments, you might consider these payments to be certain, i.e. risk-free. On the other hand, if the government bond promises certain dollar payments, you will not know how many consumption goods you will be able to buy for these dollar payments, that is the payments are risky in real terms. The payments of bonds issued by corporations are also uncertain. The future payments of derivatives such as forwards, futures, options, and swaps depend on the evolution of some underlying random variable and therefore are also uncertain.

Let us simply refer to the payments of any asset as dividends. More precisely, a "dividend" means the payment of a given asset at a given point in time. The uncertain dividend of an asset at a given point in time is naturally modeled by a random variable. If an asset provides the owner with payments at several points in time, we need a collection of random variables (one for each payment date) to represent all the dividends. Such a collection of random variables is called a stochastic process. A stochastic process is therefore the natural way to represent the uncertain flow of dividends of an asset over time. We will refer to the stochastic process representing the dividends of an asset as the dividend process of the asset.

1.2.2 Investors

In reality, only a small part of the trading in financial markets is executed directly by individuals while the majority of trades are executed by corporations and financial institutions such as pension funds, insurance companies, banks, broker firms, etc. However, these institutional investors trade

on behalf of individuals, either customers or shareholders. Productive firms issue stocks and corporate bonds to finance investments in production technology they hope will generate high earnings and, consequently, high returns to their owners in future years. In the end, the decisions taken at the company level are also driven by the desires of individuals to shift consumption opportunities across time and states. In our basic models we will assume that all investors are individuals and ignore the many good reasons for the existence of various intermediaries. For example, we will assume that assets are traded without transaction costs. We will also ignore taxes and the role of the government and central banks in the financial markets. Some authors use the term "agent" or "investor" instead of "individual," maybe in order to indicate that some investment decisions are taken by other decision-making units than individual human beings.

How should we represent an individual in an asset pricing model? We will assume that individuals basically care about their consumption of goods and services throughout their life. The consumption of a given individual at a future point in time is typically uncertain and we will therefore represent it by a random variable. The consumption of an individual at all future dates is represented by a stochastic process, the consumption process of the individual. Although real-life economies offer a large variety of consumption goods and services, in our basic models we will assume that there is only one good available for consumption and that each individual only cares about her own consumption and not the consumption of other individuals. The single consumption good is assumed to be perishable, i.e. it cannot be stored or resold but has to be consumed immediately. In more advanced models discussed in later chapters we will in fact relax these assumptions and allow for multiple consumption goods, e.g. we will introduce a durable good (like a house), and we will also discuss models in which the well-being of an individual also depends on what other individuals consume, which is often referred to as the "keeping up with the Jones'es" property. Both extensions turn out to be useful in bringing our theoretical models closer to real-life financial data but it is preferable to understand the simpler models first. Of course, the well-being of an individual will also be affected by the number of hours she works, the physical and mental challenges offered by her position, etc., but such issues will also be ignored in basic models.

We will assume that each individual is endowed with some current wealth and some future income stream from labor, gifts, inheritance, etc. For most individuals the future income will be uncertain. The income of an individual at a given future point in time is thus represented by a random variable and the income at all future dates is represented by a stochastic process, the income process. We will assume that the income process is exogenously given and hence ignore labor supply decisions.

If the individual cannot make investments at all (not even save current wealth), it will be impossible for her to currently consume more than her current wealth and impossible to consume more at a future point in time than her income at that date. Financial markets allow the individual to shift consumption opportunities from one point in time to another, e.g. from working life to retirement, and from one state of the world to another, e.g. from a state in which income is extremely high to a state where income is extremely low (much as insurance contracts do). The prices of financial assets give the prices of shifting consumption through time and states of the world. The individuals' desire to shift consumption through time and states will determine the demand and supply and hence the equilibrium prices of the financial assets. To study asset pricing we therefore have to model how individuals choose between different, uncertain consumption processes. The

preferences for consumption of an individual is typically modeled by a utility function. Since this is a text on asset pricing, we are not primarily interested in deriving the optimal consumption stream and the associated optimal strategy for trading financial assets. However, since asset prices are set by the decisions of individuals, we will have to discuss some aspects of optimal consumption and trading.

1.2.3 Equilibrium

For any given asset, i.e. any given dividend process, our aim is to characterize the "reasonable" price or the set of "reasonable" prices. A price is considered reasonable if the price is an equilibrium price. An equilibrium is characterized by two conditions: (1) supply equals demand for any asset, i.e. markets clear, (2) any investor is satisfied with her current position in the assets given her personal situation and the asset prices. Associated with any equilibrium is a set of prices for all assets and, for each investor, a trading strategy and the implied consumption strategy.

1.2.4 The time span of the model

As discussed above, the important ingredients of all basic asset pricing models are the dividends of the assets available for trade and the utility functions, current wealth, and future incomes of the individuals that can trade the assets. We will discuss asset pricing in three types of models:

1. **one-period model**: all action takes place at two points in time, the beginning of the period (time 0) and the end of the period (time 1). Assets pay dividends only at the end of the period and are traded only at the beginning of the period. The aim of the model is to characterize the prices of the assets at the beginning of the period. Individuals have some initial beginning-of-period wealth and (maybe) some end-of-period income. They can consume at both points in time.

2. **discrete-time model**: all action takes place at a finite number of points in time. Let us denote the set of these time points by $\mathcal{T} = \{0, 1, 2, \ldots, T\}$. Individuals can trade at any of these time points, except at T, and consume at any time $t \in \mathcal{T}$. Assets can pay dividends at any time in $\mathcal{T}$, except time 0. Assuming that the price of a given asset at a given point in time is ex-dividend, i.e. the value of future dividends excluding any dividend at that point in time, prices are generally non-trivial at all but the last point in time. We aim at characterizing these prices.

3. **continuous-time model**: individuals can consume at any point in time in an interval $\mathcal{T} = [0, T]$. Assets pay dividends in the interval $(0, T]$ and can be traded in $[0, T)$. Ex-dividend asset prices are non-trivial in $[0, T)$. Again, our aim is to characterize these prices.

In a one-period setting there is uncertainty about the state of the world at the end of the period. The dividends of financial assets and the incomes of the individuals at the end of the period will generally be unknown at the beginning of the period and thus modeled as random variables. Any quantity that depends on either the dividends or income will also be random variables. For example, this will be the case for the end-of-period value of portfolios and the end-of-period consumption of individuals.

Both the discrete-time model and the continuous-time model are multi-period models and can potentially capture the dynamics of asset prices. In both cases, T denotes some terminal date in the sense that we will not model what happens after time T. We assume that $T < \infty$ but under some technical conditions the analysis extends to $T = \infty$.

Financial markets are by nature dynamic and should therefore be studied in a multi-period setting. One-period models should serve only as a pedagogical first step in the derivation of the more appropriate multi-period models. Indeed, many of the important conclusions derived in one-period models carry over to multi-period models. Other conclusions do not. And some issues cannot be meaningfully studied in a one-period framework.

It is not easy to decide on whether to use a discrete-time or a continuous-time framework for studying multi-period asset pricing. Both model types have their virtues and drawbacks. Both model types are applied in theoretical research and real-life applications. We will therefore consider both modeling frameworks. The basic asset pricing results in the early chapters will be derived in both settings. Some more specific asset pricing models discussed in later chapters will only be presented in one of these frameworks. Some authors prefer to use a discrete-time model, others prefer a continuous-time model. It is comforting that, for most purposes, both models will result in identical or very similar conclusions.

At first, you might think that the discrete-time framework is more realistic. However, in real-life economies individuals can in fact consume and adjust portfolios at virtually any point in time. Individuals are certainly not restricted to consume and trade at a finite set of pre-specified points in time. Of course no individual will trade financial assets continuously due to the existence of explicit and implicit costs of such transactions. But even if we take such costs into account, the frequency and exact timing of actions can be chosen by each individual. If we are really concerned about transaction costs, it would be better to include those in a continuous-time modeling framework.

Many people will find discrete-time models easier to understand than continuous-time models and if you want to compare theoretical results with actual data it will usually be an advantage if the model is formulated with a period length closely linked to the data frequency. On the other hand, once you have learned how to deal with continuous-time stochastic processes, many results are clearer and more elegantly derived in continuous-time models than in discrete-time models. The analytical virtues of continuous-time models are basically due to the well-developed theory of stochastic calculus for continuous-time stochastic processes, but also due to the fact that integrals are easier to deal with than discrete sums, differential equations are easier to deal with than difference equations, etc.

1.3 The organization of this book

The remainder of this book is organized as follows. Chapter 2 discusses how to represent uncertainty and information flow in asset pricing models. It also introduces stochastic processes and some key results on how to deal with stochastic processes, which we will use in later chapters.

Chapter 3 shows how we can model financial assets and their dividends as well as how we can represent portfolios and trading strategies. It also defines the important concepts of arbitrage, redundant assets, and market completeness.

Chapter 4 defines the key concept of a state-price deflator both in one-period models, in discrete-

time multi-period models, and in continuous models. A state-price deflator is one way to represent the general pricing mechanism of a financial market. We can price any asset given the state-price deflator and the dividend process of that asset. Conditions for the existence and uniqueness of a state-price deflator are derived as well as a number of useful properties of state-price deflators. We will also briefly discuss alternative ways of representing the general market pricing mechanism, e.g. through a set of risk-neutral probabilities.

The state-price deflator and therefore asset prices are ultimately determined by the supply and demand of investors. Chapter 5 studies how we can represent the preferences for investors. We discuss when preferences can be represented by expected utility, how we can measure the risk aversion of an individual, and introduce some frequently used utility functions. In Chapter 6 we investigate how individual investors will make decisions on consumption and investments. We set up the utility maximization problem for the individual and characterize the solution for different relevant specifications of preferences. The solution gives an important link between state-price deflators (and, thus, the prices of financial assets) and the optimal decisions at the individual level.

Chapter 7 deals with the market equilibrium. We will discuss when market equilibria are Pareto-efficient and when we can think of the economy as having only one, representative individual instead of many individuals.

Chapter 8 further explores the link between individual consumption choice and asset prices. The very general Consumption-based Capital Asset Pricing Model (CCAPM) is derived. A simple version of the CCAPM is confronted with data and a number of extensions are discussed.

Chapter 9 studies the so-called factor models of asset pricing where one or multiple factors govern the state-price deflators and thus asset prices and returns. Some empirically successful factor models are described. It is also shown how pricing factors can be identified theoretically as a special case of the general CCAPM.

While Chapters 8 and 9 mostly focus on explaining the expected excess return of risky assets, most prominently stocks, Chapter 10 explores the implications of general asset pricing theory for the short-term interest rate and the whole term structure of interest rates. It also critically reviews some traditional views on the term structure of interest rates.

Chapter 11 shows how the information in a state-price deflator equivalently can be represented by the price of one specific asset and an appropriately risk-adjusted probability measure. This turns out to be very useful when dealing with derivative securities, which is the topic of Chapter 12.

Each chapter ends with a number of exercises, which either illustrate the concepts and conclusions of the chapter or provide additional related results.

1.4 Prerequisites

We will study asset pricing with the well-established scientific approach: make precise definitions of concepts, clear statements of assumptions, and formal derivations of results. This requires extensive use of mathematics, but not very complicated mathematics. Concepts, assumptions, and results will all be accompanied by financial interpretations. Examples will be used for illustrations. The main mathematical disciplines we will apply are linear algebra, optimization, and probability theory. Linear algebra and optimization are covered by many good textbooks on mathematics for economics as, e.g., the companion books by Sydsaeter and Hammond (2005) and Sydsaeter,

Hammond, Seierstad, and Strom (2005), and—of course—by many more general textbooks on mathematics. Probability theory is usually treated in separate textbooks... A useful reference "manual" is Sydsaeter, Strom, and Berck (2000).

Probability theory Appendix A gives a review of main concepts and definitions in probability theory. Appendix B summarizes some important results on the lognormal distribution which we shall frequently apply.

Linear algebra, vectors, and matrices We will frequently use vectors and matrices to represent a lot of information in a compact manner. For example, we will typically use a vector to represent the prices of a number of assets and use a matrix to represent the dividends of different assets in different states of the world. Therefore some basic knowledge of how to handle vectors and matrices (so-called linear algebra) are needed.

We will use boldface symbols like $\boldsymbol{x}$ to denote vectors and vectors are generally assumed to be column vectors. Matrices will be indicate by double underlining like $\underline{\underline{A}}$. We will use the symbol $^\top$ to denote the transpose of a vector or a matrix. The following is an incomplete and relatively unstructured list of basic properties of vectors and matrices.

Given two vectors $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ of the same dimension, the dot product of $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as $\boldsymbol{x} \cdot \boldsymbol{y} = x_1 y_1 + \cdots + x_n y_n = \sum_{i=1}^n x_i y_i$.

The identity matrix of dimension n is the $n \times n$ matrix $\underline{\underline{I}} = [I_{ij}]$ with 1 along the diagonal and zeros elsewhere, i.e. $I_{ii} = 1$, $i = 1, \ldots, n$, and $I_{ij} = 0$ for $i, j = 1, \ldots, n$ with $i \neq j$.

The transpose of an $m \times n$ matrix $\underline{\underline{A}} = [A_{ij}]$ is $n \times m$ matrix $\underline{\underline{A}}^\top = [A_{ji}]$ with columns and rows interchanged.

Given an $m \times n$ matrix $\underline{\underline{A}} = [A_{ij}]$ and an $n \times p$ matrix $\underline{\underline{B}} = [B_{kl}]$, the product $\underline{\underline{AB}}$ is the $m \times p$ matrix with (i, j)'th entry given by $(\underline{\underline{AB}})_{i,j} = \sum_{k=1}^n A_{ik} B_{kj}$. In particular with $m = p = 1$, we see that for two vectors $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, we have $\boldsymbol{x}^\top \boldsymbol{y} = \boldsymbol{x} \cdot \boldsymbol{y}$, so the matrix product generalizes the dot product for vectors.

$\left(\underline{\underline{AB}}\right)^\top = \underline{\underline{B}}^\top \underline{\underline{A}}^\top$.

An $n \times n$ matrix $\underline{\underline{A}}$ is said to be non-singular if there exists a matrix $\underline{\underline{A}}^{-1}$ so that $\underline{\underline{AA}}^{-1} = \underline{\underline{I}}$, where $\underline{\underline{I}}$ is the $n \times n$ identity matrix, and then $\underline{\underline{A}}^{-1}$ is called the inverse of $\underline{\underline{A}}$.

A 2×2 matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is non-singular if $ad - bc \neq 0$ and the inverse is then given by

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

If $\underline{\underline{A}}$ is non-singular, then $\underline{\underline{A}}^\top$ is non-singular and $(\underline{\underline{A}}^\top)^{-1} = (\underline{\underline{A}}^{-1})^\top$.

If $\underline{\underline{A}}$ and $\underline{\underline{B}}$ are non-singular matrices of appropriate dimensions,

$$\left(\underline{\underline{AB}}\right)^{-1} = \underline{\underline{B}}^{-1} \underline{\underline{A}}^{-1}. \tag{1.1}$$

If $\boldsymbol{a}$ and $\boldsymbol{x}$ are vectors of the same dimension, then $\frac{\partial \boldsymbol{a}^\top \boldsymbol{x}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{x}^\top \boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}$.

If $\boldsymbol{x}$ is a vector of dimension n and $\underline{\underline{A}}$ is an $n \times n$ matrix, then $\frac{\partial \boldsymbol{x}^\top \underline{\underline{A}} \boldsymbol{x}}{\partial \boldsymbol{x}} = \left(\underline{\underline{A}} + \underline{\underline{A}}^\top\right) \boldsymbol{x}$. In particular, if $\underline{\underline{A}}$ is symmetric, i.e. $\underline{\underline{A}}^\top = \underline{\underline{A}}$, we have $\frac{\partial \boldsymbol{x}^\top \underline{\underline{A}} \boldsymbol{x}}{\partial \boldsymbol{x}} = 2 \underline{\underline{A}} \boldsymbol{x}$.

Optimization Optimization problems arise naturally in all parts of economics, also in finance. To study asset pricing theory, we will have to study how individual investors make decisions about consumption and investment. Assuming that the well-being of an individual can be represented by some sort of utility function, we will have to maximize utility subject to various constraints, e.g. a budget constraint. Therefore we have to apply results on constrained optimization. The main approach for solving constrained optimization problems is the Lagrange approach; see, e.g., Sydsaeter and Hammond (2005).

Chapter 2

Uncertainty, information, and stochastic processes

2.1 Introduction

2.2 Probability space

Any model with uncertainty refers to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where

- Ω is the state space of possible outcomes. An element $\omega \in \Omega$ represents a possible realization of all uncertain objects of the model;
- $\mathcal{F}$ is a σ-algebra in Ω, i.e. a collection of subsets of Ω with the properties

  (i) $\Omega \in \mathcal{F}$,

  (ii) for any set F in $\mathcal{F}$, the complement $F^c \equiv \Omega \setminus F$ is also in $\mathcal{F}$,

  (iii) if $F_1, F_2, \cdots \in \mathcal{F}$, then the union $\cup_{n=1}^{\infty} F_n$ is in $\mathcal{F}$.

  $\mathcal{F}$ is the collection of all events (subsets of Ω) that can be assigned a probability;
- $\mathbb{P}$ is a probability measure, i.e. a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ with $\mathbb{P}(\Omega) = 1$ and the property that $\mathbb{P}(\cup_{m=1}^{\infty} A_m) = \sum_{m=1}^{\infty} \mathbb{P}(A_m)$ for any sequence $A_1, A_2, \ldots$ of disjoint events.

An uncertain object can be formally modeled as a random variable on the probability space. A random variable X on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a real-valued function on Ω which is $\mathcal{F}$-measurable in the sense that for any interval $I \subseteq \mathbb{R}$, the set $\{\omega \in \Omega \mid X(\omega) \in I\}$ belongs to $\mathcal{F}$, i.e. we can assign a probability to the event that the random variable takes on a value in I.

What is the relevant state space for an asset pricing model? A state $\omega \in \Omega$ represents a possible realization of all relevant uncertain objects over the entire time span of the model. In one-period models dividends, incomes, etc. are realized at time 1. A state defines realized values of all the dividends and incomes at time 1. In multi-period models a state defines dividends, incomes, etc. at all points in time considered in the model, i.e. all $t \in \mathcal{T}$, where either $\mathcal{T} = \{0, 1, 2, \ldots, T\}$ or $\mathcal{T} = [0, T]$. The state space must include all the possible combinations of realizations of the

uncertain objects that may affect the pricing of the assets. These uncertain objects include all the possible combinations of realizations of (a) all the future dividends of all assets, (b) all the future incomes of all individuals, and (c) any other initially unknown variables that may affect prices, e.g. variables that contain information about the future development in dividends or income. The state space Ω therefore has to be "large." If you want to allow for continuous random variables, for example dividends that are normally distributed, you will need an infinite state space. If you restrict all dividends, incomes, etc. to be discrete random variables, i.e. variables with a finite number of possible realizations, you can do with a finite state space. For some purposes we need to distinguish between an infinite state space and a finite state space.

We will sometimes assume a finite state space in which case we will take it to be $\Omega = \{1, 2, \ldots, S\}$ so that there are S possible states of which exactly one will be realized. An event is then simply a subset of Ω and $\mathcal{F}$ is the collection of all subsets of Ω. The probability measure $\mathbb{P}$ is defined by the state probabilities $p_\omega \equiv \mathbb{P}(\omega)$, $\omega = 1, 2, \ldots, S$, which we take to be strictly positive with $p_1 + \ldots p_S = 1$, of course. With a finite state space we can represent random variables with S-dimensional vectors and apply results and techniques from linear algebra. In any case we take the state probabilities as given and assume they are known to all individuals.

2.3 Information

In a one-period model all uncertainty is resolved at time $t = 1$. At time 0 we only know that the true state is an element in Ω. At time 1 we know exactly which state has been realized. In a multi-period model the uncertainty is gradually resolved. Investors will gradually know more and more about the true state. For example, the dividends of assets at a given point in time are typically unknown before that time, but known afterwards. The consumption and investment decisions taken by individuals at a given point in time will depend on the available information at that time and therefore asset prices will also depend on the information known. We will therefore have to consider how to formally represent the flow of information through time.

To illustrate how we can represent the information at different points in time, consider an example of a two-period, three-date economy with six possible outcomes simply labeled 1 through 6. In Figure 2.1 each outcome is represented by a dashed horizontal line. The probability of each outcome is written next to each line. At time 0 we assume that investors are unable to rule out any of the six outcomes—if a state could be ruled out from the start, it should not have been included in the model. This is indicated by the ellipse around the six dots/lines representing the possible outcomes. At time 1, investors have learned either (i) that the true outcome is 1 or 2, (ii) that the true outcome is 3,4, or 5, or (iii) that the true outcome is 6. At time 2, all uncertainty has been resolved so that investors will know exactly which outcome is realized.

We can represent the information available at any given point in time t by a **partition** $\mathbf{F}_t$ of Ω, which means that $\mathbf{F}_t$ is a collection of subsets $F_{t1}, F_{t2}, \ldots$ of Ω so that

(i) the union of these subsets equal the entire set Ω: $\cup_k F_{tk} = \Omega$.

(ii) the subsets are disjoint: $F_{tk} \cap F_{tl} = \emptyset$ for all $k \neq l$.

In our example, the partition $\mathbf{F}_0$ representing time 0 information (or rather lack of information)

is the trivial partition consisting only of $F_0 = \Omega$, i.e.

$$\mathbf{F}_0 = \{\Omega\}.$$

The partition $\mathbf{F}_1$ representing time 1 information consists of F_{11}, F_{12}, and F_{13}, i.e.

$$\mathbf{F}_1 = \big\{\{1,2\},\{3,4,5\},\{6\}\big\}.$$

The partition $\mathbf{F}_2$ that represents time 2 information (full information) is

$$\mathbf{F}_2 = \big\{\{1\},\{2\},\{3\},\{4\},\{5\},\{6\}\big\}.$$

In a general multi-period model with a finite state space Ω, the information flow can be summarized by a sequence $(\mathbf{F}_t)_{t\in\mathcal{T}}$ of partitions. Since investors learn more and more, the partitions should be increasingly fine, which more formally means that when $t < t'$, every set $F \in \mathbf{F}_{t'}$ is a subset of some set in $\mathbf{F}_t$.

An alternative way of representing the information flow is in terms of an **information filtration**, i.e. a sequence $(\mathcal{F}_t)_{t\in\mathcal{T}}$ of sigma-algebras on Ω. Given a partition $\mathbf{F}_t$ of Ω, we can construct a sigma-algebra $\mathcal{F}_t$ as the set of all unions of (countably many) sets in $\mathbf{F}_t$, including the "empty union", i.e. the empty set $\emptyset$. Where $\mathbf{F}_t$ contains only the disjoint "decidable" events at time t, $\mathcal{F}_t$ contains all "decidable" events at time t. For our simple two-period example above we get

$$\mathcal{F}_0 = \{\emptyset,\Omega\},$$
$$\mathcal{F}_1 = \{\emptyset,\{1,2\},\{3,4,5\},\{6\},\{1,2,3,4,5\},\{1,2,6\},\{3,4,5,6\},\Omega\},$$

while $\mathcal{F}_2$ becomes the collection of *all* possible subsets of Ω. In a general multi-period model we write $(\mathcal{F}_t)_{t\in\mathcal{T}}$ for the information filtration. We will always assume that the time 0 information is trivial, corresponding to $\mathcal{F}_0 = \{\emptyset,\Omega\}$ and that all uncertainty is resolved at or before the final date so that $\mathcal{F}_T$ is the set of all possible subsets of Ω. The fact that we learn more and more about the true state as time goes by implies that we must have $\mathcal{F}_t \subset \mathcal{F}_{t'}$ whenever $t < t'$, i.e. every set in $\mathcal{F}_t$ is also in $\mathcal{F}_{t'}$.

Above we constructed an information filtration from a sequence of partitions. We can also go from a filtration to a sequence of partitions. In each $\mathcal{F}_t$, simply remove all sets that are unions of other sets in $\mathcal{F}_t$. Therefore there is a one-to-one relationship between information filtration and a sequence of partitions.

In models with an infinite state space, the information filtration representation is preferable. In any case we will therefore generally write the formal model of uncertainty and information as a **filtered probability space** $(\Omega,\mathcal{F},\mathbb{P},(\mathcal{F}_t)_{t\in\mathcal{T}})$, where $(\Omega,\mathcal{F},\mathbb{P})$ is a probability space and $(\mathcal{F}_t)_{t\in\mathcal{T}}$ is an information filtration.

Whenever the state space is finite we can alternatively represent the uncertainty and information flow by a multinomial tree. For example we can depict the uncertainty and information flow in Figure 2.1 by the multinomial tree in Figure 2.2. Each node at a given point in time corresponds to an element in the partition representing the information. For example the node labeled F_{11} at time 1 represents the element $\{1,2\}$ of the partition $\mathbf{F}_1$. We can think of F_{11}, F_{12}, and F_{13} as the three possible "scenarios" at time 1. At time 0, there is only one possible scenario. At time 2, where all uncertainty is resolved, there are as many scenarios as states. The arrival of

new information in a given period can be thought of as the transition from one scenario at the beginning of the period to a scenario at the end of the period. In our example, there are three possible transitions in the first period. If the economy is in scenario F_{11} at time 1 (i.e. the true state is 1 or 2), there are two possible transitions over the next period, either to F_{21} (state 1) or to F_{22} (state 2). If the economy is in scenario F_{12} at time 1 (the true state is known to be 3, 4, or 5), there are three possible transitions over the next period, to F_{23} (state 3), to F_{24} (state 4), or to F_{25} (state 5). If the economy is in scenario F_{13} at time 1, the true state is already known to be state 6, and there is only one possible transition over the next period, corresponding to no new information arriving. Each state corresponds to a path through the tree.

The transitions are illustrated by the arrows in Figure 2.2. The numbers along the lines are conditional probabilities of the transitions happening. Over the first period there is really no information to condition on. The transition from F_0 to F_{11} will happen with a probability of 0.3, which is simply the sum of the probabilities of the two outcomes in F_{11}, namely the probability of 0.24 for state 1 and the probability of 0.06 for state 2. Similarly for the other transitions over the first period. The probabilities assigned to the transitions over the second period are true conditional probabilities. Conditional on the economy being in scenario F_{11} at time 1, it will move to F_{21} with a probability of $0.24/(0.24 + 0.06) = 0.8$ since that is the probability of state 1 given that the state is either 1 or 2 as represented by scenario F_{11}. Similarly for the other transitions over the second period. Of course, given the conditional probabilities in the multinomial tree, we can also recover the state probabilities. For example, the state $\omega = 5$ corresponds to a transition from F_0 to F_{12} over the first period, followed by a transition from F_{12} to F_{25} over the second period. The probability of this sequence of transitions is given by the product of the probabilities of each of the transitions, i.e. $0.4 \cdot 0.5 = 0.2$, which equals the probability of state $\omega = 5$.

In our asset pricing models we will often deal with expectations of random variables, e.g. the expectation of the dividend of an asset at a future point in time. In the computation of such an expectation we should take the information currently available into account. Hence we need to consider conditional expectations. Recall that the information at a given point in time t is represented by a σ-algebra $\mathcal{F}_t$ (or, equivalently, a partition $\mathbf{F}_t$). One can generally write the expectation of a random variable X given the σ-algebra $\mathcal{F}_t$ as $\mathrm{E}[X|\mathcal{F}_t]$. For our purposes the σ-algebra $\mathcal{F}_t$ will always represent the information at time t and we will write $\mathrm{E}_t[X]$ instead of $\mathrm{E}[X|\mathcal{F}_t]$. Since we assume that the information at time 0 is trivial, conditioning on time 0 information is the same as not conditioning on any information, hence $\mathrm{E}_0[X] = \mathrm{E}[X]$. Since we assume that all uncertainty is resolved at time T, we have $\mathrm{E}_T[X] = X$. We will sometimes use the following result:

Theorem 2.1 (The Law of Iterated Expectations) *If $\mathcal{F}$ and $\mathcal{G}$ are two σ-algebras with $\mathcal{F} \subseteq \mathcal{G}$ and X is a random variable, then $\mathrm{E}\left[\mathrm{E}[X|\mathcal{G}] \mid \mathcal{F}\right] = \mathrm{E}[X|\mathcal{F}]$. In particular, if $(\mathcal{F}_t)_{t\in\mathcal{T}}$ is an information filtration and $t' > t$, we have*

$$\mathrm{E}_t\left[\mathrm{E}_{t'}[X]\right] = \mathrm{E}_t[X].$$

Loosely speaking, the theorem says that what you expect today of some variable that will be realized in two days is equal to what you expect today that you will expect tomorrow about the same variable.

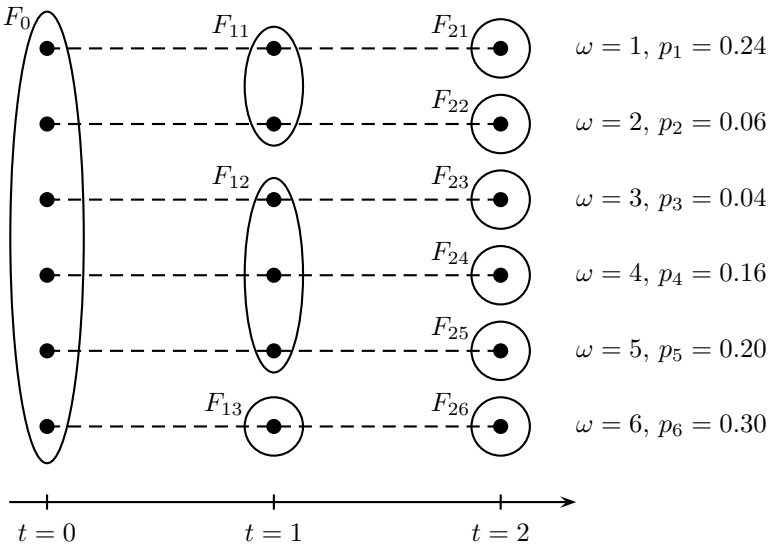


Figure 2.1: An example of a two-period economy.

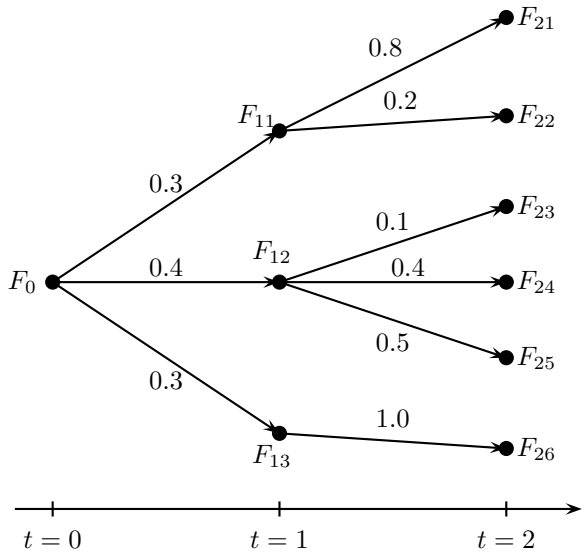


Figure 2.2: The multinomial tree version of the two-period economy in Figure 2.1.

We can define conditional variances, covariances, and correlations from the conditional expectation exactly as one defines (unconditional) variances, covariances, and correlations from (unconditional) expectations:

$$\mathrm{Var}_t[X] = \mathrm{E}_t\left[(X - \mathrm{E}_t[X])^2\right] = \mathrm{E}_t[X^2] - (\mathrm{E}_t[X])^2,$$
$$\mathrm{Cov}_t[X,Y] = \mathrm{E}_t\left[(X - \mathrm{E}_t[X])(Y - \mathrm{E}_t[Y])\right] = \mathrm{E}_t[XY] - \mathrm{E}_t[X]\,\mathrm{E}_t[Y],$$
$$\mathrm{Corr}_t[X,Y] = \frac{\mathrm{Cov}_t[X,Y]}{\sqrt{\mathrm{Var}_t[X]\,\mathrm{Var}_t[Y]}}.$$

Again the conditioning on time t information is indicated by a t subscript.

In the two-period model of Figures 2.1 and 2.2 suppose we have an asset paying state-dependent dividends at time 1 and 2 as depicted in Figures 2.3 and 2.4. Then the expected time 2 dividend computed at time 0 is

$$\mathrm{E}[D_2] = \mathrm{E}_0[D_2] = 0.24 \cdot 0 + 0.06 \cdot 20 + 0.04 \cdot 10 + 0.16 \cdot 5 + 0.2 \cdot 20 + 0.3 \cdot 20 = 12.4.$$

What is the expected time 2 dividend computed at time 1? It will depend on the information available at time 1. If the information corresponds to the event $F_{11} = \{1, 2\}$, the expected dividend is

$$\mathrm{E}_1[D_2] = 0.8 \cdot 0 + 0.2 \cdot 20 = 4.$$

If the information corresponds to the event $F_{12} = \{3, 4, 5\}$, the expected dividend is

$$\mathrm{E}_1[D_2] = 0.1 \cdot 10 + 0.4 \cdot 5 + 0.5 \cdot 20 = 13.$$

If the information corresponds to the event $F_{13} = \{6\}$, the expected dividend is

$$\mathrm{E}_1[D_2] = 1.0 \cdot 20 = 20.$$

The time 1 expectation of the time 2 dividend is therefore a random variable which is measurable with respect to the information at time 1. Note that the time 0 expectation of that random variable is

$$\mathrm{E}\left[\mathrm{E}_1[D_2]\right] = 0.3 \cdot 4 + 0.4 \cdot 13 + 0.3 \cdot 20 = 12.4 = \mathrm{E}[D_2]$$

consistent with the Law of Iterated Expectations.

2.4 Stochastic processes: definition, notation, and terminology

In one-period models all uncertain objects can be represented by a random variable. For example the dividend (at time 1) of a given asset is a random variable. In multi-period models we have to keep track of dividends, asset prices, consumption, portfolios, (labor) income, etc., throughout the time set $\mathcal{T}$. For example the dividend of a given asset, say asset i, at a particular future date $t \in \mathcal{T}$ can be represented by a random variable D_{it}. Recall that, formally, a random variable is a function from the state space Ω into $\mathbb{R}$, the set of real numbers. To represent the dividends of an asset throughout all dates, we need a collection of random variables, one for each date. Such a collection is called a **stochastic process**. (We will often just write "process" instead of "stochastic

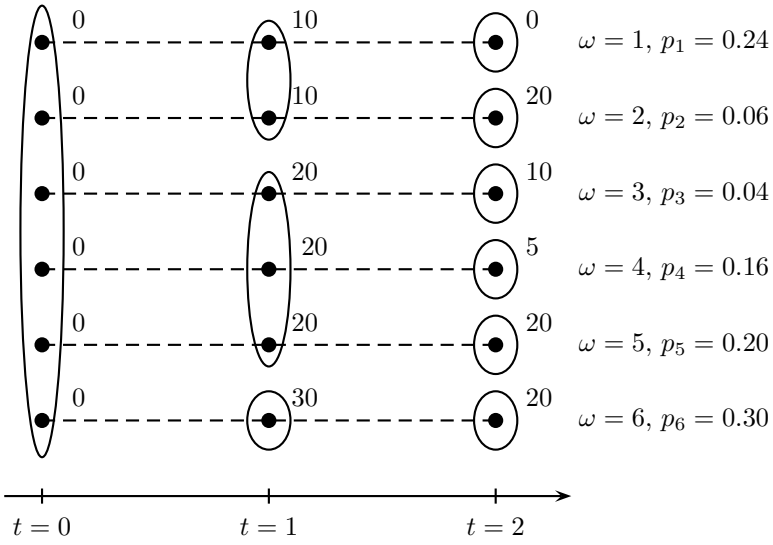


Figure 2.3: The dividends of an asset in the two-period economy.

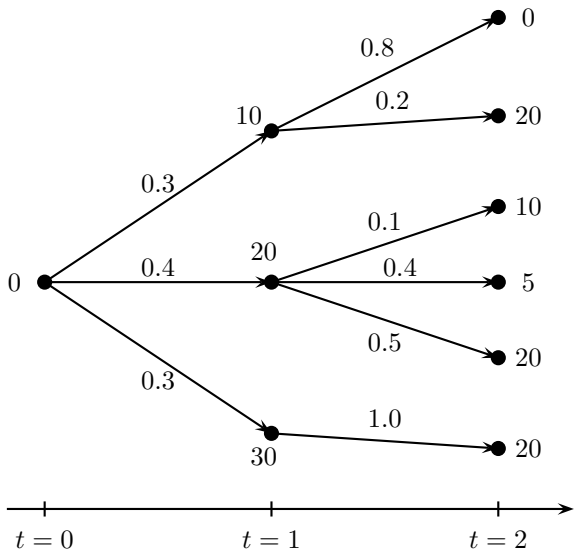


Figure 2.4: The multinomial tree version of the dividend process in Figure 2.3.

process.") The dividend of asset i is thus represented by a stochastic process $D_i = (D_{it})_{t\in\mathcal{T}}$, where each D_{it} is a random variable. We can form multi-dimensional stochastic processes by stacking one-dimensional stochastic processes. For example, we can represent the dividends of I assets by an I-dimensional stochastic process $\boldsymbol{D} = (\boldsymbol{D}_t)_{t\in\mathcal{T}}$, where $\boldsymbol{D}_t = (D_{1t}, \dots, D_{It})^\top$.

In general, for any $t \in \mathcal{T}$, the dividend at time t will be known at time t, but not before. The random variable D_{it} is then said to be $\mathcal{F}_t$-measurable, where $\mathcal{F}_t$ is the sigma-algebra representing the information at time t. If this information is equivalently represented by a partition $\mathbf{F}_t = \{F_{t1}, F_{t2}, \dots\}$, measurability means that D_{it} is constant on each of the elements F_{tj} of the partition. Note that this is indeed the case in our example in Figure 2.3. If D_{it} is $\mathcal{F}_t$-measurable for any $t \in \mathcal{T}$, the stochastic process $D_i = (D_{it})_{t\in\mathcal{T}}$ is said to be **adapted** to the information filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$. Since the dividends of assets and the income of individuals are assumed to be part of the exogenous uncertainty, it is natural to assume that dividend processes and income processes are adapted to the information filtration. In fact, most concrete models write down stochastic processes for all the exogenous variables and define the information filtration as the smallest filtration to which these exogenous processes are adapted.

An individual can choose how much to consume and which portfolio to invest in at any point in time, of course subject to a budget constraint and other feasibility constraints. The consumption and portfolio chosen at a given future date are likely to depend on the income the individual has received up to that point and her information about her future income and the future dividends of assets. The consumption rate at a given future date is therefore a random variable and the consumption rates at all dates constitute a stochastic process, the consumption process. Similarly, the portfolios chosen at different dates form a stochastic process, the portfolio process or so-called trading strategy. Representing consumption and investments by stochastic processes does not mean that we treat them as being exogenously given, but simply that the individual will condition her consumption and portfolio decisions on the information received. Since we assume that the underlying model of uncertainty includes all the uncertainty relevant for the decisions of the individuals, it is natural to require that consumption and portfolio processes are adapted to the information filtration.

Now consider prices. In a multi-period model we need to keep track of prices at all points in time. The price of a given asset at a given point in time depends on the supply and demand for that asset of all individuals, which again depends on the information individuals have at that time. Hence, we can also represent prices by adapted stochastic processes. To sum up, all the stochastic process relevant for our purposes will be adapted to the information filtration representing the resolution of all the relevant uncertainty.

Next, we introduce some further terminology often used in connection with stochastic processes. A stochastic process $X = (X_t)_{t\in\mathcal{T}}$ is said to be a **martingale** (relative to the probability measure $\mathbb{P}$ and the information filtration $(\mathcal{F}_t)_{t\in\mathcal{T}}$), if for all $t, t' \in \mathcal{T}$ with $t < t'$

$$\mathrm{E}_t[X_{t'}] = X_t,$$

i.e. no change in the value is expected.

A **sample path** of a stochastic process X is the collection of realized values $(X_t(\omega))_{t\in\mathcal{T}}$ for a given outcome $\omega \in \Omega$. The **value space** of a stochastic process is the smallest set $\mathcal{S}$ with the property that $\mathbb{P}(\{X_t \in \mathcal{S}\}) = 1$ for all $t \in \mathcal{T}$. If the value space has countably many elements,

the stochastic process is called a discrete-value process. Otherwise, it is called a continuous-value process. Of course, if the state space Ω is finite, all processes will be discrete-value processes. If you want to model continuous-value processes, you need an infinite state space.

For the modeling of most time-varying economic objects it seems reasonable to use continuous-value processes. Admittedly, stock prices are quoted on exchanges as multiples of some smallest possible unit (0.01 currency units in many countries) and interest rates are rounded off to some number of decimals, but the set of possible values of such objects is approximated very well by an interval in $\mathbb{R}$ (maybe $\mathbb{R}_+$ or $\mathbb{R}$ itself). Also, the mathematics involved in the analysis of continuous-value processes is simpler and more elegant than the mathematics for discrete-value processes. However there are economic objects that can only take on a very limited set of values. For these objects discrete-value processes should be used. An example is the credit ratings assigned by Moody's and similar agencies to debt issues of corporations.

As time goes by, we can observe the evolution in the object which the stochastic process describes. At any given time t', the previous values $(X_t)_{t\in[0,t')}$, where $X_t \in \mathcal{S}$, will be known (at least in the models we consider). These values constitute the **history** of the process up to time t'. The future values are still stochastic.

As time passes we will typically revise our expectations of the future values of the process or, more precisely, revise the probability distribution we attribute to the value of the process at any future point in time, cf. the discussion in the previous section. Suppose we stand at time t and consider the value of a process X at a future time $t' > t$. The distribution of the value of $X_{t'}$ is characterized by probabilities $\mathbb{P}(X_{t'} \in A)$ for subsets A of the value space $\mathcal{S}$. If for all $t, t' \in \mathcal{T}$ with $t < t'$ and all $A \subseteq \mathcal{S}$, we have that

$$\mathbb{P}\left(X_{t'} \in A \mid (X_s)_{s\leq t}\right) = \mathbb{P}\left(X_{t'} \in A \mid X_t\right),$$

then X is called a **Markov process**. Broadly speaking, this condition says that, given the presence, the future is independent of the past. The history contains no information about the future value that cannot be extracted from the current value.

2.5 Some discrete-time stochastic processes

In most discrete-time financial models the basic uncertainty is described by a sequence $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_T$ of random variables, one for each point in time. Think of ε_t as an exogenous shock to the financial market at time t. We assume that the shocks at different points in time are mutually independent, that each shock has a mean of zero and a variance of one. The shock at any given point in time t can be multi-variate, in which case we will write it as a vector, $\boldsymbol{\varepsilon}_t$. In that case the elements of the vector are assumed to be mutually independent. We assume that the shocks at all points in time have the same dimension. The distribution of the exogenous shocks has to be specified in the model. Typically, the shocks are assumed to be normally distributed (infinite state space) but models with a binomial or multinominal structure (finite state space) also exist.

The filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t\in\mathcal{T}})$ is defined implicitly from the assumptions on the exogenous shocks. For example, assume that the exogenous shocks $\varepsilon_1, \ldots, \varepsilon_T$ are $N(0,1)$ distributed. The state space is then the set of all possible realizations of all the T shocks, which is equivalent to $\mathbb{R}^T$. The σ-algebra $\mathcal{F}$ is the set of events that can be assigned a probability, which is

the set of (Borel-)subsets of $\mathbb{R}^T$. The probability measure $\mathbb{P}$ is defined via the normality assumption as

$$\mathbb{P}(\varepsilon_t < h) = N(h) \equiv \int_{-\infty}^{h} \frac{1}{\sqrt{2\pi}} e^{-a^2/2}\, da, \quad t = 1, \ldots, T,$$

where $N(\cdot)$ is the cumulative distribution function for an $N(0,1)$ variable. Probabilities of other events will follow from the above. The information at time t is represented by the smallest σ-algebra with respect to which the random variables $\varepsilon_1, \ldots, \varepsilon_t$ are measurable.

Stochastic processes for dividends etc. can be defined relative to the assumed exogenous shocks. It is easy to obtain non-zero means, non-unit variances, and dependencies across time. A discrete-time stochastic process $X = (X_t)_{t\in\mathcal{T}}$ is typically specified by the initial value X_0 (a constant in $\mathbb{R}$) and the increments over each period, i.e. $\Delta X_{t+1} \equiv X_{t+1} - X_t$ for each $t = 0, 1, \ldots, T-1$. The increments ΔX_{t+1} are defined in terms of the exogenous shocks $\varepsilon_1, \ldots, \varepsilon_{t+1}$ which implies that the process X is adapted.

Let us look at some discrete-time stochastic processes frequently used in asset pricing models. In all the examples we assume that the exogenous shocks $\varepsilon_1, \ldots, \varepsilon_T$ are independent, one-dimensional $N(0,1)$ distributed random variables.

Random walk: $\Delta X_{t+1} = \sigma\varepsilon_{t+1}$ or, equivalently, $X_{t+1} = X_0 + \sigma(\varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_{t+1})$. Here σ is a positive constant. A random walk is a Markov process since only the current value X_t and not the previous values $X_{t-1}, X_{t-2}, \ldots$ affect X_{t+1}. Since the expected change over any single period is zero, the expected change over any time interval will be zero, so a random walk is a martingale. Conditionally on X_t, X_{t+1} is normally distributed with mean X_t and variance σ^2. X_{t+1} is unconditionally (i.e. seen from time 0) normally distributed with mean X_0 and variance $(t+1)\sigma^2$.

Random walk with drift: $\Delta X_{t+1} = \mu + \sigma\varepsilon_{t+1}$, where μ is a constant (the drift rate) and σ is a positive constant. Also a random walk with drift is a Markov process. The expected change over any single period is μ so unless $\mu = 0$ and we are back to the random walk (without drift), the process $X = (X_t)_{t\in\mathcal{T}}$ is not a martingale. Conditionally on X_t, X_{t+1} is normally distributed with mean $\mu + X_t$ and variance σ^2. X_{t+1} is unconditionally normally distributed with mean $X_0 + (t+1)\mu$ and variance $(t+1)\sigma^2$.

Autoregressive processes: A process $X = (X_t)_{t\in\mathcal{T}}$ with

$$\Delta X_{t+1} = (1-\rho)(\mu - X_t) + \sigma\varepsilon_{t+1},$$

where $\rho \in (-1, 1)$, is said to be an autoregressive process of order 1 or simply an AR(1) process. It is a Markov process since only the current value X_t affects the distribution of the future value. The expected change over the period is positive if $X_t < \mu$ and negative if $X_t > \mu$. In any case, the value X_{t+1} is expected to be closer to μ than X_t is. The process is pulled towards μ and therefore the process is said to be mean-reverting. This is useful for modeling the dynamics of variables that tend to vary with the business cycle around some long-term average. Note, however, extreme shocks may cause the process to be pushed further away from μ.

The covariance between the subsequent values X_t and $X_{t+1} = X_t + (1-\rho)(\mu - X_t) + \sigma\varepsilon_{t+1} = \rho X_t + (1-\rho)\mu + \sigma\varepsilon_{t+1}$ is

$$\operatorname{Cov}[X_t, X_{t+1}] = \rho \operatorname{Cov}[X_t, X_t] = \rho \operatorname{Var}[X_t]$$

so that $\rho = \mathrm{Cov}[X_t, X_{t+1}]/\mathrm{Var}[X_t]$ is the so-called auto-correlation parameter. Solving backwards, we find

$$\begin{aligned}
X_{t+1} &= \rho X_t + (1-\rho)\mu + \sigma\varepsilon_{t+1} \\
&= \rho\left(\rho X_{t-1} + (1-\rho)\mu + \sigma\varepsilon_t\right) + (1-\rho)\mu + \sigma\varepsilon_{t+1} \\
&= \rho^2 X_{t-1} + (1+\rho)(1-\rho)\mu + \sigma\varepsilon_{t+1} + \rho\sigma\varepsilon_t \\
&= \dots \\
&= \rho^{k+1} X_{t-k} + (1+\rho+\cdots+\rho^k)(1-\rho)\mu + \sigma\varepsilon_{t+1} + \rho\sigma\varepsilon_t + \cdots + \rho^k\sigma\varepsilon_{t-k+1} \\
&= \rho^{k+1} X_{t-k} + (1-\rho^{k+1})\mu + \sigma\sum_{j=0}^{k}\rho^j\varepsilon_{t+1-j}.
\end{aligned}$$

More generally, a process $X = (X_t)_{t\in\mathcal{T}}$ with

$$X_{t+1} = \mu + \rho_1(X_t - \mu) + \rho_2(X_{t-1} - \mu) + \cdots + \rho_l(X_{t-l+1} - \mu) + \sigma\varepsilon_{t+1}$$

is said to be an autoregressive process of order l or simply an AR(l) process. If the order is higher than 1, the process is not a Markov process.

(G)ARCH processes: ARCH is short for Autoregressive Conditional Heteroskedasticity. An ARCH(l) process $X = (X_t)_{t\in\mathcal{T}}$ is defined by

$$X_{t+1} = \mu + \sigma_{t+1}\varepsilon_{t+1},$$

where

$$\sigma_{t+1}^2 = \delta + \sum_{i=1}^{l}\alpha_i\varepsilon_{t+1-i}^2.$$

The conditional variance depends on squares of the previous l shock terms.

GARCH is short for Generalized Autoregressive Conditional Heteroskedasticity. A GARCH(l, m) process $X = (X_t)_{t\in\mathcal{T}}$ is defined by

$$X_{t+1} = \mu + \sigma_{t+1}\varepsilon_{t+1},$$

where

$$\sigma_{t+1}^2 = \delta + \sum_{i=1}^{l}\alpha_i\varepsilon_{t+1-i}^2 + \sum_{j=1}^{m}\beta_j\sigma_{t+1-j}^2.$$

ARCH and GARCH processes are often used for detailed modeling of stock market volatility.

More generally we can define an adapted process $X = (X_t)_{t\in\mathcal{T}}$ by the initial value X_0 and the equation

$$\Delta X_{t+1} = \mu(X_t, \dots, X_0) + \sigma(X_t, \dots, X_0)\varepsilon_{t+1}, \quad t = 0, 1, \dots, T-1, \tag{2.1}$$

where μ and σ are real-valued functions. If $\varepsilon_{t+1} \sim N(0,1)$, the conditional distribution of X_{t+1} given X_t is a normal distribution with mean $X_t + \mu(X_t, \dots, X_0)$ and variance $\sigma(X_t, \dots, X_0)^2$.

We can write the stochastic processes introduced above in a different way that will ease the transition to continuous-time processes. Let $z = (z_t)_{t\in\mathcal{T}}$ denote a unit random walk starting at zero, i.e. a process with the properties

(i) $z_0 = 0$,

(ii) $\Delta z_{t+1} \equiv z_{t+1} - z_t \sim N(0,1)$ for all $t = 0, 1, \ldots, T-1$,

(iii) $\Delta z_1, \Delta z_2, \ldots \Delta z_T$ are independent.

Then we can define a random walk with drift as a process X with

$$\Delta X_{t+1} = \mu + \sigma \Delta z_{t+1},$$

an AR(1) process is defined by

$$\Delta X_{t+1} = (1-\rho)(\mu - X_t) + \sigma \Delta z_{t+1},$$

and a general adapted process is defined by

$$\Delta X_{t+1} = \mu(X_t, \ldots, X_0) + \sigma(X_t, \ldots, X_0)\Delta z_{t+1}.$$

We will see very similar equations in continuous time.

2.6 Continuous-time stochastic processes

2.6.1 Brownian motions

In the continuous-time asset pricing models we will consider in this book, the basic uncertainty in the economy is represented by the evolution of a standard Brownian motion. A (one-dimensional) stochastic process $z = (z_t)_{t\in[0,T]}$ is called a **standard Brownian motion**, if it satisfies the following conditions:

(i) $z_0 = 0$,

(ii) for all $t, t' \geq 0$ with $t < t'$: $z_{t'} - z_t \sim N(0, t'-t)$ [normally distributed increments],

(iii) for all $0 \leq t_0 < t_1 < \cdots < t_n$, the random variables $z_{t_1} - z_{t_0}$, …, $z_{t_n} - z_{t_{n-1}}$ are mutually independent [independent increments],

(iv) z has continuous sample paths.

The first three conditions are equivalent to the discrete-time case studied above. We can informally think of $dz_t \approx z_{t+dt} - z_t \sim N(0, dt)$ as an exogenous shock to the economy at time t. The state space Ω is in this case the (infinite) set of all paths of the standard Brownian motion z. The information filtration $(\mathcal{F}_t)_{t\in[0,T]}$ is generated by the standard Brownian motion z in the sense that, for each t, $\mathcal{F}_t$ is the smallest σ-algebra on which the random variable z_t is measurable. The probability measure $\mathbb{P}$ is fixed by the normality assumption.

Any continuous-time stochastic process $X = (X_t)_{t\in[0,T]}$ in the financial models in this book will be defined in terms of the standard Brownian motion by an initial constant value X_0 and an equation of the form

$$dX_t = \mu_t\, dt + \sigma_t\, dz_t.$$

Here μ_t and σ_t are known at time t (measurable with respect to $\mathcal{F}_t$) but may depend on X_s and z_s for $s \leq t$. We can informally think of dX_t as the increment $X_{t+dt} - X_t$ over the "instant" (of

length dt) following time t. Since dz_t has mean zero and variance dt, we can informally compute the conditional mean and variance of dX_t as

$$\mathrm{E}_t[dX_t] = \mu_t\,dt, \qquad \mathrm{Var}_t[dX_t] = \sigma_t^2\,dt.$$

Therefore we can interpret μ_t and σ_t^2 as the conditional mean and conditional variance of the change in the value of the process per time unit. The properties of the process X will depend on the specification of μ_t and σ_t. We will be more formal and give examples below.

The standard Brownian motion is basically the continuous-time version of a random walk with initial value 0. The standard Brownian motion is a Markov process because the increment from today to any future point in time is independent of the history of the process. The standard Brownian motion is also a martingale since the expected change in the value of the process is zero. The name Brownian motion is in honor of the Scottish botanist Robert Brown, who in 1828 observed the apparently random movements of pollen submerged in water. The often used name Wiener process is due to Norbert Wiener, who in the 1920s was the first to show the existence of a stochastic process with these properties and who developed a mathematically rigorous analysis of the process. As early as in the year 1900, the standard Brownian motion was used in a model for stock price movements by the French researcher Louis Bachelier, who derived the first option pricing formula.

The defining characteristics of a standard Brownian motion look very nice, but they have some drastic consequences. It can be shown that the sample paths of a standard Brownian motion are nowhere differentiable, which broadly speaking means that the sample paths bend at all points in time and are therefore strictly speaking impossible to illustrate. However, one can get an idea of the sample paths by simulating the values of the process at different times. If $\varepsilon_1, \ldots, \varepsilon_n$ are independent draws from a standard $N(0,1)$ distribution, we can simulate the value of the standard Brownian motion at time $0 \equiv t_0 < t_1 < t_2 < \cdots < t_n$ as follows:

$$z_{t_i} = z_{t_{i-1}} + \varepsilon_i\sqrt{t_i - t_{i-1}}, \quad i = 1, \ldots, n.$$

With more time points and hence shorter intervals we get a more realistic impression of the sample paths of the process. Figure 2.5 shows a simulated sample path for a standard Brownian motion over the interval $[0,1]$ based on a partition of the interval into 200 subintervals of equal length.[1] Note that since a normally distributed random variable can take on infinitely many values, a standard Brownian motion has infinitely many sample paths that each has a zero probability of occurring. The figure shows just one possible sample path. Note that the picture resembles typical stock price charts.

Another property of a standard Brownian motion is that the expected length of the sample path over any future time interval (no matter how short) is infinite. In addition, the expected number of times a standard Brownian motion takes on any given value in any given time interval

[1] Most spreadsheets and programming tools have a built-in procedure that generates uniformly distributed numbers over the interval $[0,1]$. Such uniformly distributed random numbers can be transformed into standard normally distributed numbers in several ways. One example: Given uniformly distributed numbers U_1 and U_2, the numbers ε_1 and ε_2 defined by

$$\varepsilon_1 = \sqrt{-2\ln U_1}\sin(2\pi U_2), \qquad \varepsilon_2 = \sqrt{-2\ln U_1}\cos(2\pi U_2)$$

will be independent standard normally distributed random numbers. This is the so-called Box-Muller transformation. See e.g. Press, Teukolsky, Vetterling, and Flannery (1992, Sec. 7.2).

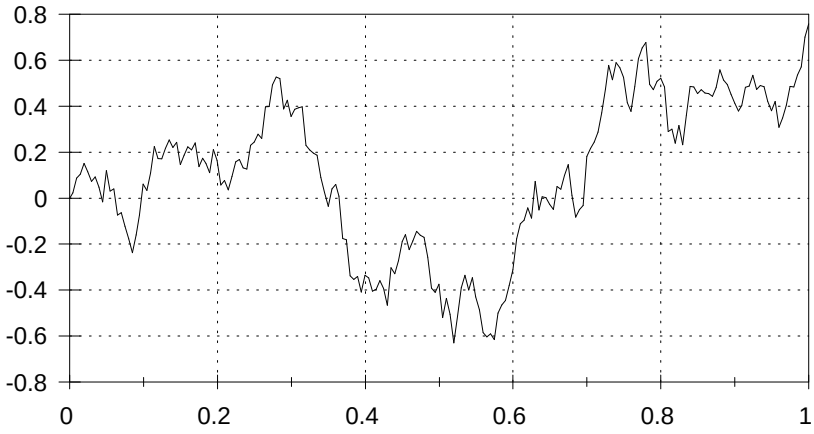


Figure 2.5: A simulated sample path of a standard Brownian motion based on 200 subintervals.

is also infinite. Intuitively, these properties are due to the fact that the size of the increment of a standard Brownian motion over an interval of length Δt is proportional to $\sqrt{\Delta t}$, in the sense that the standard deviation of the increment equals $\sqrt{\Delta t}$. When Δt is close to zero, $\sqrt{\Delta t}$ is significantly larger than Δt, so the changes are large relative to the length of the time interval over which the changes are measured.

The expected change in an object described by a standard Brownian motion equals zero and the variance of the change over a given time interval equals the length of the interval. This can easily be generalized. As before let $z = (z_t)_{t\geq 0}$ be a one-dimensional standard Brownian motion and define a new stochastic process $X = (X_t)_{t\geq 0}$ by

$$X_t = X_0 + \mu t + \sigma z_t, \quad t \geq 0, \tag{2.2}$$

where X_0, μ, and σ are constants. The constant X_0 is the initial value for the process X. It follows from the properties of the standard Brownian motion that, seen from time 0, the value X_t is normally distributed with mean μt and variance $\sigma^2 t$, i.e. $X_t \sim N(X_0 + \mu t, \sigma^2 t)$.

The change in the value of the process between two arbitrary points in time t and t', where $t < t'$, is given by

$$X_{t'} - X_t = \mu(t' - t) + \sigma(z_{t'} - z_t).$$

The change over an infinitesimally short interval $[t, t + \Delta t]$ with $\Delta t \to 0$ is often written as

$$dX_t = \mu\, dt + \sigma\, dz_t, \tag{2.3}$$

where dz_t can loosely be interpreted as a $N(0, dt)$-distributed random variable. To give this a precise mathematical meaning, it must be interpreted as a limit of the expression

$$X_{t+\Delta t} - X_t = \mu\Delta t + \sigma(z_{t+\Delta t} - z_t)$$

for $\Delta t \to 0$. The process X is called a **generalized Brownian motion** or a generalized Wiener process. This is basically the continuous-time version of a random walk with drift. The parameter

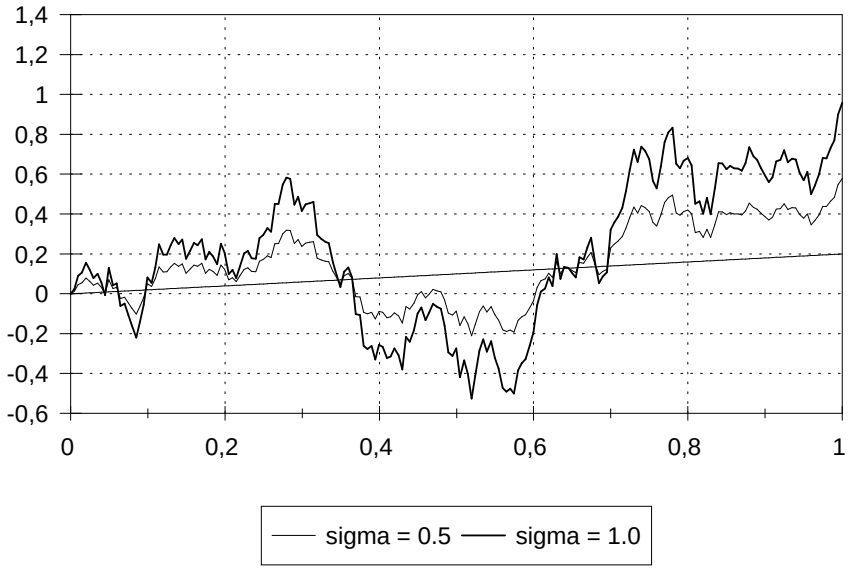


Figure 2.6: Simulation of a generalized Brownian motion with $\mu = 0.2$ and $\sigma = 0.5$ or $\sigma = 1.0$. The straight line shows the trend corresponding to $\sigma = 0$. The simulations are based on 200 subintervals.

μ reflects the expected change in the process per unit of time and is called the drift rate or simply the **drift** of the process. The parameter σ reflects the uncertainty about the future values of the process. More precisely, σ^2 reflects the variance of the change in the process per unit of time and is often called the **variance rate** of the process. σ is a measure for the standard deviation of the change per unit of time and is referred to as the **volatility** of the process.

A generalized Brownian motion inherits many of the characteristic properties of a standard Brownian motion. For example, also a generalized Brownian motion is a Markov process, and the sample paths of a generalized Brownian motion are also continuous and nowhere differentiable. However, a generalized Brownian motion is not a martingale unless $\mu = 0$. The sample paths can be simulated by choosing time points $0 \equiv t_0 < t_1 < \cdots < t_n$ and iteratively computing

$$X_{t_i} = X_{t_{i-1}} + \mu(t_i - t_{i-1}) + \varepsilon_i \sigma \sqrt{t_i - t_{i-1}}, \quad i = 1, \ldots, n,$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent draws from a standard normal distribution. Figure 2.6 show simulated sample paths for two different values of σ but the same μ. The paths are drawn using the same sequence of random numbers ε_i so that they are directly comparable. The straight line represent the deterministic trend of the process, which corresponds to imposing the condition $\sigma = 0$ and hence ignoring the uncertainty. The parameter μ determines the trend, and the parameter σ determines the size of the fluctuations around the trend.

If the parameters μ and σ are allowed to be time-varying in a deterministic way, the process X is said to be a *time-inhomogeneous* generalized Brownian motion. In differential terms such a

process can be written as defined by

$$dX_t = \mu(t)\,dt + \sigma(t)\,dz_t. \tag{2.4}$$

Over a very short interval $[t, t+\Delta t]$ the expected change is approximately $\mu(t)\Delta t$, and the variance of the change is approximately $\sigma(t)^2\Delta t$. More precisely, the increment over any interval $[t, t']$ is given by

$$X_{t'} - X_t = \int_t^{t'} \mu(u)\,du + \int_t^{t'} \sigma(u)\,dz_u. \tag{2.5}$$

The last integral is a so-called stochastic integral, which we will define (although not rigorously) and describe in a later section. There we will also state a theorem, which implies that, seen from time t, the integral $\int_t^{t'} \sigma(u)\,dz_u$ is a normally distributed random variable with mean zero and variance $\int_t^{t'} \sigma(u)^2\,du$.

2.6.2 Diffusion processes

For both standard Brownian motions and generalized Brownian motions, the future value is normally distributed and can therefore take on any real value, i.e. the value space is equal to $\mathbb{R}$. Many economic variables can only have values in a certain subset of $\mathbb{R}$. For example, prices of financial assets with limited liability are non-negative. The evolution in such variables cannot be well represented by the stochastic processes studied so far. In many situations we will instead use so-called diffusion processes.

A (one-dimensional) **diffusion process** is a stochastic process $X = (X_t)_{t\geq 0}$ for which the change over an infinitesimally short time interval $[t, t+dt]$ can be written as

$$dX_t = \mu(X_t, t)\,dt + \sigma(X_t, t)\,dz_t, \tag{2.6}$$

where z is a standard Brownian motion, but where the drift μ and the volatility σ are now functions of time and the current value of the process.[2] This expression generalizes (2.3), where μ and σ were assumed to be constants, and (2.4), where μ and σ were functions of time only. An equation like (2.6), where the stochastic process enters both sides of the equality, is called a **stochastic differential equation**. Hence, a diffusion process is a solution to a stochastic differential equation.

If both functions μ and σ are independent of time, the diffusion is said to be **time-homogeneous**, otherwise it is said to be **time-inhomogeneous**. For a time-homogeneous diffusion process, the distribution of the future value will only depend on the current value of the process and how far into the future we are looking – not on the particular point in time we are standing at. For example, the distribution of $X_{t+\delta}$ given $X_t = x$ will only depend on X and δ, but not on t. This is not the case for a time-inhomogeneous diffusion, where the distribution will also depend on t.

In the expression (2.6) one may think of dz_t as being $N(0, dt)$-distributed, so that the mean and variance of the change over an infinitesimally short interval $[t, t+dt]$ are given by

$$\mathrm{E}_t[dX_t] = \mu(X_t, t)\,dt, \qquad \mathrm{Var}_t[dX_t] = \sigma(X_t, t)^2\,dt.$$

[2] For the process X to be mathematically meaningful, the functions $\mu(x, t)$ and $\sigma(x, t)$ must satisfy certain conditions. See e.g. Øksendal (1998, Ch. 7) and Duffie (2001, App. E).

To be more precise, the change in a diffusion process over any interval $[t, t']$ is

$$X_{t'} - X_t = \int_t^{t'} \mu(X_u, u)\, du + \int_t^{t'} \sigma(X_u, u)\, dz_u. \tag{2.7}$$

Here the integrand of the first integral $\int_t^{t'} \mu(X_u, u)\, du$ depends on the values X_u for $u \in [t, t']$, which are generally unknown at time t. It is therefore natural to define the integral $\int_t^{t'} \mu(X_u, u)\, du$ as the random variable which in state $\omega \in \Omega$ has the value $\int_t^{t'} \mu(X_u(\omega), u)\, du$, which is now just the integration of a real-valued function of time. The other integral $\int_t^{t'} \sigma(X_u, u)\, dz_u$ is a so-called stochastic integral, which we will discuss in Section 2.6.5.

We will often use the informal and intuitive differential notation (2.6). The drift rate $\mu(X_t, t)$ and the variance rate $\sigma(X_t, t)^2$ are really the limits

$$\mu(X_t, t) = \lim_{\Delta t \to 0} \frac{\mathrm{E}_t\left[X_{t+\Delta t} - X_t\right]}{\Delta t},$$
$$\sigma(X_t, t)^2 = \lim_{\Delta t \to 0} \frac{\mathrm{Var}_t\left[X_{t+\Delta t} - X_t\right]}{\Delta t}.$$

A diffusion process is a Markov process as can be seen from (2.6), since both the drift and the volatility only depend on the current value of the process and not on previous values. A diffusion process is not a martingale, unless the drift $\mu(X_t, t)$ is zero for all X_t and t. A diffusion process will have continuous, but nowhere differentiable sample paths. The value space for a diffusion process and the distribution of future values will depend on the functions μ and σ. In Section 2.6.7 we will an example of a diffusion process often used in financial modeling, the so-called geometric Brownian motion. Other diffusion processes will be used in later chapters.

2.6.3 Itô processes

It is possible to define even more general continuous-path processes than those in the class of diffusion processes. A (one-dimensional) stochastic process X_t is said to be an **Itô process**, if the local increments are on the form

$$dX_t = \mu_t\, dt + \sigma_t\, dz_t, \tag{2.8}$$

where the drift μ and the volatility σ themselves are stochastic processes. A diffusion process is the special case where the values of the drift μ_t and the volatility σ_t are given as functions of t and X_t. For a general Itô process, the drift and volatility may also depend on past values of the X process and also on past and current values of other adapted processes. It follows that Itô processes are generally not Markov processes. They are generally not martingales either, unless μ_t is identically equal to zero (and σ_t satisfies some technical conditions). The processes μ and σ must satisfy certain regularity conditions for the X process to be well-defined. We will refer the reader to Øksendal (1998, Ch. 4) for these conditions. The expression (2.8) gives an intuitive understanding of the evolution of an Itô process, but it is more precise to state the evolution in the integral form

$$X_{t'} - X_t = \int_t^{t'} \mu_u\, du + \int_t^{t'} \sigma_u\, dz_u. \tag{2.9}$$

Again the first integral can be defined "state-by-state" and the second integral is a stochastic integral.

2.6.4 Jump processes

Above we have focused on processes having sample paths that are continuous functions of time, so that one can depict the evolution of the process by a continuous curve. Stochastic processes which have sample paths with discontinuities (jumps) also exist. The jumps of such processes are often modeled by Poisson processes or related processes. It is well-known that large, sudden movements in financial variables occur from time to time, for example in connection with stock market crashes. There may be many explanations of such large movements, for example a large unexpected change in the productivity in a particular industry or the economy in general, perhaps due to a technological break-through. Another source of sudden, large movements is changes in the political or economic environment such as unforseen interventions by the government or central bank. Stock market crashes are sometimes explained by the bursting of a bubble (which does not necessarily conflict with the usual assumption of rational investors). Whether such sudden, large movements can be explained by a sequence of small continuous movements in the same direction or jumps have to be included in the models is an empirical question, which is still open. While jump processes may be relevant for many purposes, they are also more difficult to deal with than processes with continuous sample paths so that it will probably be best to study models without jumps first. This book will only address continuous-path processes. An overview of financial models with jump processes is given by Cont and Tankov (2004).

2.6.5 Stochastic integrals

In (2.7) and (2.9) and similar expressions a term of the form $\int_t^{t'} \sigma_u \, dz_u$ appears. An integral of this type is called a stochastic integral or an Itô integral. For given $t < t'$, the stochastic integral $\int_t^{t'} \sigma_u \, dz_u$ is a random variable. Assuming that σ_u is known at time u, the value of the integral becomes known at time t'. The process σ is called the integrand. The stochastic integral can be defined for very general integrands. The simplest integrands are those that are piecewise constant. Assume that there are points in time $t \equiv t_0 < t_1 < \cdots < t_n \equiv t'$, so that σ_u is constant on each subinterval $[t_i, t_{i+1})$. The stochastic integral is then defined by

$$\int_t^{t'} \sigma_u \, dz_u = \sum_{i=0}^{n-1} \sigma_{t_i} \left(z_{t_{i+1}} - z_{t_i} \right). \tag{2.10}$$

If the integrand process σ is not piecewise constant, there will exist a sequence of piecewise constant processes $\sigma^{(1)}, \sigma^{(2)}, \ldots$, which converges to σ. For each of the processes $\sigma^{(m)}$, the integral $\int_t^{t'} \sigma_u^{(m)} \, dz_u$ is defined as above. The integral $\int_t^{t'} \sigma_u \, dz_u$ is then defined as a limit of the integrals of the approximating processes:

$$\int_t^{t'} \sigma_u \, dz_u = \lim_{m \to \infty} \int_t^{t'} \sigma_u^{(m)} \, dz_u. \tag{2.11}$$

We will not discuss exactly how this limit is to be understood and which integrand processes we can allow. Again the interested reader is referred to Øksendal (1998). The distribution of the integral $\int_t^{t'} \sigma_u \, dz_u$ will, of course, depend on the integrand process and can generally not be completely characterized, but the following theorem gives the mean and the variance of the integral:

Theorem 2.2 *The stochastic integral* $\int_t^{t'} \sigma_u \, dz_u$ *has the following properties:*

$$\mathrm{E}_t \left[\int_t^{t'} \sigma_u \, dz_u \right] = 0,$$
$$\mathrm{Var}_t \left[\int_t^{t'} \sigma_u \, dz_u \right] = \int_t^{t'} \mathrm{E}_t[\sigma_u^2] \, du.$$

Proof: Suppose that σ is piecewise constant and divide the interval $[t, t']$ into subintervals defined by the time points $t \equiv t_0 < t_1 < \cdots < t_n \equiv t'$ so that σ is constant on each subinterval $[t_i, t_{i+1})$ with a value σ_{t_i} which is known at time t_i. Then

$$\mathrm{E}_t \left[\int_t^{t'} \sigma_u \, dz_u \right] = \sum_{i=0}^{n-1} \mathrm{E}_t \left[\sigma_{t_i} \left(z_{t_{i+1}} - z_{t_i} \right) \right] = \sum_{i=0}^{n-1} \mathrm{E}_t \left[\sigma_{t_i} \, \mathrm{E}_{t_i} \left[\left(z_{t_{i+1}} - z_{t_i} \right) \right] \right] = 0,$$

using the Law of Iterated Expectations. For the variance we have

$$\mathrm{Var}_t \left[\int_t^{t'} \sigma_u \, dz_u \right] = \mathrm{E}_t \left[\left(\int_t^{t'} \sigma_u \, dz_u \right)^2 \right] - \left(\mathrm{E}_t \left[\int_t^{t'} \sigma_u \, dz_u \right] \right)^2 = \mathrm{E}_t \left[\left(\int_t^{t'} \sigma_u \, dz_u \right)^2 \right]$$

and

$$\begin{aligned} \mathrm{E}_t \left[\left(\int_t^{t'} \sigma_u \, dz_u \right)^2 \right] &= \mathrm{E}_t \left[\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \sigma_{t_i} \sigma_{t_j} (z_{t_{i+1}} - z_{t_i})(z_{t_{j+1}} - z_{t_j}) \right] \\ &= \sum_{i=0}^{n-1} \mathrm{E}_t \left[\sigma_{t_i}^2 (z_{t_{i+1}} - z_{t_i})^2 \right] = \sum_{i=0}^{n-1} \mathrm{E}_t \left[\sigma_{t_i}^2 \right] (t_{i+1} - t_i) = \int_t^{t'} \mathrm{E}_t[\sigma_u^2] \, du. \end{aligned}$$

If σ is not piecewise constant, we can approximate it by a piecewise constant process and take appropriate limits. □

If the integrand is a deterministic function of time, $\sigma(u)$, the integral will be normally distributed, so that the following result holds:

Theorem 2.3 *If* z *is a Brownian motion, and* $\sigma(u)$ *is a deterministic function of time, the random variable* $\int_t^{t'} \sigma(u) \, dz_u$ *is normally distributed with mean zero and variance* $\int_t^{t'} \sigma(u)^2 \, du$.

Proof: We present a sketch of the proof. Dividing the interval $[t, t']$ into subintervals defined by the time points $t \equiv t_0 < t_1 < \cdots < t_n \equiv t'$, we can approximate the integral with the sum

$$\int_t^{t'} \sigma(u) \, dz_u \approx \sum_{i=0}^{n-1} \sigma(t_i) \left(z_{t_{i+1}} - z_{t_i} \right).$$

The increment of the Brownian motion over any subinterval is normally distributed with mean zero and a variance equal to the length of the subinterval. Furthermore, the different terms in the sum are mutually independent. It is well-known that a sum of normally distributed random variables is itself normally distributed, and that the mean of the sum is equal to the sum of the means, which in the present case yields zero. Due to the independence of the terms in the sum, the variance of the sum is also equal to the sum of the variances, i.e.

$$\mathrm{Var}_t \left[\sum_{i=0}^{n-1} \sigma(t_i) \left(z_{t_{i+1}} - z_{t_i} \right) \right] = \sum_{i=0}^{n-1} \sigma(t_i)^2 \, \mathrm{Var}_t \left[z_{t_{i+1}} - z_{t_i} \right] = \sum_{i=0}^{n-1} \sigma(t_i)^2 (t_{i+1} - t_i),$$

which is an approximation of the integral $\int_t^{t'} \sigma(u)^2\,du$. The result now follows from an appropriate limit where the subintervals shrink to zero length. □

Note that the process $y = (y_t)_{t\geq 0}$ defined by $y_t = \int_0^t \sigma_u\,dz_u$ is a martingale, since

$$\begin{aligned}\mathrm{E}_t[y_{t'}] &= \mathrm{E}_t\left[\int_0^{t'} \sigma_u\,dz_u\right] = \mathrm{E}_t\left[\int_0^t \sigma_u\,dz_u + \int_t^{t'} \sigma_u\,dz_u\right] \\ &= \mathrm{E}_t\left[\int_0^t \sigma_u\,dz_u\right] + \mathrm{E}_t\left[\int_t^{t'} \sigma_u\,dz_u\right] = \int_0^t \sigma_u\,dz_u = y_t,\end{aligned}$$

so that the expected future value is equal to the current value. More generally $y_t = y_0 + \int_0^t \sigma_u\,dz_u$ for some constant y_0, is a martingale. The converse is also true in the sense that any martingale can be expressed as a stochastic integral. This is the so-called martingale representation theorem:

Theorem 2.4 *Suppose the process $M = (M_t)$ is a martingale with respect to a probability measure under which $z = (z_t)$ is a standard Brownian motion. Then a unique adapted process $\theta = (\theta_t)$ exists such that*

$$M_t = M_0 + \int_0^t \theta_u\,dz_u$$

for all t.

For a mathematically more precise statement of the result and a proof, see Øksendal (1998, Thm. 4.3.4).

Now the stochastic integral with respect to the standard Brownian motion has been defined, we can also define stochastic integrals with respect to other stochastic processes. For example, if X_t is a diffusion given by $dX_t = \mu(X_t, t)\,dt + \sigma(X_t, t)\,dz_t$ and $\alpha = (\alpha_t)_{t\in[0,T]}$ is a sufficiently "nice" stochastic process, we can define

$$\int_0^t \alpha_u\,dX_u = \int_0^t \alpha_u\mu(X_u, u)\,du + \int_0^t \alpha_u\sigma(X_u, u)\,dz_u.$$

2.6.6 Itô's Lemma

In continuous-time models a stochastic process for the dynamics of some basic quantity is often taken as given, while other quantities of interest can be shown to be functions of that basic variable. To determine the dynamics of these other variables, we shall apply Itô's Lemma, which is basically the chain rule for stochastic processes. We will state the result for a function of a general Itô process, although we will frequently apply the result for the special case of a function of a diffusion process.

Theorem 2.5 *Let $X = (X_t)_{t\geq 0}$ be a real-valued Itô process with dynamics*

$$dX_t = \mu_t\,dt + \sigma_t\,dz_t,$$

where μ and σ are real-valued processes, and z is a one-dimensional standard Brownian motion. Let $g(X, t)$ be a real-valued function which is two times continuously differentiable in X and continuously differentiable in t. Then the process $y = (y_t)_{t\geq 0}$ defined by

$$y_t = g(X_t, t)$$

is an Itô process with dynamics

$$dy_t = \left(\frac{\partial g}{\partial t}(X_t, t) + \frac{\partial g}{\partial X}(X_t, t)\mu_t + \frac{1}{2}\frac{\partial^2 g}{\partial X^2}(X_t, t)\sigma_t^2 \right) dt + \frac{\partial g}{\partial X}(X_t, t)\sigma_t \, dz_t. \tag{2.12}$$

The proof of Itô's Lemma is based on a Taylor expansion of $g(X_t, t)$ combined with appropriate limits, but a formal proof is beyond the scope of this presentation. Once again, we refer to Øksendal (1998) and similar textbooks. The result can also be written in the following way, which may be easier to remember:

$$dy_t = \frac{\partial g}{\partial t}(X_t, t)\, dt + \frac{\partial g}{\partial X}(X_t, t)\, dX_t + \frac{1}{2}\frac{\partial^2 g}{\partial X^2}(X_t, t)(dX_t)^2. \tag{2.13}$$

Here, in the computation of $(dX_t)^2$, one must apply the rules $(dt)^2 = dt \cdot dz_t = 0$ and $(dz_t)^2 = dt$, so that

$$(dX_t)^2 = (\mu_t\, dt + \sigma_t\, dz_t)^2 = \mu_t^2 (dt)^2 + 2\mu_t\sigma_t\, dt \cdot dz_t + \sigma_t^2 (dz_t)^2 = \sigma_t^2\, dt.$$

The intuition behind these rules is as follows: When dt is close to zero, $(dt)^2$ is far less than dt and can therefore be ignored. Since $dz_t \sim N(0, dt)$, we get $\mathrm{E}[dt \cdot dz_t] = dt \cdot \mathrm{E}[dz_t] = 0$ and $\mathrm{Var}[dt \cdot dz_t] = (dt)^2 \mathrm{Var}[dz_t] = (dt)^3$, which is also very small compared to dt and is therefore ignorable. Finally, we have $\mathrm{E}[(dz_t)^2] = \mathrm{Var}[dz_t] - (\mathrm{E}[dz_t])^2 = dt$, and it can be shown that[3] $\mathrm{Var}[(dz_t)^2] = 2(dt)^2$. For dt close to zero, the variance is therefore much less than the mean, so $(dz_t)^2$ can be approximated by its mean dt.

In standard mathematics, the differential of a function $y = g(t, X)$ where t and X are real variables is defined as $dy = \frac{\partial g}{\partial t}\, dt + \frac{\partial g}{\partial X}\, dX$. When X is an Itô process, (2.13) shows that we have to add a second-order term.

2.6.7 The geometric Brownian motion

The geometric Brownian motion is an important example of a diffusion process. A stochastic process $X = (X_t)_{t \geq 0}$ is said to be a **geometric Brownian motion** if it is a solution to the stochastic differential equation

$$dX_t = \mu X_t\, dt + \sigma X_t\, dz_t, \tag{2.14}$$

where μ and σ are constants. The initial value for the process is assumed to be positive, $X_0 > 0$. A geometric Brownian motion is the particular diffusion process that is obtained from (2.6) by inserting $\mu(X_t, t) = \mu X_t$ and $\sigma(X_t, t) = \sigma X_t$.

The expression (2.14) can be rewritten as

$$\frac{dX_t}{X_t} = \mu\, dt + \sigma\, dz_t,$$

which is the relative (percentage) change in the value of the process over the next infinitesimally short time interval $[t, t + dt]$. If X_t is the price of a traded asset, then dX_t/X_t is the rate of return on the asset over the next instant. The constant μ is the expected rate of return per period, while σ is the standard deviation of the rate of return per period. In this context it is often μ which is called the drift (rather than μX_t) and σ which is called the volatility (rather than σX_t). Strictly speaking, one must distinguish between the relative drift and volatility (μ and σ, respectively) and

[3] This is based on the computation $\mathrm{Var}[(z_{t+\Delta t} - z_t)^2] = \mathrm{E}[(z_{t+\Delta t} - z_t)^4] - \left(\mathrm{E}[(z_{t+\Delta t} - z_t)^2]\right)^2 = 3(\Delta t)^2 - (\Delta t)^2 = 2(\Delta t)^2$ and a passage to the limit.

the absolute drift and volatility (μX_t and σX_t, respectively). An asset with a constant expected rate of return and a constant relative volatility has a price that follows a geometric Brownian motion. For example, such an assumption is used for the stock price in the famous Black-Scholes-Merton model for stock option pricing, cf. Chapter 12. In the framework of consumption-based capital asset pricing models it is often assumed that the aggregate consumption in the economy follows a geometric Brownian motion, cf. Chapter 8.

Next, we will find an explicit expression for X_t, i.e. we will find a solution to the stochastic differential equation (2.14). We can then also determine the distribution of the future value of the process. We apply Itô's Lemma with the function $g(x,t) = \ln x$ and define the process $y_t = g(X_t, t) = \ln X_t$. Since

$$\frac{\partial g}{\partial t}(X_t,t) = 0, \qquad \frac{\partial g}{\partial x}(X_t,t) = \frac{1}{X_t}, \qquad \frac{\partial^2 g}{\partial x^2}(X_t,t) = -\frac{1}{X_t^2},$$

we get from Theorem 2.5 that

$$dy_t = \left(0 + \frac{1}{X_t}\mu X_t - \frac{1}{2}\frac{1}{X_t^2}\sigma^2 X_t^2\right)\,dt + \frac{1}{X_t}\sigma X_t\,dz_t = \left(\mu - \frac{1}{2}\sigma^2\right)\,dt + \sigma\,dz_t.$$

Hence, the process $y_t = \ln X_t$ is a generalized Brownian motion. In particular, we have

$$y_{t'} - y_t = \left(\mu - \frac{1}{2}\sigma^2\right)(t'-t) + \sigma(z_{t'} - z_t),$$

which implies that

$$\ln X_{t'} = \ln X_t + \left(\mu - \frac{1}{2}\sigma^2\right)(t'-t) + \sigma(z_{t'} - z_t).$$

Taking exponentials on both sides, we get

$$X_{t'} = X_t \exp\left\{\left(\mu - \frac{1}{2}\sigma^2\right)(t'-t) + \sigma(z_{t'} - z_t)\right\}. \tag{2.15}$$

This is true for all $t' > t \geq 0$. In particular,

$$X_t = X_0 \exp\left\{\left(\mu - \frac{1}{2}\sigma^2\right)t + \sigma z_t\right\}.$$

Since exponentials are always positive, we see that X_t can only have positive values, so that the value space of a geometric Brownian motion is $\mathcal{S} = (0, \infty)$.

Suppose now that we stand at time t and have observed the current value X_t of a geometric Brownian motion. Which probability distribution is then appropriate for the uncertain future value, say at time t'? Since $z_{t'} - z_t \sim N(0, t'-t)$, we see from (2.15) that the future value $X_{t'}$ (conditional on X_t) will be lognormally distributed. The probability density function for $X_{t'}$ (given X_t) is given by

$$f(x) = \frac{1}{x\sqrt{2\pi\sigma^2(t'-t)}} \exp\left\{-\frac{1}{2\sigma^2(t'-t)}\left(\ln\left(\frac{x}{X_t}\right) - \left(\mu - \frac{1}{2}\sigma^2\right)(t'-t)\right)^2\right\}, \quad x > 0,$$

and the mean and variance are

$$\begin{aligned}
\mathrm{E}_t[X_{t'}] &= X_t e^{\mu(t'-t)}, \\
\mathrm{Var}_t[X_{t'}] &= X_t^2 e^{2\mu(t'-t)}\left[e^{\sigma^2(t'-t)} - 1\right],
\end{aligned}$$

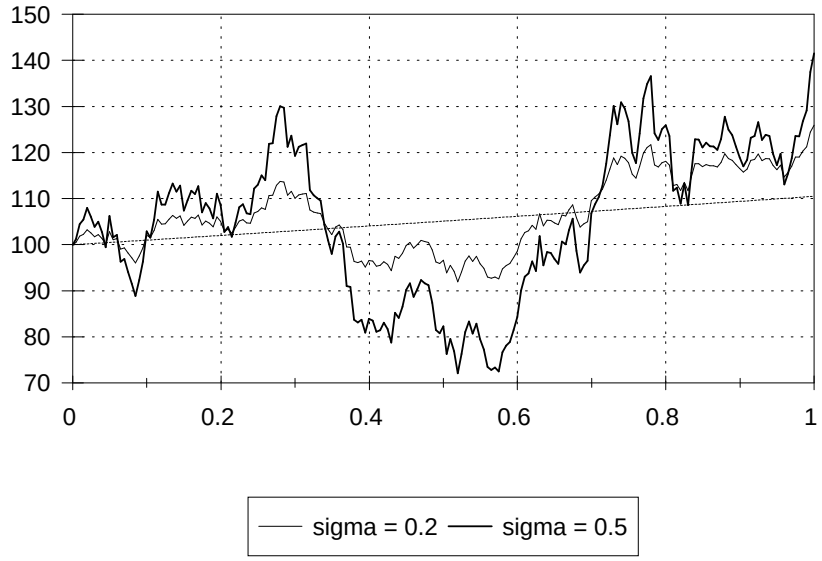


Figure 2.7: Simulation of a geometric Brownian motion with initial value $X_0 = 100$, relative drift rate $\mu = 0.1$, and a relative volatility of $\sigma = 0.2$ and $\sigma = 0.5$, respectively. The smooth curve shows the trend corresponding to $\sigma = 0$. The simulations are based on 200 subintervals of equal length, and the same sequence of random numbers has been used for the two σ-values.

cf. Appendix B.

Paths can be simulated by recursively computing either

$$X_{t_i} = X_{t_{i-1}} + \mu X_{t_{i-1}}(t_i - t_{i-1}) + \sigma X_{t_{i-1}} \varepsilon_i \sqrt{t_i - t_{i-1}}$$

or, more accurately,

$$X_{t_i} = X_{t_{i-1}} \exp\left\{\left(\mu - \frac{1}{2}\sigma^2\right)(t_i - t_{i-1}) + \sigma\varepsilon_i\sqrt{t_i - t_{i-1}}\right\}.$$

Figure 2.7 shows a single simulated sample path for $\sigma = 0.2$ and a sample path for $\sigma = 0.5$. For both sample paths we have used $\mu = 0.1$ and $X_0 = 100$, and the same sequence of random numbers.

We will consider other specific diffusions in later chapter, when we need them. For example, we shall use the Ornstein-Uhlenbeck process defined by

$$dX_t = \kappa\,(\theta - X_t)\;dt + \sigma\,dz_t,$$

which is the continuous-time equivalent of the discrete-time AR(1) process, and the square-root process defined by

$$dX_t = \kappa\,(\theta - X_t)\;dt + \sigma\sqrt{X_t}\,dz_t.$$

Such processes are used, among other things, to model the dynamics of interest rates.

2.7 Multi-dimensional processes

So far we have only considered one-dimensional processes, i.e. processes with a value space which is $\mathbb{R}$ or a subset of $\mathbb{R}$. In most asset pricing models we need to keep track of several processes, e.g. dividend and price processes for different assets, and we will often be interested in covariances and correlations between different processes.

If the exogenous shocks in the model are one-dimensional, then increments over the smallest time interval considered in the model will be perfectly correlated. In a discrete-time model where the exogenous shocks $\varepsilon_1, \ldots, \varepsilon_T$ are one-dimensional, changes in any two processes between two subsequent points in time, say t and $t+1$, will be perfectly correlated. For example, if X and Y are two general processes defined by

$$\Delta X_{t+1} = f_X(X_t) + g_X(X_t)\varepsilon_{t+1}$$

and

$$\Delta Y_{t+1} = f_Y(Y_t) + g_Y(Y_t)\varepsilon_{t+1}$$

for some nice functions f_X, g_X, f_Y, g_Y, then by using the properties of covariances and variances we get

$$\begin{gathered}\operatorname{Cov}_t[\Delta X_{t+1}, \Delta Y_{t+1}] = g_X(X_t)g_Y(Y_t)\operatorname{Var}_t[\varepsilon_{t+1}] = g_X(X_t)g_Y(Y_t),\\ \operatorname{Var}_t[\Delta X_{t+1}] = g_X(X_t)^2, \quad \operatorname{Var}_t[\Delta Y_{t+1}] = g_Y(Y_t)^2,\\ \operatorname{Corr}_t[\Delta X_{t+1}, \Delta Y_{t+1}] = 1.\end{gathered}$$

Increments in two processes over more than one sub-period are generally not perfectly correlated even with a one-dimensional shock.

In a continuous-time model where the exogenous shock process $z = (z_t)_{t\in[0,T]}$ is one-dimensional, the instantaneous increments of any two processes will be perfectly correlated. For example, if we consider the two Itô processes X and Y defined by

$$dX_t = \mu_{Xt}\,dt + \sigma_{Xt}\,dz_t, \qquad dY_t = \mu_{Yt}\,dt + \sigma_{Yt}\,dz_t,$$

then $\operatorname{Cov}_t[dX_t, dY_t] = \sigma_{Xt}\sigma_{Yt}\,dt$ so that the instantaneous correlation becomes

$$\operatorname{Corr}_t[dX_t, dY_t] = \frac{\operatorname{Cov}_t[dX_t, dY_t]}{\sqrt{\operatorname{Var}_t[dX_t]\operatorname{Var}_t[dY_t]}} = \frac{\sigma_{Xt}\sigma_{Yt}\,dt}{\sqrt{\sigma_{Xt}^2\,dt\;\sigma_{Yt}^2\,dt}} = 1.$$

Increments over any non-infinitesimal time interval are generally not perfectly correlated, i.e. for any $h > 0$ a correlation like $\operatorname{Corr}_t[X_{t+h} - X_t, Y_{t+h} - Y_t]$ is typically different from one but close to one for small h.

To obtain non-perfectly correlated changes over the shortest time period considered by the model we need an exogenous shock of a dimension higher than one, i.e. a shock vector. One can without loss of generality assume that the different components of this shock vector are mutually independent and generate non-perfect correlations between the relevant processes by varying the sensitivities of those processes towards the different exogenous shocks. In a discrete-time setting the exogenous shock is often assumed to be a multi-variate normally distributed random variable $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{Kt})^\top$ where ε_{it} and ε_{jt} for $i \neq j$ are independent one-dimensional normally distributed random variables (mean zero, unit variance). Analogously the exogenous shocks in continuous-time

settings are assumed to be generated by a multi-dimensional standard Brownian motion. Below we give a precise definition and briefly go through some other multi-dimensional continuous-time processes and a multi-dimensional version of Itô's Lemma.

A **K-dimensional standard Brownian motion** $\boldsymbol{z} = (z_1, \ldots, z_K)^\top$ is a stochastic process where the individual components z_i are mutually independent one-dimensional standard Brownian motions. If we let $\mathbf{0} = (0, \ldots, 0)^\top$ denote the zero vector in $\mathbb{R}^K$ and let $\underline{\underline{I}}$ denote the identity matrix of dimension $K \times K$ (the matrix with ones in the diagonal and zeros in all other entries), then we can write the defining properties of a K-dimensional Brownian motion $\boldsymbol{z}$ as follows:

(i) $\boldsymbol{z}_0 = \mathbf{0}$,

(ii) for all $t, t' \geq 0$ with $t < t'$: $\boldsymbol{z}_{t'} - \boldsymbol{z}_t \sim \boldsymbol{N}(\mathbf{0}, (t'-t)\underline{\underline{I}})$ [normally distributed increments],

(iii) for all $0 \leq t_0 < t_1 < \cdots < t_n$, the random variables $\boldsymbol{z}_{t_1} - \boldsymbol{z}_{t_0}$, …, $\boldsymbol{z}_{t_n} - \boldsymbol{z}_{t_{n-1}}$ are mutually independent [independent increments],

(iv) $\boldsymbol{z}$ has continuous sample paths in $\mathbb{R}^K$.

Here, $\boldsymbol{N}(\boldsymbol{a}, \underline{\underline{b}})$ denotes a K-dimensional normal distribution with mean vector $\boldsymbol{a}$ and variance-covariance matrix $\underline{\underline{b}}$. As for standard Brownian motions, we can also define multi-dimensional generalized Brownian motions, which simply are vectors of independent one-dimensional generalized Brownian motions.

A **K-dimensional diffusion process** $\boldsymbol{X} = (X_1, \ldots, X_K)^\top$ is a process with increments of the form

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, t)\, dt + \underline{\underline{\sigma}}(\boldsymbol{X}_t, t)\, d\boldsymbol{z}_t, \tag{2.16}$$

where $\boldsymbol{\mu}$ is a function from $\mathbb{R}^K \times \mathbb{R}_+$ into $\mathbb{R}^K$, and $\underline{\underline{\sigma}}$ is a function from $\mathbb{R}^K \times \mathbb{R}_+$ into the space of $K \times K$-matrices. As before, $\boldsymbol{z}$ is a K-dimensional standard Brownian motion. The evolution of the multi-dimensional diffusion can also be written componentwise as

$$\begin{aligned} dX_{it} &= \mu_i(\boldsymbol{X}_t, t)\, dt + \boldsymbol{\sigma}_i(X_t, t)^\top\, d\boldsymbol{z}_t \\ &= \mu_i(\boldsymbol{X}_t, t)\, dt + \sum_{k=1}^{K} \sigma_{ik}(\boldsymbol{X}_t, t)\, dz_{kt}, \quad i = 1, \ldots, K, \end{aligned} \tag{2.17}$$

where $\boldsymbol{\sigma}_i(\boldsymbol{X}_t, t)^\top$ is the i'th row of the matrix $\underline{\underline{\sigma}}(\boldsymbol{X}_t, t)$, and $\sigma_{ik}(\boldsymbol{X}_t, t)$ is the (i,k)'th entry (i.e. the entry in row i, column k). Since $dz_{1t}, \ldots, dz_{Kt}$ are mutually independent and all $N(0, dt)$ distributed, the expected change in the i'th component process over an infinitesimal period is

$$\mathrm{E}_t[dX_{it}] = \mu_i(\boldsymbol{X}_t, t)\, dt, \quad i = 1, \ldots, K,$$

so that μ_i can be interpreted as the drift of the i'th component. Furthermore, the covariance

between changes in the i'th and the j'th component processes over an infinitesimal period becomes

$$\begin{aligned}\operatorname{Cov}_t[dX_{it}, dX_{jt}] &= \operatorname{Cov}_t\left[\sum_{k=1}^{K}\sigma_{ik}(\boldsymbol{X}_t,t)\,dz_{kt}, \sum_{l=1}^{K}\sigma_{jl}(\boldsymbol{X}_t,t)\,dz_{lt}\right]\\ &= \sum_{k=1}^{K}\sum_{l=1}^{K}\sigma_{ik}(\boldsymbol{X}_t,t)\sigma_{jl}(\boldsymbol{X}_t,t)\operatorname{Cov}_t[dz_{kt}, dz_{lt}]\\ &= \sum_{k=1}^{K}\sigma_{ik}(\boldsymbol{X}_t,t)\sigma_{jk}(\boldsymbol{X}_t,t)\,dt\\ &= \boldsymbol{\sigma}_i(\boldsymbol{X}_t,t)^\top\boldsymbol{\sigma}_j(\boldsymbol{X}_t,t)\,dt, \quad i,j=1,\ldots,K,\end{aligned}$$

where we have applied the usual rules for covariances and the independence of the components of $\boldsymbol{z}$. In particular, the variance of the change in the i'th component process of an infinitesimal period is given by

$$\operatorname{Var}_t[dX_{it}] = \operatorname{Cov}_t[dX_{it}, dX_{it}] = \sum_{k=1}^{K}\sigma_{ik}(\boldsymbol{X}_t,t)^2\,dt = \|\boldsymbol{\sigma}_i(\boldsymbol{X}_t,t)\|^2\,dt, \quad i=1,\ldots,K.$$

The volatility of the i'th component is given by $\|\boldsymbol{\sigma}_i(\boldsymbol{X}_t,t)\|$. The variance-covariance matrix of changes of $\boldsymbol{X}_t$ over the next instant is $\underline{\underline{\Sigma}}(\boldsymbol{X}_t,t)\,dt = \underline{\underline{\sigma}}(\boldsymbol{X}_t,t)\underline{\underline{\sigma}}(\boldsymbol{X}_t,t)^\top\,dt$. The correlation between instantaneous increments in two component processes is

$$\operatorname{Corr}_t[dX_{it}, dX_{jt}] = \frac{\boldsymbol{\sigma}_i(\boldsymbol{X}_t,t)^\top\boldsymbol{\sigma}_j(\boldsymbol{X}_t,t)\,dt}{\sqrt{\|\boldsymbol{\sigma}_i(\boldsymbol{X}_t,t)\|^2\,dt\,\|\boldsymbol{\sigma}_j(\boldsymbol{X}_t,t)\|^2\,dt}} = \frac{\boldsymbol{\sigma}_i(\boldsymbol{X}_t,t)^\top\boldsymbol{\sigma}_j(\boldsymbol{X}_t,t)}{\|\boldsymbol{\sigma}_i(\boldsymbol{X}_t,t)\|\,\|\boldsymbol{\sigma}_j(\boldsymbol{X}_t,t)\|},$$

which can be any number in $[-1,1]$ depending on the elements of $\boldsymbol{\sigma}_i$ and $\boldsymbol{\sigma}_j$.

Similarly, we can define a **K-dimensional Itô process** $\boldsymbol{x} = (X_1,\ldots,X_K)^\top$ to be a process with increments of the form

$$d\boldsymbol{X}_t = \boldsymbol{\mu}_t\,dt + \underline{\underline{\sigma}}_t\,d\boldsymbol{z}_t, \tag{2.18}$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_t)$ is a K-dimensional stochastic process and $\underline{\underline{\sigma}} = (\underline{\underline{\sigma}}_t)$ is a stochastic process with values in the space of $K \times K$-matrices.

Next, we state a multi-dimensional version of Itô's Lemma, where a one-dimensional process is defined as a function of time and a multi-dimensional process.

Theorem 2.6 *Let $\boldsymbol{X} = (\boldsymbol{X}_t)_{t\geq 0}$ be an Itô process in $\mathbb{R}^K$ with dynamics $d\boldsymbol{X}_t = \boldsymbol{\mu}_t\,dt + \underline{\underline{\sigma}}_t\,d\boldsymbol{z}_t$ or, equivalently,*

$$dX_{it} = \mu_{it}\,dt + \boldsymbol{\sigma}_{it}^\top\,d\boldsymbol{z}_t = \mu_{it}\,dt + \sum_{k=1}^{K}\sigma_{ikt}\,dz_{kt}, \quad i=1,\ldots,K,$$

where $z_1,\ldots,z_K$ are independent standard Brownian motions, and μ_i and σ_{ik} are well-behaved stochastic processes.

Let $g(\boldsymbol{X},t)$ be a real-valued function for which all the derivatives $\frac{\partial g}{\partial t}$, $\frac{\partial g}{\partial X_i}$, and $\frac{\partial^2 g}{\partial X_i \partial X_j}$ exist and are continuous. Then the process $y = (y_t)_{t\geq 0}$ defined by $y_t = g(\boldsymbol{X}_t,t)$ is also an Itô process with dynamics

$$\begin{aligned}dy_t = &\left(\frac{\partial g}{\partial t}(\boldsymbol{X}_t,t) + \sum_{i=1}^{K}\frac{\partial g}{\partial X_i}(\boldsymbol{X}_t,t)\mu_{it} + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K}\frac{\partial^2 g}{\partial X_i \partial X_j}(\boldsymbol{X}_t,t)\gamma_{ijt}\right)dt\\ &+ \sum_{i=1}^{K}\frac{\partial g}{\partial X_i}(\boldsymbol{X}_t,t)\sigma_{i1t}\,dz_{1t} + \cdots + \sum_{i=1}^{K}\frac{\partial g}{\partial X_i}(\boldsymbol{X}_t,t)\sigma_{iKt}\,dz_{Kt},\end{aligned} \tag{2.19}$$

where $\gamma_{ij} = \sigma_{i1}\sigma_{j1} + \cdots + \sigma_{iK}\sigma_{jK}$ *is the covariance between the processes* X_i *and* X_j*.*

The result can also be written as

$$dy_t = \frac{\partial g}{\partial t}(\boldsymbol{X}_t, t)\, dt + \sum_{i=1}^{K} \frac{\partial g}{\partial X_i}(\boldsymbol{X}_t, t)\, dX_{it} + \frac{1}{2}\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{\partial^2 g}{\partial X_i \partial X_j}(\boldsymbol{X}_t, t)(dX_{it})(dX_{jt}), \tag{2.20}$$

where in the computation of $(dX_{it})(dX_{jt})$ one must use the rules $(dt)^2 = dt \cdot dz_{it} = 0$ for all i, $dz_{it} \cdot dz_{jt} = 0$ for $i \neq j$, and $(dz_{it})^2 = dt$ for all i. Alternatively, the result can be expressed using vector and matrix notation:

$$dy_t = \left(\frac{\partial g}{\partial t}(\boldsymbol{X}_t, t) + \left(\frac{\partial g}{\partial \boldsymbol{X}}(\boldsymbol{X}_t, t)\right)^\top \boldsymbol{\mu}_t + \frac{1}{2}\operatorname{tr}\left(\underline{\underline{\sigma}}_t^\top \left[\frac{\partial^2 g}{\partial \boldsymbol{X}^2}(\boldsymbol{X}_t, t)\right] \underline{\underline{\sigma}}_t\right)\right) dt + \left(\frac{\partial g}{\partial \boldsymbol{X}}(\boldsymbol{X}_t, t)\right)^\top \underline{\underline{\sigma}}_t\, d\boldsymbol{z}_t, \tag{2.21}$$

where

$$\frac{\partial g}{\partial \boldsymbol{X}}(\boldsymbol{X}_t, t) = \begin{pmatrix} \frac{\partial g}{\partial X_1}(\boldsymbol{X}_t, t) \\ \cdots \\ \frac{\partial g}{\partial X_K}(\boldsymbol{X}_t, t) \end{pmatrix}, \quad \frac{\partial^2 g}{\partial \boldsymbol{X}^2}(\boldsymbol{X}_t, t) = \begin{pmatrix} \frac{\partial^2 g}{\partial X_1^2}(\boldsymbol{X}_t, t) & \frac{\partial^2 g}{\partial X_1 \partial X_2}(\boldsymbol{X}_t, t) & \cdots & \frac{\partial^2 g}{\partial X_1 \partial X_K}(\boldsymbol{X}_t, t) \\ \frac{\partial^2 g}{\partial X_2 \partial X_1}(\boldsymbol{X}_t, t) & \frac{\partial^2 g}{\partial X_2^2}(\boldsymbol{X}_t, t) & \cdots & \frac{\partial^2 g}{\partial X_2 \partial X_K}(\boldsymbol{X}_t, t) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 g}{\partial X_K \partial X_1}(\boldsymbol{X}_t, t) & \frac{\partial^2 g}{\partial X_K \partial X_2}(\boldsymbol{X}_t, t) & \cdots & \frac{\partial^2 g}{\partial X_K^2}(\boldsymbol{X}_t, t) \end{pmatrix},$$

and tr denotes the trace of a quadratic matrix, i.e. the sum of the diagonal elements. For example, $\operatorname{tr}(\underline{\underline{A}}) = \sum_{i=1}^{K} A_{ii}$.

The probabilistic properties of a K-dimensional diffusion process is completely specified by the drift function $\boldsymbol{\mu}$ and the variance-covariance function $\underline{\underline{\Sigma}}$. The values of the variance-covariance function are symmetric and positive-definite matrices. Above we had $\underline{\underline{\Sigma}} = \underline{\underline{\sigma}}\,\underline{\underline{\sigma}}^\top$ for a general $(K \times K)$-matrix $\underline{\underline{\sigma}}$. But from linear algebra it is well-known that a symmetric and positive-definite matrix can be written as $\underline{\underline{\hat{\sigma}}}\,\underline{\underline{\hat{\sigma}}}^\top$ for a lower-triangular matrix $\underline{\underline{\hat{\sigma}}}$, i.e. a matrix with $\hat{\sigma}_{ik} = 0$ for $k > i$. This is the so-called Cholesky decomposition. Hence, we may write the dynamics as

$$\begin{aligned} dX_{1t} &= \mu_1(\boldsymbol{X}_t, t)\, dt + \hat{\sigma}_{11}(\boldsymbol{X}_t, t)\, dz_{1t} \\ dX_{2t} &= \mu_2(\boldsymbol{X}_t, t)\, dt + \hat{\sigma}_{21}(\boldsymbol{X}_t, t)\, dz_{1t} + \hat{\sigma}_{22}(\boldsymbol{X}_t, t)\, dz_{2t} \\ &\vdots \\ dX_{Kt} &= \mu_K(\boldsymbol{X}_t, t)\, dt + \hat{\sigma}_{K1}(\boldsymbol{X}_t, t)\, dz_{1t} + \hat{\sigma}_{K2}(\boldsymbol{X}_t, t)\, dz_{2t} + \cdots + \hat{\sigma}_{KK}(\boldsymbol{X}_t, t)\, dz_{Kt} \end{aligned} \tag{2.22}$$

We can think of building up the model by starting with X_1. The shocks to X_1 are represented by the standard Brownian motion z_1 and its coefficient $\hat{\sigma}_{11}$ is the volatility of X_1. Then we extend the model to include X_2. Unless the infinitesimal changes to X_1 and X_2 are always perfectly correlated we need to introduce another standard Brownian motion, z_2. The coefficient $\hat{\sigma}_{21}$ is fixed to match the covariance between changes to X_1 and X_2 and then $\hat{\sigma}_{22}$ can be chosen so that $\sqrt{\hat{\sigma}_{21}^2 + \hat{\sigma}_{22}^2}$ equals the volatility of X_2. The model may be extended to include additional processes in the same manner.

Some authors prefer to write the dynamics in an alternative way with a single standard Brownian

motion $\hat{z}_i$ for each component X_i such as

$$\begin{aligned} dX_{1t} &= \mu_1(\boldsymbol{X}_t, t)\, dt + V_1(\boldsymbol{X}_t, t)\, d\hat{z}_{1t} \\ dX_{2t} &= \mu_2(\boldsymbol{X}_t, t)\, dt + V_2(\boldsymbol{X}_t, t)\, d\hat{z}_{2t} \\ &\vdots \\ dX_{Kt} &= \mu_K(\boldsymbol{X}_t, t)\, dt + V_K(\boldsymbol{X}_t, t)\, d\hat{z}_{Kt} \end{aligned} \tag{2.23}$$

Clearly, the coefficient $V_i(\boldsymbol{X}_t, t)$ is then the volatility of X_i. To capture an instantaneous non-zero correlation between the different components the standard Brownian motions $\hat{z}_1, \ldots, \hat{z}_K$ have to be mutually correlated. Let ρ_{ij} be the correlation between $\hat{z}_i$ and $\hat{z}_j$. If (2.23) and (2.22) are meant to represent the same dynamics, we must have

$$\begin{aligned} V_i &= \sqrt{\hat{\sigma}_{i1}^2 + \cdots + \hat{\sigma}_{ii}^2}, \quad i = 1, \ldots, K, \\ \rho_{ii} = 1; \quad \rho_{ij} &= \frac{\sum_{k=1}^{i} \hat{\sigma}_{ik}\hat{\sigma}_{jk}}{V_i V_j},\ \rho_{ji} = \rho_{ij}, \quad i < j. \end{aligned}$$

2.8 Exercises

EXERCISE 2.1 In the two-period economy illustrated in Figures 2.1 and 2.2 consider an asset paying a dividend at time 2 given by

$$D_2 = \begin{cases} 0, & \text{for } \omega = 3, \\ 5, & \text{for } \omega \in \{1, 2, 4\}, \\ 10, & \text{for } \omega \in \{5, 6\}. \end{cases}$$

(a) What is the expectation at time 0 of D_2? What is the expectation at time 1 of D_2? Verify that the Law of Iterated Expectations holds for these expectations.

(b) What is the variance at time 0 of D_2? What is the variance at time 1 of D_2? Confirm that $\mathrm{Var}[D_2] = \mathrm{E}\left[\mathrm{Var}_1[D_2]\right] + \mathrm{Var}\left[\mathrm{E}_1[D_2]\right]$.

EXERCISE 2.2 Let $X = (X_t)$ and $Y = (Y_t)$ be the price processes of two assets with no intermediate dividends and assume that

$$\begin{aligned} dX_t &= X_t \left[0.05\, dt + 0.1\, dz_{1t} + 0.2\, dz_{2t}\right], \\ dY_t &= Y_t \left[0.07\, dt + 0.3\, dz_{1t} - 0.1\, dz_{2t}\right]. \end{aligned}$$

(a) What is the expected rate of return of each of the two assets?

(b) What is the return variance and volatility of each of the two assets?

(c) What is the covariance and the correlation between the returns on the two assets?

EXERCISE 2.3 Suppose $X = (X_t)$ is a geometric Brownian motion, $dX_t = \mu X_t\,dt + \sigma X_t\,dz_t$. What is the dynamics of the process $y = (y_t)$ defined by $y_t = (X_t)^n$? What can you say about the distribution of future values of the y process?

EXERCISE 2.4 Suppose that the continuous-time stochastic process $X = (X_t)$ is defined as

$$X_t = \frac{1}{2}\int_0^t \lambda_s^2\,ds + \int_0^t \lambda_s\,dz_s,$$

where $z = (z_t)$ is a one-dimensional standard Brownian motion and $\lambda = (\lambda_t)$ is some "nice" stochastic process.

(a) Argue that $dX_t = \frac{1}{2}\lambda_t^2\,dt + \lambda_t\,dz_t$.

(b) Suppose that the continuous-time stochastic process $\xi = (\xi_t)$ is defined as $\xi_t = \exp\{-X_t\}$. Show that $d\xi_t = -\lambda_t\xi_t\,dz_t$.

EXERCISE 2.5 (Adapted from Björk (2004).) Define the process $y = (y_t)$ by $y_t = z_t^4$, where $z = (z_t)$ is a standard Brownian motion. Find the dynamics of y. Show that

$$y_t = 6\int_0^t z_s^2\,ds + 4\int_0^t z_s^3\,dz_s.$$

Show that $\mathrm{E}[y_t] \equiv \mathrm{E}[z_t^4] = 3t^2$.

EXERCISE 2.6 (Adapted from Björk (2004).) Define the process $y = (y_t)$ by $y_t = e^{az_t}$, where a is a constant and $z = (z_t)$ is a standard Brownian motion. Find the dynamics of y. Show that

$$y_t = 1 + \frac{1}{2}a^2\int_0^t y_s\,ds + a\int_0^t y_s\,dz_s.$$

Define $m(t) = \mathrm{E}[y_t]$. Show that m satisfies the ordinary differential equation

$$m'(t) = \frac{1}{2}a^2 m(t), \quad m(0) = 1.$$

Show that $m(t) = e^{a^2t/2}$ and conclude that

$$\mathrm{E}\left[e^{az_t}\right] = e^{a^2t/2}.$$

EXERCISE 2.7 Consider the two general stochastic processes $X_1 = (X_{1t})$ and $X_2 = (X_{2t})$ defined by the dynamics

$$\begin{aligned} dX_{1t} &= \mu_{1t}\,dt + \sigma_{1t}\,dz_{1t}, \\ dX_{2t} &= \mu_{2t}\,dt + \rho_t\sigma_{2t}\,dz_{1t} + \sqrt{1-\rho_t^2}\sigma_{2t}\,dz_{2t}, \end{aligned}$$

where z_1 and z_2 are independent one-dimensional standard Brownian motions. Interpret μ_{it}, σ_{it}, and ρ_t. Define the processes $y = (y_t)$ and $w = (w_t)$ by $y_t = X_{1t}X_{2t}$ and $w_t = X_{1t}/X_{2t}$. What

is the dynamics of y and w? Concretize your answer for the special case where X_1 and X_2 are geometric Brownian motions with constant correlation, i.e. $\mu_{it} = \mu_i X_{it}$, $\sigma_{it} = \sigma_i X_{it}$, and $\rho_t = \rho$ with μ_i, σ_i, and ρ being constants.

Chapter 3

Assets, portfolios, and arbitrage

3.1 Introduction

This chapter shows how to model assets and portfolios of assets in one- and multi-period models with uncertainty. The important concepts of arbitrage, redundant assets, and market completeness are introduced.

3.2 Assets

An asset is characterized by its dividends and its price. We will always assume that the dividends of the basic assets are non-negative and that there is a positive probability of a positive dividend at the terminal date of the model. Then we can safely assume (since equilibrium prices will be arbitrage-free; see precise definition below) that the prices of the basic assets are always positive. We assume without loss of generality that assets pay no dividends at time 0. The price of an asset at a given point in time is exclusive any dividend payment at that time, i.e. prices are ex-dividend. At the last point in time considered in the model, all assets must have a zero price. We assume throughout that I basic assets are traded.

3.2.1 The one-period framework

In a one-period model any asset i is characterized by its time 0 price P_i and its time 1 dividend D_i, which is a random variable. If the realized state is $\omega \in \Omega$, asset i will give a dividend of $D_i(\omega)$. We can gather all the prices in the I-dimensional vector $\boldsymbol{P} = (P_1, \ldots, P_I)^\top$ and all the dividends in the I-dimensional random variable $\boldsymbol{D} = (D_1, \ldots, D_I)^\top$. We assume that all variances of dividends and all pairwise covariances of dividends are finite.

Instead of prices and dividends, we will often focus on returns. Returns can be defined in different ways. The **net return** on an investment over a period is simply the end-of-period outcome generated by the initial investment subtracted by the initial investment. In a one-period framework the net return on asset i is $D_i - P_i$. The **net rate of return** is the net return relative to the investment made, i.e. the net rate of return on asset i is $r_i = (D_i - P_i)/P_i$. The **gross rate of return** on asset i is defined as $R_i = D_i/P_i = 1 + r_i$, the ratio of the uncertain dividend to the price. The **log-return** or **continuously compounded rate of return** is defined as $\ln R_i = \ln(D_i/P_i)$.

Note that since the end-of-period dividend is a random variable, any of these returns will also be a random variable. In a one-period model an asset is said to be risk-free if it pays the same dividend in all states so that return (with any of the above definitions) will be non-random, i.e. known at the beginning of the period. We will write the risk-free gross return as R^f.

We can stack the returns on the different assets into vectors. For example, the gross rate of return vector is $\boldsymbol{R} = (R_1, \ldots, R_I)^\top$. Defining the $I \times I$ matrix

$$\operatorname{diag}(\boldsymbol{P}) = \begin{pmatrix} P_1 & 0 & \ldots & 0 \\ 0 & P_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & P_I \end{pmatrix}, \tag{3.1}$$

we can write the link between the dividend vector and the gross rate of return vector as

$$\boldsymbol{R} = [\operatorname{diag}(\boldsymbol{P})]^{-1} \boldsymbol{D} \quad \Leftrightarrow \quad \boldsymbol{D} = \operatorname{diag}(\boldsymbol{P})\boldsymbol{R}. \tag{3.2}$$

Given our assumptions about dividends, all the gross returns will have finite variances and all the pairwise covariances of gross returns will also be finite.

Note that given the expected dividend of asset i, $\mathrm{E}[D_i]$, finding the expected return $\mathrm{E}[R_i]$ is equivalent to finding the price P_i. We can therefore study equilibrium expected returns instead of equilibrium prices and many asset pricing models are typically formulated in terms of expected returns (as for example the classical CAPM).

3.2.2 The discrete-time framework

In a discrete-time model with $\mathcal{T} = \{0, 1, 2, \ldots, T\}$, we allow for dividends at all dates except time 0 so that the dividends of an asset are represented by an adapted, non-negative stochastic process $D_i = (D_{it})_{t \in \mathcal{T}}$ with initial value $D_{i0} = 0$. The random variable D_{it} represents the dividend payment of asset i at time t. We are interested in prices at all dates $t \in \mathcal{T}$ and let $P_i = (P_{it})_{t \in \mathcal{T}}$ denote the price process of asset i. By our assumptions, $P_{iT} = 0$ in all states. We assume that no matter what the information at time $T-1$ is, for any asset there will be a positive (conditional) probability that the terminal dividend is positive. We collect the prices and dividend processes of the I assets in I-dimensional processes $\boldsymbol{P} = (\boldsymbol{P}_t)_{t \in \mathcal{T}}$ and $\boldsymbol{D} = (\boldsymbol{D}_t)_{t \in \mathcal{T}}$ with $\boldsymbol{P}_t = (P_{1t}, \ldots, P_{It})^\top$ and similarly for $\boldsymbol{D}_t$.

The gross rate of return on asset i between two adjacent points in time, say t and time $t+1$, is defined as $R_{i,t+1} = (P_{i,t+1} + D_{i,t+1})/P_{it}$, the net rate of return is $r_{i,t+1} = R_{i,t+1} - 1$, and the log-return is $\ln R_{i,t+1}$. Note that now the relation between the expected gross rate of return $\mathrm{E}_t[R_{i,t+1}]$ and the beginning-of-period price P_{it} involves both the expected dividend $\mathrm{E}_t[D_{i,t+1}]$ and the expected future price $\mathrm{E}_t[P_{i,t+1}]$. Therefore we cannot easily switch between statements about expected returns and statements about prices. We will consider both formulations of asset pricing models and the link between them in later chapters.

We can also define returns over longer holding periods. Now, we could define the gross rate of return on asset i between time t and time $t+n$ as $R_{i,t,t+n} = (P_{i,t+n} + D_{i,t+1} + \cdots + D_{i,t+n})/P_{it}$ but, unless the intermediate dividends $D_{i,t+1}, \ldots, D_{i,t+n-1}$ are all zero, we will add values at different dates without any discounting. This is typically not very useful. More appropriately,

we can compute the return if intermediate dividends are reinvested in the asset. Assume we buy one unit of asset i at time t. For the dividend of $D_{i,t+1}$ received at time $t+1$, one can buy $D_{i,t+1}/P_{i,t+1}$ extra units of the asset so that the total holdings is $A_{t+1} \equiv 1 + D_{i,t+1}/P_{i,t+1}$ units. The total dividend received at time $t+2$ is then $(1 + D_{i,t+1}/P_{i,t+1})D_{i,t+2}$, which will buy you $(1 + D_{i,t+1}/P_{i,t+1})D_{i,t+2}/P_{i,t+2}$ additional units of the asset, bringing the total up to

$$A_{t+2} \equiv 1 + \frac{D_{i,t+1}}{P_{i,t+1}} + \left(1 + \frac{D_{i,t+1}}{P_{i,t+1}}\right)\frac{D_{i,t+2}}{P_{i,t+2}} = \left(1 + \frac{D_{i,t+1}}{P_{i,t+1}}\right)\left(1 + \frac{D_{i,t+2}}{P_{i,t+2}}\right)$$

units. Continuing like this we end up with

$$A_{t+n} \equiv \prod_{m=1}^{n}\left(1 + \frac{D_{i,t+m}}{P_{i,t+m}}\right) = \left(1 + \frac{D_{i,t+1}}{P_{i,t+1}}\right)\ldots\left(1 + \frac{D_{i,t+n}}{P_{i,t+n}}\right)$$

units at time $t+n$. The gross rate of return on asset i between time t and time $t+n$ is therefore

$$R_{i,t,t+n} = \frac{A_{t+n}P_{i,t+n}}{P_{it}} = \frac{P_{i,t+n}}{P_{it}}\prod_{m=1}^{n}\left(1 + \frac{D_{i,t+m}}{P_{i,t+m}}\right).$$

Again, we can define the corresponding net rate of return or log-return.

In the discrete-time framework an asset is said to be risk-free if the dividend at any time $t = 1,\ldots,T$ is already known at time $t-1$, no matter what information is available at time $t-1$. The risk-free gross return between time t and $t+1$ is denoted by R_t^f so that the subscript indicate the point in time at which the return will be known to investors, not the point in time at which the return is realized. Before time t, the risk-free return R_t^f for the period beginning at t is not necessarily known. Risk-free rates fluctuate over time so they are only risk-free in the short run. If you roll-over in one-period risk-free investments from time t to $t+n$, the total gross rate of return will be

$$R_{t,t+n}^f = \prod_{m=0}^{n-1} R_{t+m}^f.$$

Note that this return is risky seen at, or before, time t. An asset that provides a risk-free return from time t to time $t+n$ is a default-free, zero-coupon bond maturing at time $t+n$. If it has a face value of 1 and its time t price is denoted by B_t^{t+n}, you will get a gross rate of return of $1/B_t^{t+n}$.

3.2.3 The continuous-time framework

In a continuous-time model over the time span $\mathcal{T} = [0,T]$, the price of an asset i is represented by an adapted stochastic process $P_i = (P_{it})_{t\in[0,T]}$. In practice, no assets pay dividends continuously. However, for computational purposes it is sometimes useful to approximate a stream of frequent dividend payments by a continuous-time dividend process. On the other hand, a reasonable model should also allow for assets paying lump-sum dividends. We could capture both through a process $\mathcal{D}_i = (\mathcal{D}_{it})_{t\in[0,T]}$ where $\mathcal{D}_{it}$ is the (undiscounted) sum of all dividends of asset i up to and including time t. The total dividend received in a small interval $[t, t+dt]$ would then be $d\mathcal{D}_{it}$ and the dividend yield would be $d\mathcal{D}_{it}/P_{it}$. A lump-sum dividend at time t would correspond to a jump in $\mathcal{D}_{it}$. For notational simplicity we assume that only at time T the basic assets of the economy will pay a lump-sum dividend. The terminal dividend of asset i is modeled by a random variable D_{iT}, assumed non-negative with a positive probability of a positive value. Up to time T, dividends are paid out continuously. In general, the dividend yield can then be captured by a specification

like $d\mathcal{D}_{it}/P_{it} = \delta_{it}\,dt + \nu_{it}\,dz_{it}$ for some one- or multi-dimensional standard Brownian motion z_i, but for simplicity we assume that there is no uncertainty about the dividend yield over the next instant, i.e. $\nu_{it} = 0$. To sum up, the dividends of asset i are represented by a dividend yield process $\delta_i = (\delta_{it})_{t\in[0,T]}$ and a terminal lump-sum dividend D_i.

Next consider returns. If we buy one unit of asset i at time t and keep reinvesting the continuous dividends by purchasing extra (fractions of) units of the assets, how many units will we end up with at time $t' > t$? First divide the interval $[t, t']$ into many bits of length Δt so that dividends are cashed in and additional units bought at times $t + \Delta t, t + 2\Delta t, ...$ We can then proceed as in the discrete-time case discussed above. Let $A_{t+n\Delta t}$ denote the number of units of the asset we will have immediately after time $t + n\Delta t$. We start with $A_t = 1$ unit. At $t + \Delta t$ we receive a dividend of $\delta_{it}P_{i,t+\Delta t}\Delta t$ which we spend on buying $\delta_{it}\Delta t$ extra assets, bringing our holdings up to $A_{t+\Delta t} = 1 + \delta_{it}\Delta t$. At time $t + 2\Delta t$ we receive a total dividend of $A_{t+\Delta t}\delta_{i,t+\Delta t}P_{i,t+2\Delta t}\Delta t$, which will buy us $A_{t+\Delta t}\delta_{i,t+\Delta t}\Delta t$ extra units. Our total is now $A_{t+2\Delta t} = A_{t+\Delta t} + A_{t+\Delta t}\delta_{i,t+\Delta t}\Delta t$. Continuing like this we will find for any $s = t + n\Delta t$ for some integer $n < N$, our total holdings immediately after time $s + \Delta t$ is given by $A_{s+\Delta t} = A_s + A_s\delta_{is}\Delta t$ so that

$$\frac{A_{s+\Delta t} - A_s}{\Delta t} = \delta_{is}A_s.$$

If we go to the continuous-time limit and let $\Delta t \to 0$, the left-hand side will approach the derivative A'_s and we see that A_s must satisfy the differential equation $A'_s = \delta_{is}A_s$ as well as the initial condition $A_t = 1$. The solution is

$$A_s = \exp\left\{\int_t^s \delta_{iu}\,du\right\}.$$

So investing one unit in asset i at time t and continuously reinvesting the dividends, we will end up with $A_{t'} = \exp\{\int_t^{t'} \delta_{iu}\,du\}$ units of the asset at time t'. The gross rate of return on asset i over the interval $[t, t']$ is therefore $R_{i,t,t'} = \exp\{\int_t^{t'} \delta_{iu}\,du\}P_{i,t'}/P_{it}$.

The net rate of return per time period is $r_{i,t,t'} = (R_{i,t,t'} - 1)/(t' - t)$. Informally letting $t' \to t$, we get the instantaneous net rate of return per time period

$$r_{it} = \frac{1}{dt}\frac{dP_{it}}{P_{it}} + \delta_{it}.$$

The first term on the right-hand side is the instantaneous percentage capital gain (still unknown at time t), the second term is the dividend yield (assumed to be known at time t).

We will typically write the dynamics of the price of an asset i as

$$dP_{it} = P_{it}\left[\mu_{it}\,dt + \boldsymbol{\sigma}_{it}^\top\,d\boldsymbol{z}_t\right], \tag{3.3}$$

where $\boldsymbol{z} = (\boldsymbol{z}_t)$ is a standard Brownian motion of dimension d representing shocks to the prices, μ_{it} is then the expected capital gain so that the total expected net rate of return per time period is $\mu_{it} + \delta_{it}$, while $\boldsymbol{\sigma}_{it}$ is the vector of sensitivities of the price with respect to the exogenous shocks. The volatility is the standard deviation of instantaneous relative price changes, which is $\|\boldsymbol{\sigma}_{it}\| = \left(\sum_{j=1}^d \sigma_{ijt}^2\right)^{1/2}$. Using Itô's Lemma exactly as in Section 2.6.7, one finds that

$$P_{i,t'} = P_{it}\exp\left\{\int_t^{t'}\left(\mu_{iu} - \frac{1}{2}\|\boldsymbol{\sigma}_{iu}\|^2\right)du + \int_t^{t'}\boldsymbol{\sigma}_{iu}^\top\,d\boldsymbol{z}_u\right\}. \tag{3.4}$$

Therefore, the gross rate of return on asset i between t and t' is

$$R_{i,t,t'} = \exp\left\{\int_t^{t'} \delta_{iu}\, du\right\} \frac{P_{i,t'}}{P_{it}} = \exp\left\{\int_t^{t'} \left(\delta_{iu} + \mu_{iu} - \frac{1}{2}\|\boldsymbol{\sigma}_{iu}\|^2\right) du + \int_t^{t'} \boldsymbol{\sigma}_{iu}^\top\, d\boldsymbol{z}_u\right\}.$$

For a short period, i.e. with $t' = t + \Delta t$ for Δt small, we have

$$R_{i,t,t+\Delta t} \approx \exp\left\{\left(\delta_{it} + \mu_{it} - \frac{1}{2}\|\boldsymbol{\sigma}_{it}\|^2\right)\Delta t + \boldsymbol{\sigma}_{it}^\top\, \Delta\boldsymbol{z}_t\right\},$$

where $\Delta\boldsymbol{z}_t = \boldsymbol{z}_{t+\Delta t} - \boldsymbol{z}_t$ is d-dimensional and normally distributed with mean vector $\mathbf{0}$ and variance-covariance matrix $\underline{\underline{I}} \cdot \Delta t$ (here $\underline{\underline{I}}$ is the $d \times d$ identity matrix). The gross rate of return is thus approximately lognormally distributed with mean

$$\mathrm{E}_t\left[R_{i,t,t+\Delta t}\right] \approx \exp\left\{\left(\delta_{it} + \mu_{it} - \frac{1}{2}\|\boldsymbol{\sigma}_{it}\|^2\right)\Delta t\right\} \mathrm{E}_t\left[\exp\left\{\boldsymbol{\sigma}_{it}^\top\, \Delta\boldsymbol{z}_t\right\}\right] = e^{(\delta_{it}+\mu_{it})\Delta t},$$

cf. Appendix B.

The log-return is

$$\ln R_{i,t,t'} = \int_t^{t'} \left(\delta_{iu} + \mu_{iu} - \frac{1}{2}\|\boldsymbol{\sigma}_{iu}\|^2\right) du + \int_t^{t'} \boldsymbol{\sigma}_{iu}^\top\, d\boldsymbol{z}_u$$

with mean and variance given by

$$\begin{aligned}
\mathrm{E}_t[\ln R_{i,t,t'}] &= \int_t^{t'} \left(\delta_{iu} + \mu_{iu} - \frac{1}{2}\|\boldsymbol{\sigma}_{iu}\|^2\right) du, \\
\mathrm{Var}_t[\ln R_{i,t,t'}] &= \int_t^{t'} \mathrm{E}_t\left[\|\boldsymbol{\sigma}_{iu}\|^2\right] du,
\end{aligned}$$

according to Theorem 2.2.

We can write the dynamics of all I prices compactly as

$$d\boldsymbol{P}_t = \mathrm{diag}(\boldsymbol{P}_t)\left[\boldsymbol{\mu}_t\, dt + \underline{\underline{\sigma}}_t\, d\boldsymbol{z}_t\right], \tag{3.5}$$

where $\mathrm{diag}(\boldsymbol{P}_t)$ is defined in (3.1), $\boldsymbol{\mu}_t$ is the vector $(\mu_{1t}, \ldots, \mu_{It})^\top$ and $\underline{\underline{\sigma}}_t$ is the $I \times d$ matrix whose i'th row is $\boldsymbol{\sigma}_{it}^\top$.

A risk-free asset is an asset where the rate of return over the next instant is always known. We can think of this as an asset with a constant price and a continuous dividend yield process $r^f = (r_t^f)_{t\in[0,T]}$ *or* as an asset with a zero continuous dividend yield and a price process accumulating the interest rate payments, $P_t^f = \exp\left\{\int_0^t r_s^f\, ds\right\}$.

3.3 Portfolios and trading strategies

Individuals can trade assets at all time points of the model, except for the last date. The combination of holdings of different assets at a given point in time is called a portfolio. We assume that there are no restrictions on the portfolios that investors may form and that there are no trading costs.

3.3.1 The one-period framework

In a one-period model an individual chooses a portfolio $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)^\top$ at time 0 with θ_i being the number of units held of asset i. It is not possible to rebalance the portfolio but the individual simply cashes in the dividends at time 1. Let $D^{\boldsymbol{\theta}}$ be the random variable that represents the dividend of the portfolio $\boldsymbol{\theta}$. If state ω is realized, the total dividend from a portfolio $\boldsymbol{\theta}$ is

$$D^{\boldsymbol{\theta}}(\omega) = \sum_{i=1}^{I} \theta_i D_i(\omega) = \boldsymbol{\theta} \cdot \boldsymbol{D}(\omega),$$

i.e. $D^{\boldsymbol{\theta}} = \boldsymbol{\theta} \cdot \boldsymbol{D}$.

Denote the price or value of a portfolio $\boldsymbol{\theta}$ by $P^{\boldsymbol{\theta}}$. We will throughout this book assume that prices are linear so that

$$P^{\boldsymbol{\theta}} = \sum_{i=1}^{I} \theta_i P_i = \boldsymbol{\theta} \cdot \boldsymbol{P}.$$

This is called **the Law of One Price**. Since we ignore transaction costs, any candidate for an equilibrium pricing system will certainly have this property. In Section 3.4 we will discuss the link between the Law of One Price and the absence of arbitrage.

The fraction of the total portfolio value invested in asset i is then $\pi_i = \theta_i P_i / P^{\boldsymbol{\theta}}$ and the vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_I)^\top$ is called the portfolio weight vector. If we let $\mathbf{1} = (1, \ldots, 1)^\top$, we have $\boldsymbol{\pi} \cdot \mathbf{1} = \sum_{i=1}^{I} \pi_i = 1$. Note that $\operatorname{diag}(\boldsymbol{P})\,\mathbf{1} = \boldsymbol{P}$ and thus $P^{\boldsymbol{\theta}} = \boldsymbol{P}^\top \boldsymbol{\theta} = (\operatorname{diag}(\boldsymbol{P})\,\mathbf{1})^\top \boldsymbol{\theta} = \mathbf{1}^\top \operatorname{diag}(\boldsymbol{P})\boldsymbol{\theta}$ so that

$$\boldsymbol{\pi} = \frac{\operatorname{diag}(\boldsymbol{P})\boldsymbol{\theta}}{\boldsymbol{P}^\top \boldsymbol{\theta}} = \frac{\operatorname{diag}(\boldsymbol{P})\boldsymbol{\theta}}{\mathbf{1}^\top \operatorname{diag}(\boldsymbol{P})\boldsymbol{\theta}}. \tag{3.6}$$

Given $\boldsymbol{\theta}$ and the price vector $\boldsymbol{P}$ we can derive $\boldsymbol{\pi}$. Conversely, given $\boldsymbol{\pi}$, the total portfolio value $P^{\boldsymbol{\theta}}$, and the price vector $\boldsymbol{P}$, we can derive $\boldsymbol{\theta}$. We therefore have two equivalent ways of representing a portfolio.

The gross rate of return on a portfolio $\boldsymbol{\theta}$ is the random variable

$$R^{\boldsymbol{\theta}} = \frac{D^{\boldsymbol{\theta}}}{P^{\boldsymbol{\theta}}} = \frac{\sum_{i=1}^{I} \theta_i D_i}{P^{\boldsymbol{\theta}}} = \frac{\sum_{i=1}^{I} \theta_i P_i R_i}{P^{\boldsymbol{\theta}}} = \sum_{i=1}^{I} \frac{\theta_i P_i}{P^{\boldsymbol{\theta}}} R_i = \sum_{i=1}^{I} \pi_i R_i = \boldsymbol{\pi} \cdot \boldsymbol{R}, \tag{3.7}$$

where π_i is the portfolio weight of asset i. We observe that the gross return on a portfolio is just a weighted average of the gross rates of return on the assets in the portfolio. Similarly for the net rate of return since $\sum_{i=1}^{I} \pi_i = 1$ and thus

$$r^{\boldsymbol{\theta}} = R^{\boldsymbol{\theta}} - 1 = \left(\sum_{i=1}^{I} \pi_i R_i \right) - 1 = \sum_{i=1}^{I} \pi_i \left(R_i - 1 \right) = \sum_{i=1}^{I} \pi_i r_i = \boldsymbol{\pi} \cdot \boldsymbol{r},$$

where $\boldsymbol{r} = (r_1, \ldots, r_I)^\top$ is the vector of net rates of return on the basic assets.

3.3.2 The discrete-time framework

In a multi-period model individuals are allowed to rebalance their portfolio at any date considered in the model. A **trading strategy** is an I-dimensional adapted stochastic process $\boldsymbol{\theta} = (\boldsymbol{\theta}_t)_{t \in \mathcal{T}}$ where $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{It})^\top$ denotes the portfolio held at time t or rather immediately after trading at time t. θ_{it} is the number of units of asset i held at time t.

A trading strategy $\boldsymbol{\theta}$ generates a dividend process $D^{\boldsymbol{\theta}}$. Immediately before time t the portfolio is given by $\boldsymbol{\theta}_{t-1}$ so the investor will receive dividends $\boldsymbol{\theta}_{t-1} \cdot \boldsymbol{D}_t$ at time t, and then rebalance the portfolio to $\boldsymbol{\theta}_t$ immediately after time t. The net gain or dividend at time t is therefore equal to

$$D_t^{\boldsymbol{\theta}} = \boldsymbol{\theta}_{t-1} \cdot \boldsymbol{D}_t - (\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \cdot \boldsymbol{P}_t = \boldsymbol{\theta}_{t-1} \cdot (\boldsymbol{P}_t + \boldsymbol{D}_t) - \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t, \quad t = 1, 2, \ldots, T-1. \tag{3.8}$$

We can think of this as a budget constraint saying that the sum of the withdrawn dividend and our additional investment $(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \cdot \boldsymbol{P}_t$ has to equal the dividends we receive from the current portfolio. The terminal dividend is

$$D_T^{\boldsymbol{\theta}} = \boldsymbol{\theta}_{T-1} \cdot \boldsymbol{D}_T. \tag{3.9}$$

Given the Law of One Price, the initial price of the trading strategy is $P^{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 \cdot \boldsymbol{P}_0$. We can let $D_0^{\boldsymbol{\theta}} = -P^{\boldsymbol{\theta}}$ so that the dividend process $D^{\boldsymbol{\theta}}$ is defined at all $t \in \mathcal{T}$.

For $t = 1, \ldots, T$ define

$$V_t^{\boldsymbol{\theta}} = \boldsymbol{\theta}_{t-1} \cdot (\boldsymbol{P}_t + \boldsymbol{D}_t)$$

which is the time t value of the portfolio chosen at the previous trading date. This is the value of the portfolio just after dividends are received at time t and before the portfolio is rebalanced. Define $V_0^{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 \cdot \boldsymbol{P}_0$. We call $V^{\boldsymbol{\theta}} = (V_t^{\boldsymbol{\theta}})_{t \in \mathcal{T}}$ the **value process** of the trading strategy $\boldsymbol{\theta}$. According to (3.8), we have $V_t^{\boldsymbol{\theta}} = D_t^{\boldsymbol{\theta}} + \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t$ for $t = 1, \ldots, T$ and, in particular, $V_T^{\boldsymbol{\theta}} = D_T^{\boldsymbol{\theta}}$. The change in the value of the trading strategy between two adjacent dates is

$$\begin{aligned} V_{t+1}^{\boldsymbol{\theta}} - V_t^{\boldsymbol{\theta}} &= \boldsymbol{\theta}_t \cdot (\boldsymbol{P}_{t+1} + \boldsymbol{D}_{t+1}) - \boldsymbol{\theta}_{t-1} \cdot (\boldsymbol{P}_t + \boldsymbol{D}_t) \\ &= \boldsymbol{\theta}_t \cdot (\boldsymbol{P}_{t+1} + \boldsymbol{D}_{t+1}) - D_t^{\boldsymbol{\theta}} - \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t \\ &= \boldsymbol{\theta}_t \cdot (\boldsymbol{P}_{t+1} - \boldsymbol{P}_t + \boldsymbol{D}_{t+1}) - D_t^{\boldsymbol{\theta}}. \end{aligned} \tag{3.10}$$

The first term on the right-hand side of the last expression is the net return on the portfolio $\boldsymbol{\theta}_t$ from time t to $t+1$, the latter term is the net dividend we have withdrawn at time t.

The trading strategy is said to be **self-financing** if all the intermediate dividends are zero, i.e. if $D_t^{\boldsymbol{\theta}} = 0$ for $t = 1, \ldots, T-1$. Using (3.8) this means that

$$(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \cdot \boldsymbol{P}_t = \boldsymbol{\theta}_{t-1} \cdot \boldsymbol{D}_t, \quad t = 1, \ldots, T-1.$$

The left-hand side is the extra investment due to the rebalancing at time t, the right-hand side is the dividend received at time t. A self-financing trading strategy requires an initial investment of $P^{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 \cdot \boldsymbol{P}_0$ and generates a terminal dividend of $D_T^{\boldsymbol{\theta}} = \boldsymbol{\theta}_{T-1} \cdot \boldsymbol{D}_T$. At the intermediate dates no money is invested or withdrawn so increasing the investment in some assets must be fully financed by dividends or selling off other assets. If $\boldsymbol{\theta}$ is self-financing, we have $V_t^{\boldsymbol{\theta}} = P_t^{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t$, the time t price of the portfolio $\boldsymbol{\theta}_t$, for $t = 0, 1, \ldots, T-1$. Moreover, the change in the value of the trading strategy is just the net return, cf. (3.10).

Portfolio weights...

Returns...

3.3.3 The continuous-time framework

Also in the continuous-time framework with $\mathcal{T} = [0, T]$ a trading strategy is an I-dimensional adapted stochastic process $\boldsymbol{\theta} = (\boldsymbol{\theta}_t)_{t \in \mathcal{T}}$ where $\boldsymbol{\theta}_t = (\theta_{1t}, \ldots, \theta_{It})^\top$ denotes the portfolio held at time t or rather immediately after trading at time t.

Consistent with the discrete-time framework we define the value of the trading strategy $\boldsymbol{\theta}$ at any given time t as the price of the portfolio just chosen at that time plus any lump-sum dividends received at that time. Since we only allow for lump-sum dividends at the terminal date, we define

$$V_t^{\boldsymbol{\theta}} = \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t, \quad t < T \tag{3.11}$$

and $V_T^{\boldsymbol{\theta}} = \boldsymbol{\theta}_T \cdot \boldsymbol{D}_T \equiv D_T^{\boldsymbol{\theta}}$, the terminal lump-sum dividend. The time 0 value is the cost of initiating the trading strategy, which we can think of as a negative initial dividend, $D_0^{\boldsymbol{\theta}} = -V_0^{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 \cdot \boldsymbol{P}_0$. We assume that between time 0 and time T no lump-sum dividends can be withdrawn from the investment but funds can be withdrawn at a continuous rate as represented by the process $\alpha^{\boldsymbol{\theta}} = (\alpha_t^{\boldsymbol{\theta}})$ describing the rate with which we withdraw funds from our investments. Intermediate lump-sum withdrawals could be allowed at the expense of additional notational complexity. For later use note that an application of Itô's Lemma implies that the increment to the value process is given by

$$dV_t^{\boldsymbol{\theta}} = \boldsymbol{\theta}_t \cdot d\boldsymbol{P}_t + d\boldsymbol{\theta}_t \cdot \boldsymbol{P}_t + d\boldsymbol{\theta}_t \cdot d\boldsymbol{P}_t. \tag{3.12}$$

Assume for a moment that we do not change our portfolio over a small time interval $[t, t + \Delta t]$. The total funds withdrawn over this interval is $\alpha_t^{\boldsymbol{\theta}} \Delta t$. Let $\Delta\boldsymbol{\theta}_{t+\Delta t} = \boldsymbol{\theta}_{t+\Delta t} - \boldsymbol{\theta}_t$ and $\Delta\boldsymbol{P}_{t+\Delta t} = \boldsymbol{P}_{t+\Delta t} - \boldsymbol{P}_t$. The total dividends received from the portfolio $\boldsymbol{\theta}_t$ over the interval is $\sum_{i=1}^{I} \theta_{it}\delta_{it}P_{it}\Delta t$, which we can rewrite as $\boldsymbol{\theta}_t^\top \operatorname{diag}(\boldsymbol{P}_t)\boldsymbol{\delta}_t \Delta t$, where $\operatorname{diag}(\boldsymbol{P}_t)$ is the matrix defined in (3.1). Then the budget constraint over this interval is

$$\alpha_t^{\boldsymbol{\theta}} \Delta t + \Delta\boldsymbol{\theta}_{t+\Delta t} \cdot \boldsymbol{P}_{t+\Delta t} = \boldsymbol{\theta}_t^\top \operatorname{diag}(\boldsymbol{P}_t)\boldsymbol{\delta}_t \Delta t.$$

The left-hand side is the sum of the funds we withdraw and the net extra investment. The right-hand side is the funds we receive in dividends. Let us add and subtract $(\Delta\boldsymbol{\theta}_{t+\Delta t}) \cdot \boldsymbol{P}_t$ in the above equation. Rearranging we obtain

$$\alpha_t^{\boldsymbol{\theta}} \Delta t + \Delta\boldsymbol{\theta}_{t+\Delta t} \cdot \Delta\boldsymbol{P}_{t+\Delta t} + \Delta\boldsymbol{\theta}_{t+\Delta t} \cdot \boldsymbol{P}_t = \boldsymbol{\theta}_t^\top \operatorname{diag}(\boldsymbol{P}_t)\boldsymbol{\delta}_t \Delta t.$$

The equivalent equation for an infinitesimal interval $[t, t + dt]$ is

$$\alpha_t^{\boldsymbol{\theta}}\, dt + d\boldsymbol{\theta}_t \cdot d\boldsymbol{P}_t + d\boldsymbol{\theta}_t \cdot \boldsymbol{P}_t = \boldsymbol{\theta}_t^\top \operatorname{diag}(\boldsymbol{P}_t)\boldsymbol{\delta}_t\, dt.$$

Using this, we can rewrite the value dynamics in (3.12) as

$$dV_t^{\boldsymbol{\theta}} = \boldsymbol{\theta}_t \cdot d\boldsymbol{P}_t + \boldsymbol{\theta}_t^\top \operatorname{diag}(\boldsymbol{P}_t)\boldsymbol{\delta}_t\, dt - \alpha_t^{\boldsymbol{\theta}}\, dt.$$

Substituting in (3.5), this implies that

$$dV_t^{\boldsymbol{\theta}} = \boldsymbol{\theta}_t^\top \operatorname{diag}(\boldsymbol{P}_t) \left[(\boldsymbol{\mu}_t + \boldsymbol{\delta}_t)\, dt + \underline{\underline{\sigma}}_t\, d\boldsymbol{z}_t\right] - \alpha_t^{\boldsymbol{\theta}}\, dt. \tag{3.13}$$

As discussed earlier we can define a portfolio weight vector

$$\boldsymbol{\pi}_t = \frac{\operatorname{diag}(\boldsymbol{P}_t)\boldsymbol{\theta}_t}{\boldsymbol{P}_t^\top \boldsymbol{\theta}_t} = \frac{\operatorname{diag}(\boldsymbol{P}_t)\boldsymbol{\theta}_t}{V_t^{\boldsymbol{\theta}}}.$$

The value dynamics can therefore be rewritten as

$$dV_t^{\boldsymbol{\theta}} = V_t^{\boldsymbol{\theta}} \boldsymbol{\pi}_t^\top \left[(\boldsymbol{\mu}_t + \boldsymbol{\delta}_t)\, dt + \underline{\underline{\sigma}}_t\, d\boldsymbol{z}_t\right] - \alpha_t^{\boldsymbol{\theta}}\, dt. \tag{3.14}$$

A trading strategy is called **self-financing** if no funds are withdrawn, i.e. $\alpha_t^{\boldsymbol{\theta}} \equiv 0$. In that case the value dynamics is simply

$$dV_t^{\boldsymbol{\theta}} = \boldsymbol{\theta}_t^\top \operatorname{diag}(\boldsymbol{P}_t) \left[(\boldsymbol{\mu}_t + \boldsymbol{\delta}_t)\, dt + \underline{\underline{\sigma}}_t\, d\boldsymbol{z}_t \right]. \tag{3.15}$$

This really means that for any $t \in (0, T)$,

$$\begin{aligned} V_t^{\boldsymbol{\theta}} &= \boldsymbol{\theta}_0 \cdot \boldsymbol{P}_0 + \int_0^t \boldsymbol{\theta}_s^\top \operatorname{diag}(\boldsymbol{P}_s) \left[(\boldsymbol{\mu}_s + \boldsymbol{\delta}_s)\, ds + \underline{\underline{\sigma}}_s\, d\boldsymbol{z}_s \right] \\ &= \boldsymbol{\theta}_0 \cdot \boldsymbol{P}_0 + \int_0^t \boldsymbol{\theta}_s^\top \operatorname{diag}(\boldsymbol{P}_s) (\boldsymbol{\mu}_s + \boldsymbol{\delta}_s)\, ds + \int_0^t \boldsymbol{\theta}_s^\top \operatorname{diag}(\boldsymbol{P}_s) \underline{\underline{\sigma}}_s\, d\boldsymbol{z}_s. \end{aligned} \tag{3.16}$$

Returns...

3.4 Arbitrage

We have already made an assumption about prices, namely that prices obey the Law of One Price, i.e. prices are linear. We will now make the slightly stronger assumption that prices are set so that there is no arbitrage. An arbitrage is basically a risk-free profit.

3.4.1 The one-period framework

In the one-period framework we define an arbitrage as a portfolio $\boldsymbol{\theta}$ satisfying one of the following two conditions:

(i) $P^{\boldsymbol{\theta}} < 0$ and $D^{\boldsymbol{\theta}} \geq 0$;

(ii) $P^{\boldsymbol{\theta}} \leq 0$ and $D^{\boldsymbol{\theta}} \geq 0$ with $\mathbb{P}\left(D^{\boldsymbol{\theta}} > 0\right) > 0$.

Here $D^{\boldsymbol{\theta}}$ is the random variable that represents the dividend of the portfolio $\boldsymbol{\theta}$. The inequality $D^{\boldsymbol{\theta}} \geq 0$ means that the dividend will be non-negative no matter which state is realized, i.e. $D^{\boldsymbol{\theta}}(\omega) \geq 0$ for all $\omega \in \Omega$. (In a finite-state economy this can be replaced by the condition $\boldsymbol{D}^{\boldsymbol{\theta}} \geq 0$ on the dividend vector, which means that all elements of the vector are non-negative. The condition $\mathbb{P}\left(D^{\boldsymbol{\theta}} > 0\right) > 0$ can be replaced by the condition $\boldsymbol{D}_\omega^{\boldsymbol{\theta}} > 0$ for some state ω.)

An arbitrage offers something for nothing. It offers a non-negative dividend no matter which state is realized and its price is non-positive so that you do not have to pay anything. Either you get something today (case (i)) or you get something at the end in some state (case (ii)). This is clearly attractive to any greedy individual, i.e. any individual preferring more to less. Therefore, a market with arbitrage cannot be a market in equilibrium. Since we are interested in equilibrium pricing systems, we need only to care about pricing systems that do not admit arbitrage.

Absence of arbitrage implies that the law of one price holds. To see this, first suppose that $P^{\boldsymbol{\theta}} < \boldsymbol{\theta} \cdot \boldsymbol{P}$. Then an arbitrage can be formed by purchasing the portfolio $\boldsymbol{\theta}$ for the price of $P^{\boldsymbol{\theta}}$ and, for each $i = 1, \ldots, I$ selling θ_i units of asset i at a unit price of P_i. The end-of-period net dividend from this position will be zero no matter which state is realized. The total initial price of the position is $P^{\boldsymbol{\theta}} - \boldsymbol{\theta} \cdot \boldsymbol{P}$, which is negative. Hence, in the absence of arbitrage, we cannot have that $P^{\boldsymbol{\theta}} < \boldsymbol{\theta} \cdot \boldsymbol{P}$. The inequality $P^{\boldsymbol{\theta}} > \boldsymbol{\theta} \cdot \boldsymbol{P}$ can be ruled out by a similar argument.

On the other hand, the law of one price does not rule out arbitrage. For example, suppose that there are two possible states. Asset 1 gives a dividend of 0 in state 1 and a dividend of 1 in state 2

and costs 0.9. Asset 2 gives a dividend of 1 in state 1 and a dividend of 2 in state 2 and costs 1.6. Suppose the law of one price holds so that the price of any portfolio $(\theta_1, \theta_2)^\top$ is $0.9\theta_1 + 1.6\theta_2$. Consider the portfolio $(-2, 1)^\top$ i.e. a short position in two units of asset 1 and a long position of on unite of asset 2. The dividend of this portfolio will be 1 in state 1 and 0 in state 2 and the price is $0.9 \cdot (-2) + 1.6 \cdot 1 = -0.2$. This portfolio is clearly an arbitrage.

3.4.2 The discrete-time and continuous-time frameworks

In both the discrete-time and the continuous-time framework we define an arbitrage to be a self-financing trading strategy $\boldsymbol{\theta}$ satisfying one of the following two conditions:

(i) $V_0^{\boldsymbol{\theta}} < 0$ and $V_T^{\boldsymbol{\theta}} \geq 0$ with probability one,

(ii) $V_0^{\boldsymbol{\theta}} \leq 0$, $V_T^{\boldsymbol{\theta}} \geq 0$ with probability one, and $V_T^{\boldsymbol{\theta}} > 0$ with strictly positive probability.

As we have seen above, $V_T^{\boldsymbol{\theta}} = D_T^{\boldsymbol{\theta}}$ and $V_0^{\boldsymbol{\theta}} = -D_0^{\boldsymbol{\theta}}$ for self-financing trading strategies. We can therefore equivalently define an arbitrage in terms of dividends. A self-financing trading strategy is an arbitrage if it generates non-negative initial and terminal dividends with one of them being strictly positive with a strictly positive probability. Due to our assumptions on the dividends of the individual assets, the absence of arbitrage will imply that the prices of individual assets are strictly positive.

Ruling out arbitrages defined in (i) and (ii) will also rule out shorter term risk-free gains. Suppose for example that we can construct a trading strategy with a non-positive initial value (i.e. a non-positive price), always non-negative values, and a strictly positive value at some time $t < T$. Then this strictly positive value can be invested in any asset until time T generating a strictly positive terminal value with a strictly positive probability. The focus on self-financing trading strategies is therefore no restriction. Note that the definition of an arbitrage implies that a self-financing trading strategy with a terminal dividend of zero (in any state) must have a value process identically equal to zero.

3.4.3 Continuous-time doubling strategies

In a continuous-time setting it is theoretically possible to construct some strategies that generate something for nothing. These are the so-called doubling strategies, which were apparently first mentioned in a finance setting by Harrison and Kreps (1979). Think of a series of coin tosses numbered $n = 1, 2, \ldots$. The n'th coin toss takes place at time $1 - 1/n \in [0, 1)$. In the n'th toss, you get $\alpha 2^{n-1}$ if heads comes up, and looses $\alpha 2^{n-1}$ otherwise, where α is some positive number. You stop betting the first time heads comes up. Suppose heads comes up the first time in toss number $(k+1)$. Then in the first k tosses you have lost a total of $\alpha(1+2+\cdots+2^{k-1}) = \alpha(2^k-1)$. Since you win $\alpha 2^k$ in toss number $k+1$, your total profit will be $\alpha 2^k - \alpha(2^k - 1) = \alpha$. Since the probability that heads comes up eventually is equal to one, you will gain α with probability one. The gain is obtained before time 1 and can be made as large as possible by increasing α.

Similar strategies, with future dividends appropriately discounted, can be constructed in continuous-time models of financial markets—at least if a risk-free asset is traded—but are clearly impossible to implement in real life. As shown by Dybvig and Huang (1988), doubling strategies can be ruled out by requiring that trading strategies have values that are bounded from below, i.e. that some

constant K exists such that $V_t^{\boldsymbol{\theta}} \geq -K$ for all t. A trading strategy satisfying such a condition is said to be credit-constrained. A lower bound is reasonable since nobody can borrow an infinite amount of money. If you have a limited borrowing potential, the doubling strategy described above cannot be implemented. If you have no future income at all, $K = 0$ seems reasonable. An alternative way of eliminating doubling strategies is to impose the condition that the value process of the trading strategy has finite variance, cf. Duffie (2001). For a doubling strategy the variance of the value process is in fact infinite.

It seems evident that any greedy investor would implement a doubling strategy, if possible, since the investor will make a positive net return with a probability of one in finite time. However, Omberg (1989) shows that a doubling strategy may in fact generate an expected utility of minus infinity for risk-averse investors. In some events of zero probability a doubling strategy may result in outcomes associated with a utility of minus infinity. When multiplying zero and minus infinity in order to compute the expected utility, the result is indeterminate. Omberg computes the actual expected utility of the doubling strategy by taking an appropriate limit and finds that this is minus infinity for commonly used utility functions that are unbounded from below. Although this questions the above definition of an arbitrage, we will stick to that definition which is also the standard of the literature.

In the rest of this book we will—often implicitly—assume that some conditions are imposed so that doubling strategies are not implementable or that nobody wants to implement them.

3.5 Redundant assets

An asset is said to be redundant if its dividends can be replicated by a trading strategy in other assets.

For example, in the one-period framework asset i is redundant if a portfolio $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)^\top$ exists with $\theta_i = 0$ and

$$D_i = D^{\boldsymbol{\theta}} \equiv \theta_1 D_1 + \cdots + \theta_{i-1} D_{i-1} + \theta_{i+1} D_{i+1} + \cdots + \theta_I D_I.$$

Recall that the dividends are random variables so the above equation really means that

$$D_i(\omega) = \theta_1 D_1(\omega) + \cdots + \theta_{i-1} D_{i-1}(\omega) + \theta_{i+1} D_{i+1}(\omega) + \cdots + \theta_I D_I(\omega), \quad \forall \omega \in \Omega.$$

Such a portfolio is called a replicating portfolio for asset i.

If an asset i is redundant, its price follows immediately from the law of one price:

$$P_i = \theta_1 P_1 + \cdots + \theta_{i-1} P_{i-1} + \theta_{i+1} P_{i+1} + \cdots + \theta_I P_I.$$

We can thus focus on pricing the non-redundant assets, then the prices of all the other assets, the redundant assets, follow.

Note that the number of non-redundant assets cannot exceed the number of states. If there are more assets than states, there will be some redundant asset.

Example 3.1 Consider a one-period economy with three possible end-of-period states and four traded assets. The dividends are given in Table 3.1. With four assets and three states at least one

	state-contingent dividend		
	state 1	state 2	state 3
Asset 1	1	1	1
Asset 2	0	1	2
Asset 3	4	0	1
Asset 4	9	0	1

Table 3.1: The state-contingent dividends of the assets considered in Example 3.1.

asset is redundant. The dividend vector of asset 4 can be written as a non-trivial linear combination of the dividend vectors of assets 1, 2, and 3 since

$$\begin{pmatrix} 9 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} + 2 \begin{pmatrix} 4 \\ 0 \\ 1 \end{pmatrix}.$$

A portfolio of one unit of asset 1, minus one unit of asset 2, and two units of asset 3 perfectly replicates the dividend of asset 4, which is therefore redundant. In terms of random variables, we have the relation

$$D_1 - D_2 + 2D_3 = D_4$$

among the dividends of the four assets. On the other hand, asset 1 is redundant since it can be perfectly replicated by a portfolio of one unit of asset 2, minus two units of asset 3, and one unit of asset 4. Similarly, asset 2 is redundant and asset 3 is redundant. Hence, either of the four assets can be removed without affecting the set of dividend vectors that can be generated by forming portfolios. Note that once one of the assets has been removed, neither of the three remaining assets will be redundant anymore. Whether an asset is redundant or not depends on the set of other assets available for trade. This implies that we must remove redundant assets one by one: first we remove one redundant asset, then we look for another asset which is still redundant – if we find one, we can remove that, etc. □

In the multi-period model an asset is said to be redundant if its dividend process can be generated by a trading strategy in the other assets. In the discrete-time framework, asset i is redundant if there exists a trading strategy $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_{it} = 0$ for all t and all ω so that

$$D_{it} = D_t^{\boldsymbol{\theta}} \equiv \boldsymbol{\theta}_{t-1} \cdot (\boldsymbol{D}_t + \boldsymbol{P}_t) - \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t, \quad t = 1, \ldots, T.$$

Such a $\boldsymbol{\theta}$ is called a replicating trading strategy for asset i.

Just as in the one-period setting, redundant assets are uniquely priced by no-arbitrage.

Theorem 3.1 *If $\boldsymbol{\theta}$ is a replicating trading strategy for asset i, the unique arbitrage-free price of asset i at any time t is*

$$P_{it} = \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t.$$

Proof: The trading strategy $\hat{\boldsymbol{\theta}}$ defined by $\hat{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t - \boldsymbol{e}_i$, where $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top$, is self-financing and $V_T^{\hat{\boldsymbol{\theta}}} = 0$. No-arbitrage implies that $V_t^{\hat{\boldsymbol{\theta}}} = 0$. The result now follows since $V_t^{\hat{\boldsymbol{\theta}}} =$

$\hat{\boldsymbol{\theta}}_t \cdot \boldsymbol{P}_t = \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t - P_{it}$. □

The above definition of redundancy can be generalized to a continuous-time setting, where the theorem is also valid.

The theorem is useful for the pricing of derivatives and applied, e.g., in the Cox, Ross, and Rubinstein (1979) binomial model (see Exercise 3.1 at the end of this chapter) and the Black and Scholes (1973) continuous-time model for the pricing of stock options.

3.6 Marketed dividends and market completeness

By forming portfolios and trading strategies investors can generate other dividends than those of the individual basic assets. Any dividend that can be generated by trading the basic assets is said to be a marketed dividend. If all the dividends you can think of are marketed, the financial market is said to be complete. Otherwise the financial market is said to be incomplete. We will see in later chapters that some important results will depend on whether the financial market is complete or incomplete. Below we provide formal definitions and characterize complete markets.

3.6.1 The one-period framework

In the one-period framework dividends are random variables. A random variable is said to be a marketed dividend or spanned by traded assets if it is identical to the dividend of some portfolio of the traded assets. The dividend of a portfolio $\boldsymbol{\theta}$ is the random variable $D^{\boldsymbol{\theta}} = \boldsymbol{\theta} \cdot \boldsymbol{D}$. The set of marketed dividends is thus

$$\mathcal{M} = \{x | x = \boldsymbol{\theta} \cdot \boldsymbol{D} \text{ for some portfolio } \boldsymbol{\theta}\}. \tag{3.17}$$

Note that $\mathcal{M}$ is a subset of the set $\mathcal{L}$ of all random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with finite variance.

If some of the basic assets are redundant, they will not help us in generating dividends. Suppose that there are $k \leq I$ non-redundant assets. The k-dimensional random variable of dividends of these assets is denoted by $\hat{\boldsymbol{D}}$ and a portfolio of these assets is denoted by a k-dimensional vector $\hat{\boldsymbol{\theta}}$. We can obtain exactly the same dividends by using only the non-redundant assets as by using all dividends so

$$\mathcal{M} = \left\{x | x = \hat{\boldsymbol{\theta}} \cdot \hat{\boldsymbol{D}} \text{ for some portfolio } \hat{\boldsymbol{\theta}}\right\}. \tag{3.18}$$

The financial market is said to be complete if any dividend is marketed, i.e. if any random variable x is the dividend of some portfolio. In symbols, the market is complete if $\mathcal{M} = \mathcal{L}$.

With k non-redundant assets, the set of marketed dividends will be a k-dimensional linear subspace in $\mathcal{L}$. You have k choice variables, namely how much to invest in each of the k non-redundant assets. The dimension of $\mathcal{L}$ is the number of possible states. For each state ω you have to make sure that the dividend of the portfolio will equal the desired dividend $x(\omega)$. Whether the market is complete or not is therefore determined by the relation between the number of states and the number of non-redundant assets. If the state space is infinite, the market is clearly incomplete.

With a finite state space $\Omega = \{1, 2, \ldots, S\}$, there can be at most S non-redundant assets, i.e. $k \leq S$. If $k < S$, the market will be incomplete. If $k = S$, the market will be complete. Details

follow now: With a finite state space dividends can be represented by S-dimensional vectors and the dividends of all the I basic assets by the $I \times S$ dividend matrix $\underline{\underline{D}}$. A marketed dividend is then an S-dimensional vector $\boldsymbol{x}$ for which a portfolio $\boldsymbol{\theta}$ can be found such that

$$\underline{\underline{D}}^\top \boldsymbol{\theta} = \boldsymbol{x}$$

Removing redundant assets we eliminate rows in the dividend matrix $\underline{\underline{D}}$. As long as one of the assets is redundant, the rows of $\underline{\underline{D}}$ will be linearly dependent. The maximum number of linearly independent rows of a matrix is called the *rank* of the matrix. It can be shown that this is also the maximum number of linearly independent columns of the matrix. With k non-redundant assets, the rank of $\underline{\underline{D}}$ is equal to k. Removing the rows corresponding to those assets from $\underline{\underline{D}}$ we obtain a matrix $\underline{\underline{\hat{D}}}$ of dimension $k \times S$, where $k \leq S$ since we cannot have more than S linearly independent S-dimensional vectors. Then the set of marketed dividend vectors is

$$\mathcal{M} = \left\{ \underline{\underline{\hat{D}}}^\top \hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}} \in \mathbb{R}^k \right\}$$

since we can attain the same dividend vectors by forming portfolios of only the non-redundant assets as by forming portfolios of all the assets.

In the finite-state economy, the market is complete if $\mathcal{M} = \mathbb{R}^S$, i.e. any state-contingent dividend can be generated by forming portfolios of the traded assets. The market is complete if and only if for any $\boldsymbol{x} \in \mathbb{R}^S$, we can find $\boldsymbol{\theta} \in \mathbb{R}^I$ such that

$$\underline{\underline{D}}^\top \boldsymbol{\theta} = \boldsymbol{x}.$$

Market completeness is thus a question of when we can solve S equations in I unknowns. From linear algebra we have the following result:

Theorem 3.2 *With a finite state space, $\Omega = \{1, 2, \ldots, S\}$, the market is complete if and only if the rank of the $I \times S$ dividend matrix $\underline{\underline{D}}$ is equal to S.*

Clearly, a necessary (but not sufficient) condition for a complete market is that $I \geq S$, i.e. that there are at least as many assets as states. If the market is complete, the "pruned" dividend matrix $\underline{\underline{\hat{D}}}$ will be a non-singular $S \times S$ matrix.

Example 3.2 The market considered in Example 3.1 is complete since there are as many non-redundant assets as there are states. Any potential dividend vector can be formed by some portfolio of three of the traded assets. For example, if we let $\underline{\underline{D}}$ be the dividend matrix of the first three assets, we can generate any dividend vector $\boldsymbol{x}$ by solving $\underline{\underline{D}}^\top \boldsymbol{\theta} = \boldsymbol{x}$ for the portfolio $\boldsymbol{\theta}$ of the first three assets, i.e. $\boldsymbol{\theta} = \left(\underline{\underline{D}}^\top\right)^{-1} \boldsymbol{x}$. In the present case, we have

$$\underline{\underline{D}} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 4 & 0 & 1 \end{pmatrix}, \quad \underline{\underline{D}}^\top = \begin{pmatrix} 1 & 0 & 4 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{pmatrix}, \quad (\underline{\underline{D}}^\top)^{-1} = \begin{pmatrix} 0.2 & 1.6 & -0.8 \\ -0.2 & -0.6 & 0.8 \\ 0.2 & -0.4 & 0.2 \end{pmatrix}.$$

For example, the portfolio providing a dividend vector of $(5, -10, 5)^\top$ is given by

$$\boldsymbol{\theta} = \begin{pmatrix} 0.2 & 1.6 & -0.8 \\ -0.2 & -0.6 & 0.8 \\ 0.2 & -0.4 & 0.2 \end{pmatrix} \begin{pmatrix} 5 \\ -10 \\ 5 \end{pmatrix} = \begin{pmatrix} -19 \\ 9 \\ 6 \end{pmatrix}.$$

□

3.6.2 Multi-period framework

In a multi-period setting the market is said to be complete if for any random variable X there is a trading strategy $\boldsymbol{\theta}$ such that $D_T^{\boldsymbol{\theta}} = X$ in all states. This implies that any terminal payment you may think of can be obtained by some trading strategy. If you think of a payment before the terminal date T, this can be transformed into a terminal payment by investing it in a given asset until time T. Hence the definition covers all relevant dates and the market will thus be complete if any adapted stochastic process can be generated by some trading strategy in the given assets.

More formally, let $\mathfrak{L}$ denote the set of all random variables (with finite variance) whose outcome can be determined from the exogenous shocks to the economy over the entire time span $\mathfrak{T}$. On the other hand, let $\mathfrak{M}$ denote the set of possible time T values that can be generated by forming self-financing trading strategies in the financial market, i.e.

$$\mathfrak{M} = \left\{ V_T^{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta} \text{ self-financing } \right\}.$$

Of course, for any trading strategy $\boldsymbol{\theta}$ the terminal value $V_T^{\boldsymbol{\theta}}$ is a random variable, whose outcome is not determined until time T. Imposing relevant technical conditions imposed on trading strategies, the terminal value will have finite variance, so $\mathfrak{M}$ is always a subset of $\mathfrak{L}$. If, in fact, $\mathfrak{M}$ is equal to $\mathfrak{L}$, the financial market is complete. If not, it is said to be incomplete.

How many assets do we need in order to have a complete market? In the one-period model, Theorem 3.2 tells us that with S possible states we need at least S sufficiently different assets for the market to be complete. To generalize this result to the multi-period setting we must be careful. Consider once again the two-period model illustrated in Figures 2.1 and 2.2 in Chapter 2. Here there are six possible outcomes, i.e. Ω has six elements. Hence, one might think that we need access to trade in at least six sufficiently different assets in order for the market to be complete. This is not correct. We can do with fewer assets. This is based on two observations: (i) the uncertainty is not revealed completely at once, but little by little, and (ii) we can trade dynamically in the assets. In the example there are three possible transitions of the economy from time 0 to time 1. From our one-period analysis we know that three sufficiently different assets are enough to "span" this uncertainty. From time 1 to time 2 there are either two, three, or one possible transition of the economy, depending on which state the economy is in at time 1. At most, we need three sufficiently different assets to span the uncertainty over this period. In total, we can generate any dividend process if we just have access to three sufficiently different assets in both periods.

In the one-period model, "sufficiently different" means that the matrix of the dividends of the assets in a given period is of full rank. In the discrete-time multi-period model the payment of a given asset at the end of each period is the sum of the price and the dividend in that period. The relevant matrix is therefore the matrix of possible values of price plus dividend at the end of the period. This matrix tells us how the assets will react to the exogenous shocks over that subperiod. If we have at least as many assets that respond sufficiently differently to the shocks as we have possible realizations of these shocks, we can completely hedge the shocks.

The continuous-time financial market with uncertainty generated by a d-dimensional standard Brownian motion is complete if there are $d+1$ sufficiently different assets traded at all times. For example this is true with an instantaneously risk-free asset plus d risky assets having a price sensitivity matrix $\underline{\underline{\sigma}}_t$ of rank d. The formal proof of this result is pretty complicated and will

not be given here. We refer the interested reader to Harrison and Pliska (1981, 1983) and Duffie (2001). However, the result is quite intuitive given the following observations:

- For continuous changes over an instant, only means and variances matter.
- We can approximate the d-dimensional shock $d\boldsymbol{z}_t$ by a random variable that takes on $d+1$ possible values and has the same mean and variance as $d\boldsymbol{z}_t$.
- For example, a one-dimensional shock dz_t has mean zero and variance dt. This is also true for a random variable ε which equals $\sqrt{dt}$ with a probability of $1/2$ and equals $-\sqrt{dt}$ with a probability of $1/2$.
- With continuous trading, we can adjust our exposure to the exogenous shocks every instant.

Over each instant we can thus think of the model with uncertainty generated by a d-dimensional standard Brownian motion as a discrete-time model with $d+1$ states. Therefore it only takes $d+1$ sufficiently different assets to complete the market. For example, if all market prices are affected by a single shock, a one-dimensional Brownian motion, the market will be complete if you can always find two assets with different sensitivities towards that shock. One of the assets could be a risk-free asset. Note however that when counting the number of shocks you should include shocks to variables that contain information about future asset prices.

Let us take a simple, concrete example. Suppose that assets pay no intermediate dividends and that the price dynamics of an arbitratey asset, asset i, is given by

$$dP_{it} = P_{it}\left[\mu_i\,dt + \sigma_i\,dz_t\right],$$

where $z = (z_t)$ is a one-dimensional standard Brownian motion and μ_i, σ_i are constants. Suppose that assets i and j have different sensitivities towards the shock, i.e. that $\sigma_i \neq \sigma_j$. From (3.14), it follows that a self-financing trading strategy with a weight of π_t on asset i and $1-\pi_t$ on asset j will generate a value process V_t with dynamics

$$dV_t = V_t\left[(\pi_t(\mu_i-\mu_j)+\mu_j)\;dt + (\pi_t(\sigma_i-\sigma_j)+\sigma_j)\;dz_t\right].$$

We can obtain any desired sensitivity towards the shock. The portfolio weight $\pi_t = (\nu_t-\sigma_j)/(\sigma_i-\sigma_j)$ will generate a portfolio return sensitivity of ν_t. Now suppose that each μ_i is a function of some variable x_t and that the dynamics of x_t is affected by another shock represented by a standard Brownian motion $\hat{z} = (\hat{z}_t)$ independent of z, e.g.

$$dx_t = m(x_t)\,dt + v(x_t)\left(\rho\,dz_t + \sqrt{1-\rho^2}\,d\hat{z}_t\right).$$

Then the full uncertainty of the model is generated by a two-dimensional shock $(z,\hat{z})$. Since the instantaneous price changes are affected only by z, the market is now incomplete. You cannot hedge against the shock process $\hat{z}$.

For a moment think about a continuous-time model where there can be a jump in the value of one of the key variables. Suppose that in case of a jump the variable can jump to K different values. The change over an instant in such a variable cannot be represented only by the expectation and the variance of the jump. To obtain any desired exposure to the jump risk you need K assets which react sufficiently different to the jump. Typical models with jump risk assume that the

size of the jump is normally or lognormally distributed. In both cases there are infinitely many possible realizations of the jump size and, consequently, a market with finitely many assets will be incomplete.

3.6.3 Discussion

As we shall see in later sections, some fundamental results require that the market is complete. It is therefore relevant to assess the realism of this property. Market completeness means that every risk is traded. Individuals can insure against all risks (relevant for asset prices) through trading in financial assets. Clearly, individuals would like to have that opportunity so if the market is incomplete there will be an incentive to create new non-redundant assets that will help complete the market. In fact, at least part of the many new assets that have been introduced in the financial markets over the last decades do help complete the market, e.g. assets with dividends depending on the stock market volatility, the weather at a given location, or the number of natural catastrophes. On the other hand, some risks are difficult to market. For example, due to the obvious information asymmetry, it is unlikely that individual labor income risk can be fully insured in the financial market. Maybe you would like to obtain full income insurance but who should provide that, given that you probably know a lot more about your potential income than anybody else—and that you can influence your own income while others cannot. Hence, we should not expect that real-life financial markets are complete in the strict sense.

If the market is incomplete, it is not possible to adjust your exposure to one or several shocks included in the model. But maybe investors do not care about those shocks. Then the model is said to be effectively complete. Before solving for prices and optimal decisions of the individuals, it is generally impossible to decide whether an incomplete market is really effectively complete. We will return to this discussion and some formal results on that in Chapter 7.

3.7 Concluding remarks

To be added...

3.8 Exercises

EXERCISE 3.1 Consider a one-period model with only two possible end-of-period states. Three assets are traded in an arbitrage-free market. Asset 1 is a risk-free asset with a price of 1 and an end-of-period dividend of R^f, the risk-free gross rate of return. Asset 2 has a price of S and offers a dividend of uS in state 1 and dS in state 2.

(a) Show that if the inequality $d < R^f < u$ does not hold, there will be an arbitrage.

Asset 3 is a call-option on asset 2 with an exercise price of K. The dividend of asset 3 is therefore $C_u \equiv \max(uS - K, 0)$ in state 1 and $C_d \equiv \max(dS - K, 0)$ in state 2.

(b) Show that a portfolio consisting of θ_1 units of asset 1 and θ_2 units of asset 2, where

$$\theta_1 = (R^f)^{-1} \frac{uC_d - dC_u}{u - d}, \quad \theta_2 = \frac{C_u - C_d}{(u - d)S}$$

will generate the same dividend as the option.

(c) Show that the no-arbitrage price of the option is given by

$$C = (R^f)^{-1} \left(qC_u + (1-q)C_d \right),$$

where $q = (R^f - d)/(u - d)$.

EXERCISE 3.2 Imagine a one-period economy with two possible end-of-period states that are equally likely. Two assets are traded. Asset 1 has an initial price of 1 and pays off 1 in state 1 and 2 in state 2. Asset 2 has an initial price of 3 and gives a payoff of 2 in state 1 and a payoff k in state 2, where k is some constant.

(a) Argue that if $k = 4$, the Law of One Price does not hold. Is the Law of One Price violated for other values of k?

(b) For what values of k is the market complete?

(c) For what values of k is the market free of arbitrage?

(d) Assume $k = 8$. Is it possible to obtain a risk-free dividend? If so, what is the risk-free rate?

EXERCISE 3.3 Verify Equation (3.4).

EXERCISE 3.4 In a one-period two-state economy the risk-free interest rate over the period is 25%. An asset that pays out 100 in state 1 and 200 in state 2 trades at a price of 110.

(a) What is the no-arbitrage price of a second risky asset that pays out 200 in state 1 and 100 in state 2?

(b) If this second risky asset trades at a higher price than what you computed in (a), how can you obtain a risk-free profit?

Chapter 4

State prices

4.1 Introduction

If you want to price a set of assets, you could take them one by one and evaluate the dividends of each asset separately. However, to evaluate all assets in a consistent way (e.g. avoiding arbitrage) it is a better strategy first to figure out what your general pricing rule should be and subsequently you can apply that to any given dividend stream. The general pricing rule can be represented by a state-price deflator, which is the topic of this chapter. Basically, a state-price deflator contains information about the valuation of additional payments in different states and at different points in time. Combining that with the state- and time-dependent dividends of any asset, you can compute a value or price of that asset.

Section 4.2 defines the state-price deflator in each of our general frameworks (one-period, discrete-time, continuous-time) and derives some immediate consequences for prices and expected returns. Further important properties of state-price deflators are obtained in Section 4.3. Section 4.4 explains the difference between real and nominal state-price deflators. Finally, Section 4.5 gives a preview of some alternative ways of representing the information in a state-price deflator. These alternatives are preferable for some purposes and will be studied in more detail in later chapters.

The concept of state prices was introduced and studied by Arrow (1951, 1953, 1970), Debreu (1954), Negishi (1960), and Ross (1978).

4.2 Definitions and immediate consequences

This section gives a formal definition of a state-price deflator. Some authors use the name stochastic discount factor, event-price deflator, or pricing kernel instead of state-price deflator.

4.2.1 The one-period framework

A state-price deflator is a random variable ζ with the properties that

(i) ζ has finite variance,

(ii) $\zeta > 0$, i.e. $\zeta(\omega) > 0$ for all states $\omega \in \Omega$,

(iii) the price of the I basic assets are given by

$$P_i = \mathrm{E}[\zeta D_i], \quad i = 1, 2, \ldots, I, \tag{4.1}$$

or, more compactly, $\boldsymbol{P} = \mathrm{E}[\zeta \boldsymbol{D}]$.

The finite variance assumptions on both the state-price deflator and the dividends ensure that the expectation $\mathrm{E}[\zeta D_i]$ is finite.

We get a similar pricing equation for portfolios:

$$P^{\boldsymbol{\theta}} = \sum_{i=1}^{I} \theta_i P_i = \sum_{i=1}^{I} \theta_i \,\mathrm{E}[\zeta D_i] = \mathrm{E}\left[\zeta \left(\sum_{i=1}^{I} \theta_i D_i\right)\right] = \mathrm{E}\left[\zeta D^{\boldsymbol{\theta}}\right]. \tag{4.2}$$

In terms of gross rates of returns, $R_i = D_i/P_i$, a state-price deflator ζ has the property that

$$1 = \mathrm{E}[\zeta R_i], \quad i = 1, 2, \ldots, I, \tag{4.3}$$

or, $\mathbf{1} = \mathrm{E}[\zeta \boldsymbol{R}]$ in vector notation.

For a risk-free portfolio with a dividend of 1, the price P^f is

$$P^f = \mathrm{E}\left[\zeta\right]$$

and the risk-free gross rate of return is thus

$$R^f = \frac{1}{P^f} = \frac{1}{\mathrm{E}\left[\zeta\right]}. \tag{4.4}$$

Exploiting the definition of a covariance, the pricing condition (4.1) can be rewritten as

$$P_i = \mathrm{E}[\zeta]\,\mathrm{E}[D_i] + \mathrm{Cov}[D_i, \zeta]. \tag{4.5}$$

A dividend of a given size is valued more highly in a state for which the state-price deflator is high than in a state where the deflator is low. If a risk-free asset is traded, we can rewrite this as

$$P_i = \frac{\mathrm{E}[D_i] + R^f \,\mathrm{Cov}[D_i, \zeta]}{R^f}, \tag{4.6}$$

i.e. the value of a future dividend is given by the expected dividend adjusted by a covariance term, discounted at the risk-free rate. If the dividend is positively [negatively] covarying with the state-price deflator, the expected dividend is adjusted downwards [upwards]. Defining the dividend-beta of asset i with respect to the state-price deflator as

$$\beta[D_i, \zeta] = \frac{\mathrm{Cov}[D_i, \zeta]}{\mathrm{Var}[\zeta]} \tag{4.7}$$

and $\eta = -\,\mathrm{Var}[\zeta]/\,\mathrm{E}[\zeta] < 0$, we can rewrite the above pricing equation as

$$P_i = \frac{\mathrm{E}[D_i] - \beta[D_i, \zeta]\eta}{R^f}. \tag{4.8}$$

Some basic finance textbooks suggest that a future uncertain dividend can be valued by taking the expected dividend and discount it by a discount rate reflecting the risk of the dividend. For asset i, this discount rate $\hat{R}_i$ is implicitly defined by $P_i = \mathrm{E}[D_i]/\hat{R}_i$ and combining this with the above equations, we must have

$$\hat{R}_i = \frac{R^f \,\mathrm{E}[D_i]}{\mathrm{E}[D_i] - \beta[D_i, \zeta]\eta} = R^f \frac{1}{1 - \frac{\beta[D_i, \zeta]\eta}{\mathrm{E}[D_i]}}. \tag{4.9}$$

The risk-adjusted discount rate does not depend on the scale of the dividend in the sense that the risk-adjusted discount rate for a dividend of kD_i for any constant k is the same as for a dividend of D_i. Note that the risk-adjusted discount rate will be smaller than R^f if the dividend is negatively covarying with the stare-price deflator. In fact, if $\mathrm{E}[D_i] < \beta[D_i, \zeta]\eta$, the risk-adjusted gross discount rate will be negative! While this possibility is rarely realized in textbooks, it is not really surprising. Some assets or investments will have a negative expected future dividend but still a positive value today. This is the case for most insurance contracts. The lesson here is to be careful if you want to value assets by discounting expected dividends by risk-adjusted discount rates.

For the gross rate of return, $R_i = D_i/P_i$, we get

$$1 = \mathrm{E}[\zeta]\,\mathrm{E}[R_i] + \mathrm{Cov}[R_i, \zeta]$$

implying that

$$\mathrm{E}[R_i] = \frac{1}{\mathrm{E}[\zeta]} - \frac{\mathrm{Cov}[R_i, \zeta]}{\mathrm{E}[\zeta]}. \tag{4.10}$$

If a risk-free asset is available, the above equation specializes to

$$\mathrm{E}[R_i] - R^f = -\frac{\mathrm{Cov}[R_i, \zeta]}{\mathrm{E}[\zeta]}, \tag{4.11}$$

which again can be rewritten as

$$\mathrm{E}[R_i] - R^f = \frac{\mathrm{Cov}[R_i, \zeta]}{\mathrm{Var}[\zeta]} \left(-\frac{\mathrm{Var}[\zeta]}{\mathrm{E}[\zeta]} \right) = \beta[R_i, \zeta]\eta, \tag{4.12}$$

where the return-beta $\beta[R_i, \zeta]$ is defined as $\mathrm{Cov}[R_i, \zeta]/\,\mathrm{Var}[\zeta]$ corresponding to the definition of the market-beta of an asset in the traditional CAPM. An asset with a positive [negative] return-beta with respect to the state-price deflator will have an expected return smaller [larger] than the risk-free return. Of course, we can also write the covariance as the product of the correlation and the standard deviations so that

$$\mathrm{E}[R_i] - R^f = -\rho[R_i, \zeta]\sigma[R_i]\frac{\sigma[\zeta]}{\mathrm{E}[\zeta]}, \tag{4.13}$$

and the Sharpe ratio of asset i is

$$\frac{\mathrm{E}[R_i] - R^f}{\sigma[R_i]} = -\rho[R_i, \zeta]\frac{\sigma[\zeta]}{\mathrm{E}[\zeta]}. \tag{4.14}$$

The expressions involving expected returns and return-betas above are not directly useful if you want to value a future dividend. For that purpose the equations with expected dividends and dividend-betas are superior. On the other hand the return-expressions are better for empirical studies, where you have a historical record of observations of returns and, for example, of a potential state-price deflator.

Example 4.1 Let ζ be a state-price deflator and consider a dividend given by

$$D_i = a + b\zeta + \varepsilon,$$

where a, b are constants, and where ε is a random variable with mean zero and $\mathrm{Cov}[\varepsilon, \zeta] = 0$. What is the price of this dividend? We can use the original pricing condition to get

$$P_i = \mathrm{E}[D_i\zeta] = \mathrm{E}\left[(a + b\zeta + \varepsilon)\,\zeta\right] = a\,\mathrm{E}[\zeta] + b\,\mathrm{E}[\zeta^2],$$

using $\mathrm{E}[\varepsilon\zeta] = \mathrm{Cov}[\varepsilon,\zeta] + \mathrm{E}[\varepsilon]\,\mathrm{E}[\zeta] = 0 + 0 = 0$. Alternatively, we can compute

$$\mathrm{E}[D_i] = a + b\,\mathrm{E}[\zeta], \qquad \mathrm{Cov}[D_i,\zeta] = b\,\mathrm{Var}[\zeta],$$

and use (4.5) to get

$$\begin{aligned} P_i &= \mathrm{E}[\zeta]\,(a + b\,\mathrm{E}[\zeta]) + b\,\mathrm{Var}[\zeta] \\ &= a\,\mathrm{E}[\zeta] + b\,\mathrm{E}[\zeta^2] \end{aligned}$$

using the identity $\mathrm{Var}[\zeta] = \mathrm{E}[\zeta^2] - (\mathrm{E}[\zeta])^2$. □

Some empirical tests of asset pricing models focus on excess returns, where returns on all assets are measured relative to the return on a fixed benchmark asset or portfolio. Let $\bar{R}$ be the return on the benchmark. The excess return on asset i is then $R_i^e \equiv R_i - \bar{R}$. Since the above equations hold for both asset i and the benchmark, we get that

$$\mathrm{E}[\zeta R_i^e] = 0$$

and

$$\mathrm{E}[R_i^e] = -\frac{\mathrm{Cov}[R_i^e,\zeta]}{\mathrm{E}[\zeta]} = \beta[R_i^e,\zeta]\eta, \tag{4.15}$$

where $\beta[R_i^e,\zeta] = \beta[R_i,\zeta] - \beta[\bar{R},\zeta]$.

If the state space is finite, $\Omega = \{1,2,\ldots,S\}$, we can alternatively represent a general pricing rule by a *state-price vector*, which is an S-dimensional vector $\boldsymbol{\psi} = (\psi_1,\ldots,\psi_S)^\top$ with the properties

(i) $\boldsymbol{\psi} > 0$, i.e. $\psi_\omega > 0$ for all $\omega = 1,2,\ldots,S$,

(ii) the price of the I assets are given by

$$P_i = \boldsymbol{\psi}\cdot\boldsymbol{D}_i = \sum_{\omega=1}^{S}\psi_\omega D_{i\omega}, \quad i = 1,2,\ldots,I, \tag{4.16}$$

or, more compactly, $\boldsymbol{P} = \underline{\underline{D}}\boldsymbol{\psi}$, where $\underline{\underline{D}}$ is the dividend matrix of all the basic assets.

Suppose we can construct an Arrow-Debreu asset for state ω, i.e. a portfolio paying 1 in state ω and nothing in all other states. The price of this portfolio will be equal to ψ_ω, the "state price for state ω." The price of a risk-free dividend of 1 is $P^f \equiv \boldsymbol{\psi}\cdot\mathbf{1} = \sum_{\omega=1}^{S}\psi_\omega$ so that the gross risk-free rate of return is

$$R^f = \frac{1}{\sum_{\omega=1}^{S}\psi_\omega} = \frac{1}{\boldsymbol{\psi}\cdot\mathbf{1}}.$$

There is a one-to-one correspondence between state-price vectors and state-price deflators. With a finite state space a state-price deflator is equivalent to a vector $\boldsymbol{\zeta} = (\zeta_1,\zeta_2,\ldots,\zeta_S)^\top$ and we can rewrite (4.1) as

$$P_i = \sum_{\omega=1}^{S} p_\omega\zeta_\omega D_{i\omega}, \quad i = 1,2,\ldots,I.$$

We can then define a state-price vector $\boldsymbol{\psi}$ by

$$\psi_\omega = \zeta_\omega p_\omega. \tag{4.17}$$

Conversely, given a state-price vector this equation defines a state-price deflator. With infinitely many states we cannot meaningfully define state-price vectors but we can still define state-price deflators in terms of random variables.

Example 4.2 Consider the same market as in Example 3.1. Suppose there is a state-price vector $\boldsymbol{\psi} = (0.3, 0.2, 0.3)^\top$. Then we can compute the prices of the four assets as $\boldsymbol{P} = \underline{\underline{D}}\boldsymbol{\psi}$, i.e.

$$\begin{pmatrix} P_1 \\ P_2 \\ P_3 \\ P_4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 4 & 0 & 1 \\ 9 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0.3 \\ 0.2 \\ 0.3 \end{pmatrix} = \begin{pmatrix} 0.8 \\ 0.8 \\ 1.5 \\ 3 \end{pmatrix}.$$

In particular, the gross risk-free rate of return is $1/(0.3+0.2+0.3) = 1.25$ corresponding to a 25% risk-free net rate of return.

If the state probabilities are 0.5, 0.25, and 0.25, respectively, the state-price deflator corresponding to the state-price vector is given by

$$\zeta_1 = \frac{0.3}{0.5} = 0.6, \quad \zeta_2 = \frac{0.2}{0.25} = 0.8, \quad \zeta_3 = \frac{0.3}{0.25} = 1.2.$$

□

4.2.2 The discrete-time framework

In the discrete-time multi-period framework with time set $\mathcal{T} = \{0, 1, 2, \ldots, T\}$ a state-price deflator is an adapted stochastic process $\zeta = (\zeta_t)_{t\in\mathcal{T}}$ such that

(i) $\zeta_0 = 1$,

(ii) $\zeta_t > 0$ for all $t = 1, 2, \ldots, T$,

(iii) for any $t \in \mathcal{T}$, ζ_t has finite variance,

(iv) for any basic asset $i = 1, \ldots, I$ and any $t \in \mathcal{T}$, the price satisfies

$$P_{it} = \mathrm{E}_t\left[\sum_{s=t+1}^{T} D_{is}\frac{\zeta_s}{\zeta_t}\right]. \tag{4.18}$$

The condition (i) is just a normalization. The condition (iii) is purely technical and will ensure that some relevant expectations exist. Condition (iv) gives the price at time t in terms of all the future dividends and the state-price deflator. This condition will also hold for all trading strategies.

The pricing condition implies a link between the price of an asset at two different points in time, say $t < t'$. From (4.18) we have

$$P_{it'} = \mathrm{E}_{t'}\left[\sum_{s=t'+1}^{T} D_{is}\frac{\zeta_s}{\zeta_{t'}}\right].$$

We can now rewrite the price P_{it} as follows:

$$\begin{aligned}
P_{it} &= \mathrm{E}_t\left[\sum_{s=t+1}^{T} D_{is}\frac{\zeta_s}{\zeta_t}\right] \\
&= \mathrm{E}_t\left[\sum_{s=t+1}^{t'} D_{is}\frac{\zeta_s}{\zeta_t} + \sum_{s=t'+1}^{T} D_{is}\frac{\zeta_s}{\zeta_t}\right] \\
&= \mathrm{E}_t\left[\sum_{s=t+1}^{t'} D_{is}\frac{\zeta_s}{\zeta_t} + \frac{\zeta_{t'}}{\zeta_t}\sum_{s=t'+1}^{T} D_{is}\frac{\zeta_s}{\zeta_{t'}}\right] \\
&= \mathrm{E}_t\left[\sum_{s=t+1}^{t'} D_{is}\frac{\zeta_s}{\zeta_t} + \frac{\zeta_{t'}}{\zeta_t}\,\mathrm{E}_{t'}\left[\sum_{s=t'+1}^{T} D_{is}\frac{\zeta_s}{\zeta_{t'}}\right]\right] \\
&= \mathrm{E}_t\left[\sum_{s=t+1}^{t'} D_{is}\frac{\zeta_s}{\zeta_t} + P_{it'}\frac{\zeta_{t'}}{\zeta_t}\right].
\end{aligned} \tag{4.19}$$

Here the fourth equality follows from the Law of Iterated Expectations, Theorem 2.1. Conversely, Equation (4.19) implies Equation (4.18).

A particularly simple version of (4.19) occurs for $t' = t+1$:

$$P_{it} = \mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\left(P_{i,t+1} + D_{i,t+1}\right)\right], \tag{4.20}$$

or in terms of the gross rate of return $R_{i,t+1} = (P_{i,t+1} + D_{i,t+1})/P_{it}$:

$$1 = \mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}R_{i,t+1}\right]. \tag{4.21}$$

These equations show that the ratio ζ_{t+1}/ζ_t acts as a one-period state-price deflator between time t and $t+1$ in the sense of the definition in the one-period framework, except that the price at the end of the period (zero in the one-period case) is added to the dividend. Of course, given the state-price deflator process $\zeta = (\zeta_t)$ we know the state-price deflators ζ_{t+1}/ζ_t for each of the subperiods. (Note that these will depend on the realized value ζ_t which is part of the information available at time t.) Conversely, a sequence of "one-period state-price deflators" ζ_{t+1}/ζ_t and the normalization $\zeta_0 = 1$ define the entire state-price deflator process.

If an asset provides a risk-free gross rate of return R_t^f over the period between t and $t+1$ that return will be known at time t and we get that

$$1 = \mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}R_t^f\right] = R_t^f\,\mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right] \tag{4.22}$$

and hence

$$R_t^f = \left(\mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right]\right)^{-1}. \tag{4.23}$$

Similar to the one-period case, the above equations lead to an expression for the expected excess return of an asset over a single period:

$$\mathrm{E}_t[R_{i,t+1}] - R_t^f = -\frac{\mathrm{Cov}_t\left[R_{i,t+1}, \frac{\zeta_{t+1}}{\zeta_t}\right]}{\mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right]} = \beta_t\left[R_{i,t+1}, \frac{\zeta_{t+1}}{\zeta_t}\right]\eta_t, \tag{4.24}$$

where

$$\beta_t\left[R_{i,t+1}, \frac{\zeta_{t+1}}{\zeta_t}\right] = \frac{\mathrm{Cov}_t\left[R_{i,t+1}, \frac{\zeta_{t+1}}{\zeta_t}\right]}{\mathrm{Var}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right]}, \quad \eta_t = -\frac{\mathrm{Var}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right]}{\mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right]}.$$

The same relation holds for net rates of return. Just substitute $R_{i,t+1} = 1 + r_{i,t+1}$ and $R_t^f = 1 + r_t^f$ on the left-hand side and observe that the ones cancel. On the right-hand side use that $\mathrm{Cov}_t[R_{i,t+1}, x] = \mathrm{Cov}_t[1 + r_{i,t+1}, x] = \mathrm{Cov}_t[r_{i,t+1}, x]$ for any random variable x. Therefore we can work with gross or net returns as we like in such expressions. The conditional Sharpe ratio of asset i becomes

$$\frac{\mathrm{E}_t[R_{i,t+1}] - R_t^f}{\sigma_t[R_{i,t+1}]} = -\rho_t\left[R_{i,t+1}, \frac{\zeta_{t+1}}{\zeta_t}\right]\frac{\sigma_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right]}{\mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right]}. \tag{4.25}$$

These expressions are useful for empirical purposes.

For the valuation of a future dividend stream we have to apply (4.18). Applying the covariance definition once again, we can rewrite the price as

$$P_{it} = \sum_{s=t+1}^{T}\left(\mathrm{E}_t[D_{is}]\,\mathrm{E}_t\left[\frac{\zeta_s}{\zeta_t}\right] + \mathrm{Cov}_t\left[D_{is}, \frac{\zeta_s}{\zeta_t}\right]\right).$$

By definition of a state-price deflator, a zero-coupon bond maturing at time s with a face value of 1 will have a time t price of

$$B_t^s = \mathrm{E}_t\left[\frac{\zeta_s}{\zeta_t}\right].$$

Define the corresponding (annualized) yield $\hat{y}_t^s$ by

$$B_t^s = \frac{1}{(1+\hat{y}_t^s)^{s-t}} \quad \Leftrightarrow \quad \hat{y}_t^s = (B_t^s)^{-1/(s-t)} - 1. \tag{4.26}$$

Note that this is a risk-free rate of return between time t and time s. Now we can rewrite the above price expression as

$$\begin{aligned} P_{it} &= \sum_{s=t+1}^{T} B_t^s\left(\mathrm{E}_t[D_{is}] + \frac{\mathrm{Cov}_t\left[D_{is}, \frac{\zeta_s}{\zeta_t}\right]}{B_t^s}\right) \\ &= \sum_{s=t+1}^{T}\frac{\mathrm{E}_t[D_{is}] + \frac{\mathrm{Cov}_t\left[D_{is}, \frac{\zeta_s}{\zeta_t}\right]}{\mathrm{E}_t\left[\frac{\zeta_s}{\zeta_t}\right]}}{(1+\hat{y}_t^s)^{s-t}}. \end{aligned} \tag{4.27}$$

Each dividend is valued by discounting an appropriately risk-adjusted expected dividend by the risk-free return over the period. This generalizes the result in the one-period framework.

Example 4.3 Assume that the state-price deflator $\zeta = (\zeta_t)$ satisfies

$$\mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right] = \mu_\zeta, \quad \mathrm{Var}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right] = \sigma_\zeta^2$$

for each $t = 0, 1, 2, \ldots, T-1$. In particular, $\mathrm{E}_t\left[(\zeta_{t+1}/\zeta_t)^2\right] = \sigma_\zeta^2 + \mu_\zeta^2$. Consider an uncertain stream of dividends $D = (D_t)$, where the dividend growth rate is given by

$$\frac{D_{t+1}}{D_t} = a + b\frac{\zeta_{t+1}}{\zeta_t} + \varepsilon_{t+1}, \quad t = 0, 1, \ldots, T-1,$$

where $\varepsilon_1, \ldots, \varepsilon_T$ are independent with $\mathrm{E}_t[\varepsilon_{t+1}] = 0$ and $\mathrm{E}_t[\varepsilon_{t+1}\zeta_{t+1}] = 0$ for all t.

First note that

$$\begin{aligned}\mathrm{E}_t\left[\frac{D_{t+1}}{D_t}\frac{\zeta_{t+1}}{\zeta_t}\right] &= \mathrm{E}_t\left[\left(a+b\frac{\zeta_{t+1}}{\zeta_t}+\varepsilon_{t+1}\right)\frac{\zeta_{t+1}}{\zeta_t}\right] = a\,\mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right]+b\,\mathrm{E}_t\left[\left(\frac{\zeta_{t+1}}{\zeta_t}\right)^2\right]\\ &= a\mu_\zeta + b\left(\sigma_\zeta^2+\mu_\zeta^2\right)\equiv A\end{aligned}$$

for every $t=0,1,\ldots,T-1$. Together with the Law of Iterated Expectations this implies that for each $s>t$

$$\begin{aligned}\mathrm{E}_t\left[\frac{D_s}{D_t}\frac{\zeta_s}{\zeta_t}\right] &= \mathrm{E}_t\left[\frac{D_{s-1}}{D_t}\frac{\zeta_{s-1}}{\zeta_t}\frac{D_s}{D_{s-1}}\frac{\zeta_s}{\zeta_{s-1}}\right]\\ &= \mathrm{E}_t\left[\frac{D_{s-1}}{D_t}\frac{\zeta_{s-1}}{\zeta_t}\,\mathrm{E}_{s-1}\left[\frac{D_s}{D_{s-1}}\frac{\zeta_s}{\zeta_{s-1}}\right]\right]\\ &= A\,\mathrm{E}_t\left[\frac{D_{s-1}}{D_t}\frac{\zeta_{s-1}}{\zeta_t}\right]\\ &= \ldots\\ &= A^{s-t}.\end{aligned}$$

The price-dividend ratio of the asset is therefore

$$\begin{aligned}\frac{P_t}{D_t} &= \mathrm{E}_t\left[\sum_{s=t+1}^{T}\frac{D_s}{D_t}\frac{\zeta_s}{\zeta_t}\right] = \sum_{s=t+1}^{T}\mathrm{E}_t\left[\frac{D_s}{D_t}\frac{\zeta_s}{\zeta_t}\right]\\ &= \sum_{s=t+1}^{T}A^{s-t} = A+A^2+\cdots+A^{T-t}\\ &= \frac{A}{1-A}\left(1-A^{T-t}\right).\end{aligned}$$

The price is thus

$$P_t = D_t\frac{A}{1-A}\left(1-A^{T-t}\right).$$

□

Some authors formulate the important pricing condition (iv) as follows: for all basic assets $i=1,\ldots,I$ the state-price deflated gains process $G_i^\zeta=(G_{it}^\zeta)_{t\in\mathcal{T}}$ defined by

$$G_{it}^\zeta = \sum_{s=1}^{t}D_{is}\zeta_s + P_{it}\zeta_t \tag{4.28}$$

is a martingale. This means that for $t<t'$, $G_{it}^\zeta=\mathrm{E}[G_{it'}^\zeta]$, i.e.

$$\sum_{s=1}^{t}D_{is}\zeta_s + P_{it}\zeta_t = \mathrm{E}_t\left[\sum_{s=1}^{t'}D_{is}\zeta_s + P_{it'}\zeta_{t'}\right].$$

Subtracting the sum on the left-hand side and dividing by ζ_t yield (4.19).

4.2.3 The continuous-time framework

Similar to the discrete-time framework we define a state-price deflator in the continuous-time framework as an adapted stochastic process $\zeta=(\zeta_t)$ with

(i) $\zeta_0 = 1$,

(ii) $\zeta_t > 0$ for all $t \in [0, T]$,

(iii) for each t, ζ_t has finite variance,

(iv) for any basic asset $i = 1, \ldots, I$ and any $t \in [0, T)$, the price satisfies

$$P_{it} = \mathrm{E}_t \left[\int_t^T \delta_{is} P_{is} \frac{\zeta_s}{\zeta_t} \, ds + D_{iT} \frac{\zeta_T}{\zeta_t} \right]. \tag{4.29}$$

Under technical conditions, Equation (4.29) will also hold for all trading strategies. As in the discrete-time case the pricing equation (4.29) implies that for any $t < t' < T$

$$P_{it} = \mathrm{E}_t \left[\int_t^{t'} \delta_{is} P_{is} \frac{\zeta_s}{\zeta_t} \, ds + P_{it'} \frac{\zeta_{t'}}{\zeta_t} \right]. \tag{4.30}$$

Again, the condition (iv) can be reformulated as follows: for all basic assets $i = 1, \ldots, I$ the state-price deflated gains process $G_i^\zeta = (G_{it}^\zeta)_{t \in \mathcal{T}}$ defined by

$$G_{it}^\zeta = \begin{cases} \int_0^t \delta_{is} P_{is} \zeta_s \, ds + P_{it} \zeta_t & \text{for } t < T, \\ \int_0^T \delta_{is} P_{is} \zeta_s \, ds + D_{iT} \zeta_T & \text{for } t = T \end{cases} \tag{4.31}$$

is a martingale.

If we invest in one unit of asset i at time t and keep reinvesting the continuously paid dividends in the asset, we will end up at time T with $\exp\{\int_t^T \delta_{iu} \, du\}$ units of the asset, cf. the argument in Section 3.2.3. Therefore, we have the following relation:

$$P_{it} = \mathrm{E}_t \left[e^{\int_t^T \delta_{iu} \, du} D_{iT} \frac{\zeta_T}{\zeta_t} \right]. \tag{4.32}$$

As in the discrete-time case we can derive information about the short-term risk-free rate of return and the expected returns on the risky assets from the state-price deflator. First we do this informally by considering a discrete-time approximation with period length Δt and let $\Delta t \to 0$ at some point. In such an approximate model the risk-free one-period gross rate of return satisfies

$$\frac{1}{R_t^f} = \mathrm{E}_t \left[\frac{\zeta_{t+\Delta t}}{\zeta_t} \right],$$

according to (4.23). In terms of the annualized continuously compounded one-period risk-free rate r_t^f, the left-hand side is given by $\exp\{-r_t^f \Delta t\} \approx 1 - r_t^f \Delta t$. Subtracting one on both sides of the equation and changing signs, we get

$$r_t^f \Delta t = -\mathrm{E}_t \left[\frac{\zeta_{t+\Delta t} - \zeta_t}{\zeta_t} \right].$$

Dividing by Δt and letting $\Delta t \to 0$ we get

$$r_t^f = -\lim_{\Delta t \to 0} \frac{1}{\Delta t} \mathrm{E}_t \left[\frac{\zeta_{t+\Delta t} - \zeta_t}{\zeta_t} \right] = -\frac{1}{dt} \mathrm{E}_t \left[\frac{d\zeta_t}{\zeta_t} \right].$$

The left-hand side is the continuously compounded short-term risk-free interest rate. The right-hand side is minus the relative drift of the state-price deflator.

To obtain an expression for the current expected return on a risky asset start with the equivalent of (4.21) in the approximating discrete-time model, i.e.

$$\mathrm{E}_t\left[\frac{\zeta_{t+\Delta}}{\zeta_t}R_{i,t+\Delta t}\right]=1,$$

which implies that

$$\mathrm{E}_t\left[R_{i,t+\Delta t}\right]\mathrm{E}_t\left[\frac{\zeta_{t+\Delta}}{\zeta_t}\right]-1=-\operatorname{Cov}_t\left[R_{i,t+\Delta t},\frac{\zeta_{t+\Delta}}{\zeta_t}\right].$$

For small Δt, we have

$$\begin{aligned}\mathrm{E}_t\left[R_{i,t+\Delta t}\right]&\approx e^{(\delta_{it}+\mu_{it})\Delta t},\\ \mathrm{E}_t\left[\frac{\zeta_{t+\Delta t}}{\zeta_t}\right]&\approx e^{-r_t^f\Delta t},\\ \operatorname{Cov}_t\left[R_{i,t+\Delta t},\frac{\zeta_{t+\Delta}}{\zeta_t}\right]\approx\operatorname{Cov}_t\left[\frac{P_{i,t+\Delta t}}{P_{it}},\frac{\zeta_{t+\Delta t}}{\zeta_t}\right]&=\operatorname{Cov}_t\left[\frac{P_{i,t+\Delta t}-P_{it}}{P_{it}},\frac{\zeta_{t+\Delta t}-\zeta_t}{\zeta_t}\right].\end{aligned}$$

If we substitute these expressions into the previous equation, use $e^x\approx 1+x$ on the left-hand side, divide by Δt, and let $\Delta t\to 0$ we arrive at

$$\mu_{it}+\delta_{it}-r_t^f=-\lim_{\Delta t\to 0}\frac{1}{\Delta t}\operatorname{Cov}_t\left[\frac{P_{i,t+\Delta t}-P_{it}}{P_{it}},\frac{\zeta_{t+\Delta t}-\zeta_t}{\zeta_t}\right]=-\frac{1}{dt}\operatorname{Cov}_t\left[\frac{dP_{it}}{P_{it}},\frac{d\zeta_t}{\zeta_t}\right].\tag{4.33}$$

The left-hand side is the expected excess rate of return over the next instant. The right-hand side is the current rate of covariance between the return on the asset and the relative change of the state-price deflator.

Now let us give a more rigorous treatment. Write the dynamics of a state-price deflator as

$$d\zeta_t=-\zeta_t\left[m_t\,dt+\boldsymbol{\lambda}_t^\top\,d\boldsymbol{z}_t\right]\tag{4.34}$$

for some relative drift m and some "sensitivity" vector $\boldsymbol{\lambda}$. First focus on the risk-free asset. By the pricing condition in the definition of a state-price deflator, the process G_f^ζ defined by $G_{ft}^\zeta=\zeta_t\exp\{\int_0^t r_u^f\,du\}$ has to be a martingale, i.e. have a zero drift. By Itô's Lemma,

$$dG_{ft}^\zeta=G_{ft}^\zeta\left[(-m_t+r_t^f)\,dt-\boldsymbol{\lambda}_t^\top\,d\boldsymbol{z}_t\right]$$

so we conclude that $m_t=r_t^f$, i.e. the relative drift of a state-price deflator is equal to the negative of the continuously compounded short-term risk-free interest rate.

Next, for any risky asset i the process G_i^ζ defined by (4.31) must be a martingale. From Itô's Lemma and the dynamics of P_i and ζ given in (3.3) and (4.34), we get

$$\begin{aligned}dG_{it}^\zeta&=\delta_{it}P_{it}\zeta_t\,dt+\zeta_t\,dP_{it}+P_{it}d\zeta_t+(d\zeta_t)(dP_{it})\\&=P_{it}\zeta_t\left[(\mu_{it}+\delta_{it}-m_t-\boldsymbol{\sigma}_{it}^\top\boldsymbol{\lambda}_t)\,dt+(\boldsymbol{\lambda}_t+\boldsymbol{\sigma}_{it})^\top\,d\boldsymbol{z}_t\right].\end{aligned}$$

With a risk-free asset, we know that $m_t=r_t^f$, so setting the drift equal to zero, we conclude that the equation

$$\mu_{it}+\delta_{it}-r_t^f=\boldsymbol{\sigma}_{it}^\top\boldsymbol{\lambda}_t\tag{4.35}$$

must hold for any asset i. This is equivalent to (4.33). In compact form, the condition on $\boldsymbol{\lambda}$ can be written

$$\boldsymbol{\mu}_t+\boldsymbol{\delta}_t-r_t^f\mathbf{1}=\underline{\sigma}_t\boldsymbol{\lambda}_t.\tag{4.36}$$

A (nice) process $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_t)$ satisfying this equation is called a *market price of risk*. If the price of the i'th asset is only sensitive to the j'th exogenous shock, Equation (4.35) reduces to

$$\mu_{it} + \delta_{it} - r_t^f = \sigma_{ijt}\lambda_{jt},$$

implying that

$$\lambda_{jt} = \frac{\mu_{it} + \delta_{it} - r_t^f}{\sigma_{ijt}}.$$

Therefore, λ_{jt} is the compensation in terms of excess expected return per unit of risk stemming from the j'th exogenous shock. This explains the name market price of risk. Summing up, the dynamics of a state-price deflator is of the form

$$d\zeta_t = -\zeta_t \left[r_t^f \, dt + \boldsymbol{\lambda}_t^\top \, d\boldsymbol{z}_t \right], \tag{4.37}$$

where $\boldsymbol{\lambda}$ is a market price of risk.

4.3 Properties of state-price deflators

After defining state-price deflators, two questions arise naturally: When does a state-price deflator exist? When is it unique? We will answer these questions in the two following subsections.

4.3.1 Existence

Here is the answer to the existence question:

Theorem 4.1 *A state-price deflator exists if and only if prices admit no arbitrage.*

Since we have already concluded that we should only consider no-arbitrage prices, we can safely assume the existence of a state-price deflator.

Let us take the easy part of the proof first: if a state-price deflator exists, prices do not admit arbitrage. Let us just think of the one-period framework so that the state-price deflator is a strictly positive random variable and the price of a random dividend of D_i is $P_i = \mathrm{E}[\zeta D_i]$. If D_i is non-negative in all states, it is clear that ζD_i will be non-negative in all states and, consequently, the expectation of ζD_i will be non-negative. This rules out arbitrage of type (i), cf. the definition of arbitrage in Section 3.4. If, furthermore, the set of states $A = \{\omega \in \Omega : D_i(\omega) > 0\}$ has strictly positive probability, it is clear that ζD_i will be strictly positive on a set of strictly positive probability and otherwise non-negative, so the expectation of ζD_i must be strictly positive. This rules out type (ii) arbitrage. The same argument applies to the discrete-time framework. In the continuous-time setting the argument should be slightly adjusted in order to incorporate the lower bound on the value process that will rule out doubling strategies. It is not terribly difficult, but involves local martingales and super-martingales which we will not discuss here. The interested reader is referred to Duffie (2001, p. 105).

How can we show the other and more important implication that no arbitrage guarantees the existence of a state-price vector or state-price deflator? In Chapter 6 we will do that by constructing a state-price deflator from the solution to the utility maximization problem of any individual. The solution will only exist in absence of arbitrage. It is possible to prove the existence of a state-price

deflator without formally introducing individuals and solving their utility maximization problems. The alternative proof is based on the so-called *Separating Hyperplane Theorem* and involves no economics.

In the one-period framework with a finite state space $\Omega = \{1, 2, \ldots, S\}$, the alternative argument goes as follows. A given portfolio $\boldsymbol{\theta}$ has an initial dividend of $-P^{\boldsymbol{\theta}}$ and a terminal dividend given by the S-dimensional vector $\boldsymbol{D}^{\boldsymbol{\theta}}$. Let M denote the set of all $(S+1)$ dimensional pairs $(-P^{\boldsymbol{\theta}}, \boldsymbol{D}^{\boldsymbol{\theta}})$ that is generated by all portfolios, i.e. all $\boldsymbol{\theta} \in \mathbb{R}^I$. Observe that M is a closed and convex subset of $L \equiv \mathbb{R} \times \mathbb{R}^S$. Let K be the positive orthant of L, i.e. $K \equiv \mathbb{R}_+ \times \mathbb{R}^S_+$, where $\mathbb{R}_+ = [0, \infty)$. K is also a closed and convex subset of L. Note that there is no arbitrage if and only if the only common element of K and M is the zero element $(0, \mathbf{0})$.

Assume no arbitrage, i.e. $K \cap M = \{(0, \mathbf{0})\}$. By the Separating Hyperplane Theorem (see, e.g., Rockafellar 1970) there exists a non-zero linear functional $F : L \to \mathbb{R}$ with the property that $F(z) = 0$ for all $z \in M$ and $F(x) > 0$ for all non-zero $x \in K$. Hence, we can find a strictly positive φ_0 in $\mathbb{R}$ and an S-dimensional vector $\boldsymbol{\varphi}$ with strictly positive elements such that $F(d_0, \boldsymbol{d}) = \varphi_0 d_0 + \boldsymbol{\varphi} \cdot \boldsymbol{d}$ for all $(d_0, \boldsymbol{d})$ in L. Since $(-P^{\boldsymbol{\theta}}, \boldsymbol{D}^{\boldsymbol{\theta}}) \in M$ for any portfolio $\boldsymbol{\theta}$, we have that

$$0 = F(-P^{\boldsymbol{\theta}}, \boldsymbol{D}^{\boldsymbol{\theta}}) = -\varphi_0 P^{\boldsymbol{\theta}} + \boldsymbol{\varphi} \cdot \boldsymbol{D}^{\boldsymbol{\theta}},$$

and hence

$$P^{\boldsymbol{\theta}} = \frac{1}{\varphi_0} \boldsymbol{\varphi} \cdot \boldsymbol{D}^{\boldsymbol{\theta}}.$$

The vector $\boldsymbol{\psi} = \boldsymbol{\varphi}/\varphi_0$ is therefore a state-price vector.

Just assuming that prices obey the law of one price would give us a vector $\boldsymbol{\psi}$ satisfying $\boldsymbol{\psi} \cdot \boldsymbol{D}_i = P_i$ for all i. Imposing the stronger no-arbitrage condition ensures us that we can find a strictly positive vector $\boldsymbol{\psi}$ with that property, i.e. a state-price vector. Dividing by state probabilities, $\zeta_\omega = \psi_\omega / p_\omega$, we obtain a state-price deflator.

Example 4.4 Consider again the market in Example 3.1 and ignore asset 4, which is redundant. Suppose the market prices of the three remaining assets are 1.1, 2.2, and 0.6, respectively. Can you find a state-price vector $\boldsymbol{\psi}$? The only candidate is the solution to the equation system $\underline{\underline{D}}\boldsymbol{\psi} = \boldsymbol{P}$, i.e.

$$\boldsymbol{\psi} = \left(\underline{\underline{D}}\right)^{-1} \boldsymbol{P} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 4 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1.1 \\ 2.2 \\ 0.6 \end{pmatrix} = \begin{pmatrix} -0.1 \\ 0.2 \\ 1 \end{pmatrix},$$

which is not strictly positive. Hence there is no state-price vector for this market. Then there must be an arbitrage, but where? We can see that the three assets are priced such that the implicit value of an Arrow-Debreu asset for state 1 is negative. The portfolio of the three assets that replicates this Arrow-Debreu asset is given by

$$\boldsymbol{\theta} = \left(\underline{\underline{D}}^\top\right)^{-1} \boldsymbol{e}_1 = \begin{pmatrix} 0.2 & 1.6 & -0.8 \\ -0.2 & -0.6 & 0.8 \\ 0.2 & -0.4 & 0.2 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.2 \\ -0.2 \\ 0.2 \end{pmatrix},$$

which indeed has a price of $0.2 \cdot 1.1 - 0.2 \cdot 2.2 + 0.2 \cdot 0.6 = -0.1$. You get 0.1 today for an asset that pays you one if state 1 is realized and nothing in other states. This is clearly an arbitrage. □

4.3.2 Uniqueness

Here is a general result on the uniqueness of state-price deflators:

Theorem 4.2 *Assume prices admit no arbitrage. Then there is a unique state-price deflator if and only if the market is complete. If the market is incomplete, several state-price deflators exist.*

In a one-period framework the theorem can be explained as follows. In the absence of arbitrage, some state-price deflator ζ exists. If ε is a random variable with $\mathrm{E}[\varepsilon D_i] = 0$ for all i, then

$$\mathrm{E}[(\zeta + \varepsilon)D_i] = \mathrm{E}[\zeta D_i] + \mathrm{E}[\varepsilon D_i] = \mathrm{E}[\zeta D_i] = P_i.$$

If ε is strictly positive with finite variance, then $\zeta + \varepsilon$ will be a valid state-price deflator. When can we find such an ε? If the market is complete, the dividends of the basic assets span all random variables so it will be impossible to find a random variable ε not identically equal to zero with $\mathrm{E}[\varepsilon D_i] = 0$. Therefore ζ must be the only state-price deflator.

Let us be a bit more precise. Consider, for simplicity, a one-period framework with a finite state space and look for a state-price vector $\boldsymbol{\psi}$, i.e. a strictly positive solution to the equation system $\underline{\underline{D}}\boldsymbol{\psi} = \boldsymbol{P}$. If the market is complete and there are no redundant assets, the dividend matrix $\underline{\underline{D}}$ is an $S \times S$ non-singular matrix so the only solution to the equation system is

$$\boldsymbol{\psi}^* = \underline{\underline{D}}^{-1}\boldsymbol{P}.$$

If there are redundant assets, they will be uniquely priced by no-arbitrage so it will be sufficient to look for solutions to $\underline{\underline{\hat{D}}}\boldsymbol{\psi} = \hat{\boldsymbol{P}}$, where $\underline{\underline{\hat{D}}}$ is the dividend matrix and $\hat{\boldsymbol{P}}$ the price vector of the non-redundant assets. In the case of a complete market the matrix $\underline{\underline{\hat{D}}}$ is non-singular. The unique solution to the equation system is then

$$\boldsymbol{\psi}^* = \underline{\underline{\hat{D}}}^{-1}\hat{\boldsymbol{P}}. \tag{4.38}$$

Whether the market is complete or not, the S-dimensional vector

$$\boldsymbol{\psi}^* = \underline{\underline{\hat{D}}}^{\top}\left(\underline{\underline{\hat{D}}}\,\underline{\underline{\hat{D}}}^{\top}\right)^{-1}\hat{\boldsymbol{P}} \tag{4.39}$$

is a solution since

$$\underline{\underline{\hat{D}}}\boldsymbol{\psi}^* = \underline{\underline{\hat{D}}}\,\underline{\underline{\hat{D}}}^{\top}\left(\underline{\underline{\hat{D}}}\,\underline{\underline{\hat{D}}}^{\top}\right)^{-1}\hat{\boldsymbol{P}} = \hat{\boldsymbol{P}}.$$

If $\boldsymbol{\psi}^*$ is strictly positive, we can therefore conclude that it will be a state-price vector. Note that $\boldsymbol{\psi}^*$ is in fact exactly equal to the dividend generated by the portfolio $\hat{\boldsymbol{\theta}}^* = \left(\underline{\underline{\hat{D}}}\,\underline{\underline{\hat{D}}}^{\top}\right)^{-1}\hat{\boldsymbol{P}}$, i.e. if $\boldsymbol{\psi}^*$ is strictly positive, it is a state-price vector in the set of marketed dividend vectors.

In the special case of a complete and arbitrage-free market the elements of $\boldsymbol{\psi}^*$ will be strictly positive. Why? Since the market is complete, we can construct an Arrow-Debrey asset for any state. The dividend vector of the Arrow-Debreu asset for state ω is $\boldsymbol{e}_\omega = (0, \ldots, 0, 1, 0, \ldots, 0)^{\top}$, where the 1 is the ω'th element of the vector. The price of this portfolio is $\boldsymbol{\psi}^* \cdot \boldsymbol{e}_\omega = \psi^*_\omega$. To avoid arbitrage, ψ^*_ω must be strictly positive. This argument works for all $\omega = 1, \ldots, S$. Hence, $\boldsymbol{\psi}^*$ is a state-price vector if the market is complete.

In fact, if the market is complete, $\boldsymbol{\psi}^*$ is the *only* state-price vector. Suppose that both $\boldsymbol{\psi}^*$ and $\boldsymbol{\psi}^{**}$ are state-price vectors. That $\boldsymbol{\psi}^*$ is a state-price vector implies that the price of the Arrow-Debreu asset for state ω is $\boldsymbol{\psi}^* \cdot \boldsymbol{e}_\omega = \psi^*_\omega$. That $\boldsymbol{\psi}^{**}$ is a state-price vector implies that the price

of the Arrow-Debreau asset for state ω is $\boldsymbol{\psi}^{**} \cdot \boldsymbol{e}_\omega = \psi^{**}_\omega$. Hence, we can conclude that $\psi^*_\omega = \psi^{**}_\omega$. This argument works for any ω. Therefore the two vectors $\boldsymbol{\psi}^*$ and $\boldsymbol{\psi}^{**}$ are identical.

Now let us turn to state-price deflators. Due to the one-to-one correspondence between state-price vectors and state-price deflators we expect to find similar results. Define the S-dimensional vector $\boldsymbol{\zeta}^*$ by

$$\boldsymbol{\zeta}^* = \underline{\underline{\hat{D}}}^\top \left(\mathrm{E}\left[\hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \right] \right)^{-1} \hat{\boldsymbol{P}}. \tag{4.40}$$

To see the meaning of this, let us for simplicity assume that none of the basic assets are redundant so that $\boldsymbol{\zeta}^* = \underline{\underline{D}}^\top \left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top \right] \right)^{-1} \boldsymbol{P}$. Recall that $\boldsymbol{D}$ is the I-dimensional random variable for which the i'th component is given by the random dividend of asset i. Hence, $\boldsymbol{D}\boldsymbol{D}^\top$ is an $I \times I$ matrix of random variables with the (i,j)'th entry given by $D_i D_j$, i.e. the product of the random dividend of asset i and the random dividend of asset j. The expectation of a matrix of random variables is equal to the matrix of expectations of the individual random variables. So $\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top \right]$ is also an $I \times I$ matrix. For the general case we see from the definition that $\boldsymbol{\zeta}^*$ is in fact the dividend vector generated by the portfolio

$$\hat{\boldsymbol{\theta}}^* = \left(\mathrm{E}\left[\hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \right] \right)^{-1} \hat{\boldsymbol{P}}$$

of the non-redundant assets. We can think of $\boldsymbol{\zeta}^*$ as a random variable ζ^* given by

$$\zeta^* = \hat{\boldsymbol{D}}^\top \left(\mathrm{E}\left[\hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \right] \right)^{-1} \hat{\boldsymbol{P}}. \tag{4.41}$$

We can see that

$$\mathrm{E}\left[\hat{\boldsymbol{D}} \zeta^* \right] = \mathrm{E}\left[\hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \left(\mathrm{E}\left[\hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \right] \right)^{-1} \hat{\boldsymbol{P}} \right] = \mathrm{E}\left[\hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \right] \left(\mathrm{E}\left[\hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \right] \right)^{-1} \hat{\boldsymbol{P}} = \hat{\boldsymbol{P}}.$$

It follows that ζ^* is a state-price deflator if it takes only strictly positive values. It can be shown that no other state-price deflator can be written as the dividend of a portfolio of traded assets. In a complete market, ζ^* will be a state-price deflator and it will be unique.

Recall that there is a one-to-one relation between state-price vectors and state-price deflators. In general ζ^* is not the state-price deflator associated with ψ^*. However, this will be so if the market is complete. To see this, let $\mathrm{diag}(\boldsymbol{p})$ denote the diagonal $S \times S$ matrix with the state probabilities along the diagonal and zeros away from the diagonal. In general, $\mathrm{E}\left[\hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \right] = \underline{\underline{\hat{D}}}\, \mathrm{diag}(\boldsymbol{p}) \underline{\underline{\hat{D}}}^\top$ and the state-price vector associated with a given state-price deflator $\boldsymbol{\zeta}$ is $\mathrm{diag}(\boldsymbol{p})\boldsymbol{\zeta}$, cf. (4.17). With a complete market, $\underline{\underline{\hat{D}}}$ is a non-singular $S \times S$ matrix so

$$\begin{aligned} \boldsymbol{\zeta}^* &= \underline{\underline{\hat{D}}}^\top \left(\mathrm{E}\left[\hat{\boldsymbol{D}} \hat{\boldsymbol{D}}^\top \right] \right)^{-1} \hat{\boldsymbol{P}} = \underline{\underline{\hat{D}}}^\top \left[\underline{\underline{\hat{D}}}\, \mathrm{diag}(\boldsymbol{p}) \underline{\underline{\hat{D}}}^\top \right]^{-1} \hat{\boldsymbol{P}} \\ &= \underline{\underline{\hat{D}}}^\top \left(\underline{\underline{\hat{D}}}^\top \right)^{-1} [\mathrm{diag}(\boldsymbol{p})]^{-1} \underline{\underline{\hat{D}}}^{-1} \hat{\boldsymbol{P}} = [\mathrm{diag}(\boldsymbol{p})]^{-1} \underline{\underline{\hat{D}}}^{-1} \hat{\boldsymbol{P}}, \end{aligned}$$

and the state-price vector associated with $\boldsymbol{\zeta}^*$ is

$$\mathrm{diag}(\boldsymbol{p})\boldsymbol{\zeta}^* = \underline{\underline{\hat{D}}}^{-1} \hat{\boldsymbol{P}} = \boldsymbol{\psi}^*.$$

If the market is complete and arbitrage-free we have identified the unique state-price vector $\boldsymbol{\psi}^*$ and the unique state-price deflator ζ^*. But it is important to realize the following: $\boldsymbol{\psi}^*$ and ζ^* are defined in terms of the prices of the basic assets. Observing the prices and state-contingent dividends of the basic assets, we can extract the state-price deflator. If you want to compute prices of the basic assets from their state-contingent dividends, $\boldsymbol{\psi}^*$ and ζ^* are not useful. We

need to add more structure to link the state-price vector and deflator to other variables, e.g. the consumption and portfolio decisions of the individuals in the economy. This is what concrete asset pricing models have to do. Further discussions of ζ^* and what we can learn about state prices from observed prices or returns follow later in this chapter.

Example 4.5 Consider again the complete market first studied in Example 3.1. We ignore asset 4, which is in any case redundant, and let $\underline{\underline{D}}$ be the dividend matrix and $\boldsymbol{P}$ the price vector of the first three assets. If we assume that $\boldsymbol{P} = (0.8, 0.8, 1.5)^\top$, we can compute the unique state-price vector $\boldsymbol{\psi}^*$ as

$$\boldsymbol{\psi}^* = \underline{\underline{D}}^{-1}\boldsymbol{P} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \\ 4 & 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0.8 \\ 0.8 \\ 1.5 \end{pmatrix} = \begin{pmatrix} 0.3 \\ 0.2 \\ 0.3 \end{pmatrix},$$

which is consistent with the results of Example 4.2. The portfolio generating this dividend vector is $\boldsymbol{\theta}^* = (0.14, 0.06, 0.04)^\top$.

Since the market is complete, the unique state-price deflator ζ^* is the one associated with $\boldsymbol{\psi}^*$. From Example 4.2, we have $\zeta_1^* = 0.6$, $\zeta_2^* = 0.8$, $\zeta_3^* = 1.2$. The portfolio generating this dividend is $\boldsymbol{\theta}^* = (0.44, 0.36, 0.04)^\top$.

Suppose now that only assets 1 and 2 were traded with the same prices as above, $P_1 = P_2 = 0.8$. Then

$$\underline{\underline{D}} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 2 \end{pmatrix}, \quad \underline{\underline{D}}\underline{\underline{D}}^\top = \begin{pmatrix} 3 & 3 \\ 3 & 5 \end{pmatrix}, \quad \left(\underline{\underline{D}}\underline{\underline{D}}^\top\right)^{-1} = \begin{pmatrix} \frac{5}{6} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

and we get

$$\boldsymbol{\psi}^* = \underline{\underline{D}}^\top \left(\underline{\underline{D}}\,\underline{\underline{D}}^\top\right)^{-1} \boldsymbol{P} = \begin{pmatrix} \frac{4}{15} \\ \frac{4}{15} \\ \frac{4}{15}, \end{pmatrix},$$

which is strictly positive and therefore a state-price vector. It is the dividend of the portfolio that only consists of $4/15 \approx 0.2667$ units of asset 1. But we can find many other state-price vectors. We have to look for strictly positive solutions $(\psi_1, \psi_2, \psi_3)^\top$ of the two equations

$$\psi_1 + \psi_2 + \psi_3 = 0.8, \quad \psi_2 + 2\psi_3 = 0.8.$$

Subtracting one equation from the other we see that we need to have $\psi_1 = \psi_3$. Any vector of the form $\boldsymbol{\psi} = (a, 0.8 - 2a, a)^\top$ with $0 < a < 0.4$ will be a valid state-price vector. (This includes, of course, the state-price vector in the three-asset market.) All these vectors will generate the same price on any marketed dividend but different prices on non-marketed dividends. For example, the value of the dividend of asset 3, which we now assume is not traded, will be $\boldsymbol{\psi} \cdot (4, 0, 1)^\top = 5a$, which can be anywhere in the interval $(0, 2)$.

Let us compute ζ^* in the two-asset market. The computations needed for $\mathrm{E}[\boldsymbol{D}\boldsymbol{D}^\top]$ are given in Table 4.1. We get

$$\mathrm{E}[\boldsymbol{D}\boldsymbol{D}^\top] = \begin{pmatrix} 1 & 0.75 \\ 0.75 & 1.25 \end{pmatrix}, \quad (\mathrm{E}[\boldsymbol{D}\boldsymbol{D}^\top])^{-1} = \begin{pmatrix} 1.8182 & -1.0909 \\ -1.0909 & 1.4545 \end{pmatrix},$$

$$\boldsymbol{\theta}^* = (\mathrm{E}[\boldsymbol{D}\boldsymbol{D}^\top])^{-1}\boldsymbol{P} = \begin{pmatrix} 0.5818 \\ 0.2909 \end{pmatrix}, \quad \boldsymbol{\zeta}^* = \underline{\underline{D}}^\top\boldsymbol{\theta}^* = \begin{pmatrix} 0.5818 \\ 0.8727 \\ 1.1636 \end{pmatrix}.$$

	state 1	state 2	state 3	
probabilities	0.5	0.25	0.25	
	state-contingent values			expectation
D_1^2	1	1	1	1
$D_1 D_2$	0	1	2	0.75
D_2^2	0	1	4	1.25

Table 4.1: Computation of expectations for $\boldsymbol{\zeta}^*$ in Example 4.5.

Since $\boldsymbol{\zeta}^* > 0$, it is a valid state-price deflator. (Note that this is not the state-price deflator associated with the state-price vector $\boldsymbol{\psi}^*$ computed above.) We have infinitely many state-price deflators for this market. Given any state-price vector $\boldsymbol{\psi} = (a, 0.8 - 2a, a)^\top$ for $0 < a < 0.4$, the associated state-price deflator is given by $\zeta_\omega = \psi_\omega / p_\omega$, i.e.

$$\boldsymbol{\zeta} = \begin{pmatrix} a/0.5 \\ (0.8 - 2a)/0.25 \\ a/0.25 \end{pmatrix} = \begin{pmatrix} 2a \\ 3.2 - 8a \\ 4a \end{pmatrix}.$$

Letting $b = 2a$, any state-price deflator is of the form $\boldsymbol{\zeta} = (b, 3.2 - 4b, 2b)$ for $0 < b < 0.8$. □

In the multi-period discrete-time framework all the above observations and conclusions hold in each period.

Now consider the continuous-time framework and assume that an instantaneously risk-free asset is traded. We have seen above that state-price deflators are closely related to market prices of risk. Whenever we have a market price of risk $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_t)$, i.e. a nice process satisfying

$$\boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r_t^f \mathbf{1} = \underline{\underline{\sigma}}_t \boldsymbol{\lambda}_t, \tag{4.36}$$

then a state-price deflator can be defined by $\zeta_0 = 1$ and

$$d\zeta_t = -\zeta_t \left[r_t^f \, dt + \boldsymbol{\lambda}_t^\top \, d\boldsymbol{z}_t \right], \tag{4.42}$$

or, equivalently,

$$\zeta_t = \exp\left\{ -\int_0^t r_s^f \, ds - \frac{1}{2} \int_0^t \|\boldsymbol{\lambda}_s\|^2 \, ds - \int_0^t \boldsymbol{\lambda}_s^\top \, d\boldsymbol{z}_s \right\}. \tag{4.43}$$

The number of state-price deflators is therefore determined by the number of solutions to (4.36), which again depends on the rank of the matrix $\underline{\underline{\sigma}}_t$.

Suppose the rank of $\underline{\underline{\sigma}}_t$ equals k for all t. If $k < d$, there are several solutions to (4.36). We can write one solution as

$$\boldsymbol{\lambda}_t^* = \hat{\underline{\underline{\sigma}}}_t^\top \left(\hat{\underline{\underline{\sigma}}}_t \hat{\underline{\underline{\sigma}}}_t^\top \right)^{-1} \left(\hat{\boldsymbol{\mu}}_t + \hat{\boldsymbol{\delta}}_t - r_t^f \mathbf{1} \right), \tag{4.44}$$

where $\hat{\underline{\underline{\sigma}}}_t$ is the $k \times d$ matrix obtained from $\underline{\underline{\sigma}}_t$ by removing rows corresponding to redundant assets, i.e. rows that can be written as a linear combination of other rows in the matrix. Similarly, $\hat{\boldsymbol{\mu}}_t$ and $\hat{\boldsymbol{\delta}}_t$ are the k-dimensional vectors that are left after deleting from $\boldsymbol{\mu}_t$ and $\boldsymbol{\delta}_t$, respectively, the elements corresponding to the redundant assets. In the special case where $k = d$, we have

$$\boldsymbol{\lambda}_t^* = \hat{\underline{\underline{\sigma}}}_t^{-1} \left(\hat{\boldsymbol{\mu}}_t + \hat{\boldsymbol{\delta}}_t - r_t^f \mathbf{1} \right).$$

Let ζ^* be the state-price deflator associated with $\boldsymbol{\lambda}^*$, i.e.

$$\zeta_t^* = \exp\left\{-\int_0^t r_s^f\,ds - \frac{1}{2}\int_0^t \|\boldsymbol{\lambda}_s^*\|^2\,ds - \int_0^t (\boldsymbol{\lambda}_s^*)^\top\,d\boldsymbol{z}_s\right\}. \tag{4.45}$$

In the one-period framework the (candidate) state-price deflator ζ^* was shown to be the dividend of some portfolio of traded assets. What about the continuous-time framework? Consider the self-financing trading strategy given by the fractions of wealth $\boldsymbol{\pi}_t^* = \left(\hat{\underline{\underline{\sigma}}}_t\hat{\underline{\underline{\sigma}}}_t^\top\right)^{-1}\left(\hat{\boldsymbol{\mu}}_t + \hat{\boldsymbol{\delta}}_t - r_t^f\mathbf{1}\right)$ in the non-redundant assets and the fraction $1-(\boldsymbol{\pi}_t^*)^\top\mathbf{1}$ in the instantaneously risk-free asset. The dynamics of the value V_t^* of this trading strategy is given by

$$\begin{aligned} dV_t^* &= V_t^*\left[\left(r_t^f + (\boldsymbol{\pi}_t^*)^\top\left(\hat{\boldsymbol{\mu}}_t + \hat{\boldsymbol{\delta}}_t - r_t^f\mathbf{1}\right)\right)dt + (\boldsymbol{\pi}_t^*)^\top\hat{\underline{\underline{\sigma}}}_t\,d\boldsymbol{z}_t\right] \\ &= V_t^*\left[\left(r_t^f + \|\boldsymbol{\lambda}_t^*\|^2\right)dt + (\boldsymbol{\lambda}_t^*)^\top\,d\boldsymbol{z}_t\right]. \end{aligned} \tag{4.46}$$

cf. (3.14). It can be shown that $\boldsymbol{\pi}_t^*$ is the trading strategy with the highest expected continuously compounded growth rate, i.e. the trading strategy maximizing $\mathrm{E}[\ln(V_T^{\boldsymbol{\pi}}/V_0^{\boldsymbol{\pi}})]$, and it is therefore referred to as the *growth-optimal trading strategy*. Consequently, $\boldsymbol{\lambda}_t^*$ defined in (4.44) is the relative sensitivity vector of the value of the growth-optimal trading strategy. One can show that $\zeta_t^* = V_0^*/V_t^*$ (see Exercise 4.11), so we have a state-price deflator defined in terms of the value of a trading strategy.

If $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_t)$ is a nice d-dimensional stochastic process with $\underline{\underline{\sigma}}_t\boldsymbol{\varepsilon}_t = \mathbf{0}$ for all t, then $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_t^* + \boldsymbol{\varepsilon}_t$ defines a market price of risk since

$$\underline{\underline{\sigma}}_t\boldsymbol{\lambda}_t = \underline{\underline{\sigma}}_t(\boldsymbol{\lambda}_t^* + \boldsymbol{\varepsilon}_t) = \underline{\underline{\sigma}}_t\boldsymbol{\lambda}_t^* + \underline{\underline{\sigma}}_t\boldsymbol{\varepsilon}_t = \underline{\underline{\sigma}}_t\boldsymbol{\lambda}_t^* = \boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r_t^f\mathbf{1}.$$

If the market is incomplete, it will be possible to find such an $\boldsymbol{\varepsilon}_t$ and, hence, there will be more than one state-price deflator.

An example of an incomplete market is a market where the traded assets are only immediately affected by $j < d$ of the d exogenous shocks. Decomposing the d-dimensional standard Brownian motion $\boldsymbol{z}$ into $(\boldsymbol{Z}, \hat{\boldsymbol{Z}})$, where $\boldsymbol{Z}$ is j-dimensional and $\hat{\boldsymbol{Z}}$ is $(d-j)$-dimensional, the dynamics of the traded risky assets can be written as

$$d\boldsymbol{P}_t = \mathrm{diag}(\boldsymbol{P}_t)\left[\boldsymbol{\mu}_t\,dt + \underline{\underline{\sigma}}_t\,d\boldsymbol{Z}_t\right].$$

For example, the dynamics of r_t^f, $\boldsymbol{\mu}_t$, or $\underline{\underline{\sigma}}_t$ may be affected by the non-traded risks $\hat{\boldsymbol{Z}}$, representing non-hedgeable risk in interest rates, expected returns, and volatilities and correlations, respectively. Or other variables important for the investor, e.g. his labor income, may be sensitive to $\hat{\boldsymbol{Z}}$. Let us assume for simplicity that there are j risky assets and the $j \times j$ matrix $\underline{\underline{\sigma}}_t$ is non-singular (i.e. there are no redundant assets). Then we can define a unique market price of risk associated with the traded risks by the j-dimensional vector

$$\boldsymbol{\Lambda}_t = \left(\underline{\underline{\sigma}}_t\right)^{-1}(\boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r_t\mathbf{1}),$$

but for any well-behaved $(d-j)$-dimensional process $\hat{\boldsymbol{\Lambda}}$, the process $\boldsymbol{\lambda} = (\boldsymbol{\Lambda}, \hat{\boldsymbol{\Lambda}})$ will be a market price of risk for all risks. Each choice of $\hat{\boldsymbol{\Lambda}}$ generates a valid market price of risk process and hence a valid state-price deflator.

4.3.3 Convex combinations of state-price deflators

A convex combination of some objects (such as vectors, random variables, stochastic process, etc.) $x_1, x_2, \ldots, x_L$ is given by

$$x = \sum_{l=1}^{L} \alpha_l x_l,$$

where $\alpha_1, \ldots, \alpha_L$ are positive constants summing up to one. The following theorem says that if you have a number of state-price deflators, any convex combination of those deflators will also be a state-price deflator. This is true both in the one-period, the discrete-time, and the continuous-time framework. In particular, it tells you that once you have two different state-price deflators you can generate infinitely many state-price deflators. The proof of the theorem is left as Exercise 4.3.

Theorem 4.3 *If $\zeta_1, \ldots, \zeta_L$ are state-price deflators, and $\alpha_1, \ldots, \alpha_L > 0$ with $\sum_{l=1}^{L} \alpha_l = 1$, then the convex combination*

$$\zeta = \sum_{l=1}^{L} \alpha_l \zeta_l$$

is also a state-price deflator.

4.3.4 The candidate deflator $\boldsymbol{\zeta}^*$ and the Hansen-Jagannathan bound

The candidate deflator $\boldsymbol{\zeta}^*$ from the one-period framework is interesting for empirical studies so it is worthwhile to study it more closely. Let us compute the return associated with the dividend ζ^*. This return turns out to be important in later sections. For notational simplicity suppose that no assets are redundant so that

$$\zeta^* = \boldsymbol{D}^\top \left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} \boldsymbol{P}, \quad \boldsymbol{\theta}^* = \left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} \boldsymbol{P}.$$

Let us first rewrite ζ^* and $\boldsymbol{\theta}^*$ in terms of the gross returns of the assets instead of prices and dividends. Using (3.2), we get

$$\begin{aligned}
\left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} &= \left(\mathrm{E}\left[\mathrm{diag}(\boldsymbol{P})\boldsymbol{R}\boldsymbol{R}^\top \mathrm{diag}(\boldsymbol{P})\right]\right)^{-1} \\
&= \left(\mathrm{diag}(\boldsymbol{P})\, \mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right] \mathrm{diag}(\boldsymbol{P})\right)^{-1} \\
&= \left[\mathrm{diag}(\boldsymbol{P})\right]^{-1} \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \left[\mathrm{diag}(\boldsymbol{P})\right]^{-1}, \qquad (4.47)
\end{aligned}$$

using the facts that prices are non-random and that $(AB)^{-1} = B^{-1}A^{-1}$ for non-singular matrices A and B. Consequently,

$$\boldsymbol{\theta}^* = \left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} \boldsymbol{P} = \left[\mathrm{diag}(\boldsymbol{P})\right]^{-1} \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1},$$

applying that $\left[\mathrm{diag}(\boldsymbol{P})\right]^{-1}\boldsymbol{P} = \mathbf{1}$, and

$$\begin{aligned}
\zeta^* &= \boldsymbol{D}^\top \left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} \boldsymbol{P} \\
&= \boldsymbol{D}^\top \left[\mathrm{diag}(\boldsymbol{P})\right]^{-1} \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1} \\
&= \left(\left[\mathrm{diag}(\boldsymbol{P})\right]^{-1} \boldsymbol{D}\right)^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1} \\
&= \boldsymbol{R}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1} \\
&= \mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \boldsymbol{R},
\end{aligned}$$

using (3.2) and various rules from matrix algebra. The portfolio weight vector is obtained by substituting $\boldsymbol{\theta}^*$ into (3.6). Since

$$\operatorname{diag}(\boldsymbol{P})\boldsymbol{\theta}^* = \operatorname{diag}(\boldsymbol{P})[\operatorname{diag}(\boldsymbol{P})]^{-1} \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1} = \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1},$$

we get

$$\boldsymbol{\pi}^* = \frac{\left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1}}{\mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1}}. \tag{4.48}$$

The gross return on this portfolio is

$$R^* = (\boldsymbol{\pi}^*)^\top \boldsymbol{R} = \frac{\mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \boldsymbol{R}}{\mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1}}. \tag{4.49}$$

This is the gross return corresponding to the dividend ζ^*.

We can also compute R^* directly as ζ^* divided by the price of ζ^* (well-defined since it is a dividend), i.e. $R^* = \zeta^*/P(\zeta^*)$. We can rewrite ζ^* as

$$\zeta^* = \mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \boldsymbol{R},$$

and the price of ζ^* is

$$\begin{aligned} P(\zeta^*) = \mathrm{E}[\zeta^*\zeta^*] &= \mathrm{E}\left[\boldsymbol{P}^\top \left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} \boldsymbol{D}\boldsymbol{D}^\top \left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} \boldsymbol{P}\right] \\ &= \boldsymbol{P}^\top \left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} \boldsymbol{P} = \mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1}. \end{aligned}$$

Dividing ζ^* by $P(\zeta^*)$ we get (4.49).

In Exercise 4.4 you are asked to show the properties collected in the following lemma:

Lemma 4.1 *R^* has the following properties:*

$$\mathrm{E}[R^*] = \frac{\mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathrm{E}[\boldsymbol{R}]}{\mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1}} \tag{4.50}$$

$$\mathrm{E}[(R^*)^2] = \frac{1}{\mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathbf{1}} = \frac{1}{P(\zeta^*)}, \tag{4.51}$$

$$\mathrm{E}[R^* R^i] = \mathrm{E}[(R^*)^2], \quad i = 1, \ldots, N. \tag{4.52}$$

In particular,

$$\frac{\mathrm{E}[R^*]}{\mathrm{E}[(R^*)^2]} = \mathbf{1}^\top \left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1} \mathrm{E}[\boldsymbol{R}]. \tag{4.53}$$

Any random variable ζ that satisfies $P_i = \mathrm{E}[\zeta D_i]$ for all assets i can be decomposed as

$$\zeta = \mathrm{E}[\zeta] + (\boldsymbol{P} - \mathrm{E}[\zeta]\,\mathrm{E}[\boldsymbol{D}])^\top \underline{\underline{\Sigma}}_D^{-1} (\boldsymbol{D} - \mathrm{E}[\boldsymbol{D}]) + \varepsilon, \tag{4.54}$$

where $\underline{\underline{\Sigma}}_D = \mathrm{Var}[\boldsymbol{D}]$ and ε is a random variable with $\mathrm{E}[\boldsymbol{D}\varepsilon] = \mathbf{0}$ and $\mathrm{E}[\varepsilon] = 0$. In particular $\mathrm{Cov}[\boldsymbol{D}, \varepsilon] = \mathbf{0}$. So taking variances we get

$$\begin{aligned} \mathrm{Var}[\zeta] &= (\boldsymbol{P} - \mathrm{E}[\zeta]\,\mathrm{E}[\boldsymbol{D}])^\top \underline{\underline{\Sigma}}_D^{-1} \mathrm{Var}\left[\boldsymbol{D} - \mathrm{E}[\boldsymbol{D}]\right] \underline{\underline{\Sigma}}_D^{-1} (\boldsymbol{P} - \mathrm{E}[\zeta]\,\mathrm{E}[\boldsymbol{D}]) + \mathrm{Var}[\varepsilon] \\ &\geq (\boldsymbol{P} - \mathrm{E}[\zeta]\,\mathrm{E}[\boldsymbol{D}])^\top \underline{\underline{\Sigma}}_D^{-1} (\boldsymbol{P} - \mathrm{E}[\zeta]\,\mathrm{E}[\boldsymbol{D}]) \\ &= (\mathbf{1} - \mathrm{E}[\zeta]\,\mathrm{E}[\boldsymbol{R}])^\top \underline{\underline{\Sigma}}^{-1} (\mathbf{1} - \mathrm{E}[\zeta]\,\mathrm{E}[\boldsymbol{R}]), \end{aligned}$$

where $\underline{\underline{\Sigma}} = \mathrm{Var}[\boldsymbol{R}]$. The possible combinations of expectation and standard deviation of state-price deflators form a hyperbolic region in $(\mathrm{E}[\zeta], \sigma[\zeta])$-space. This result is due to Hansen and Jagannathan (1991) and the right-hand side of the above inequality (or the boundary of the hyperbolic region) is called the Hansen-Jagannathan bound. In Exercise 4.5 you are asked to show that ζ^* satisfies

$$\zeta^* = \mathrm{E}[\zeta^*] + (\boldsymbol{P} - \mathrm{E}[\zeta^*]\,\mathrm{E}[\boldsymbol{D}])^\top \underline{\underline{\Sigma}}_D^{-1} (\boldsymbol{D} - \mathrm{E}[\boldsymbol{D}])\,. \tag{4.55}$$

Note that no ε is added on the right-hand side. It follows that ζ^* satisfies the Hansen-Jagannathan bound with equality.

4.4 Nominal and real state-price deflators

It is important to distinguish between real and nominal dividends and prices. A nominal dividend [price] is the dividend [price] in units of a given currency, e.g. US dollars or Euros. The corresponding real dividend [price] is the number of units of consumption goods which can be purchased for the nominal dividend [price]. For simplicity assume that the economy only offers a single consumption good and let F_t denote the price of the good in currency units at time t. (More broadly we can think of F_t as the value of the Consumer Price Index at time t.) A nominal dividend of $\tilde{D}_t$ then corresponds to a real dividend of $D_t = \tilde{D}_t/F_t$. A nominal price of $\tilde{P}_t$ corresponds to a real price of $P_t = \tilde{P}_t/F_t$.

A state-price deflator basically links future dividends to current prices. We can define a nominal state-price deflator so that the basic pricing condition holds for nominal prices and dividends and, similarly, define a real state-price deflator so that the basic pricing condition holds for real prices and dividends. If we continue to indicate nominal quantities by a tilde and real quantities without a tilde, the definitions of state-price deflators given earlier in this chapter characterize real state-price deflators.

Consider a multi-period discrete-time economy where $\zeta = (\zeta_t)$ is a real state-price deflator so that, in particular,

$$P_{it} = \mathrm{E}_t\left[\sum_{s=t+1}^{T} D_{is}\frac{\zeta_s}{\zeta_t}\right],$$

cf. (4.18). Substituting in $P_{it} = \tilde{P}_{it}/F_t$ and $D_{is} = \tilde{D}_{is}/F_s$ and multiplying through by F_t, we obtain

$$\tilde{P}_{it} = \mathrm{E}_t\left[\sum_{s=t+1}^{T} \tilde{D}_{is}\frac{\zeta_s/F_s}{\zeta_t/F_t}\right].$$

Now it is clear that a nominal state-price deflator $\tilde{\zeta} = (\tilde{\zeta}_t)$ should be defined from a real state-price deflator $\zeta = (\zeta_t)$ as

$$\tilde{\zeta}_t = \frac{\zeta_t}{F_t}, \qquad \text{all } t \in \mathcal{T}. \tag{4.56}$$

Then the nominal state-price deflator will link nominal dividends to nominal prices in the same way that a real state-price deflator links real dividends to real prices. This relation also works in the continuous-time framework. Note that the nominal state-price deflator is positive with an initial value of $\tilde{\zeta}_0 = 1/F_0$.

In a discrete-time framework the gross nominal return on asset i between time t and time $t+1$ is $\tilde{R}_{i,t+1} = \left(\tilde{P}_{i,t+1} + \tilde{D}_{i,t+1}\right)/\tilde{P}_{it}$. The link between the gross real return and the gross nominal

return follows from

$$\begin{aligned} R_{i,t+1} &= \frac{P_{i,t+1} + D_{i,t+1}}{P_{it}} = \frac{\tilde{P}_{i,t+1}/F_{t+1} + \tilde{D}_{i,t+1}/F_{t+1}}{\tilde{P}_{it}/F_t} \\ &= \frac{\tilde{P}_{i,t+1} + \tilde{D}_{i,t+1}}{\tilde{P}_{it}} \frac{F_t}{F_{t+1}} = \tilde{R}_{i,t+1} \frac{F_t}{F_{t+1}}. \end{aligned} \tag{4.57}$$

In terms of the net rates of return $r_{i,t+1} = R_{i,t+1} - 1$, $\tilde{r}_{i,t+1} = \tilde{R}_{i,t+1} - 1$, and the percentage inflation rate $\varphi_{t+1} = F_{t+1}/F_t - 1$ we have

$$1 + r_{i,t+1} = \frac{1 + \tilde{r}_{i,t+1}}{1 + \varphi_{t+1}},$$

which implies that

$$r_{i,t+1} = \frac{1 + \tilde{r}_{i,t+1}}{1 + \varphi_{t+1}} - 1 = \frac{\tilde{r}_{i,t+1} - \varphi_{t+1}}{1 + \varphi_{t+1}} \approx \tilde{r}_{i,t+1} - \varphi_{t+1}. \tag{4.58}$$

The above equations show how to obtain real returns from nominal returns and inflation. Given a time series of nominal returns and inflation, it is easy to compute the corresponding time series of real returns.

The realized gross inflation rate F_{t+1}/F_t is generally not known in advance. Therefore the real return on a nominally risk-free asset is generally stochastic (and conversely). The link between the nominally risk-free gross return $\tilde{R}_t^f$ and the real risk-free gross return R_t^f is

$$\begin{aligned} \frac{1}{\tilde{R}_t^f} &= \mathrm{E}_t\left[\frac{\tilde{\zeta}_{t+1}}{\tilde{\zeta}_t}\right] = \mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\frac{F_t}{F_{t+1}}\right] \\ &= \mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}\right]\mathrm{E}_t\left[\frac{F_t}{F_{t+1}}\right] + \mathrm{Cov}_t\left[\frac{\zeta_{t+1}}{\zeta_t}, \frac{F_t}{F_{t+1}}\right] \\ &= \frac{1}{R_t^f}\mathrm{E}_t\left[\frac{F_t}{F_{t+1}}\right] + \mathrm{Cov}_t\left[\frac{\zeta_{t+1}}{\zeta_t}, \frac{F_t}{F_{t+1}}\right]. \end{aligned}$$

We can obtain a more elegant expression in a continuous-time framework. Let $\zeta = (\zeta_t)$ denote a real state-price deflator, which evolves over time according to

$$d\zeta_t = -\zeta_t\left[r_t^f\, dt + \boldsymbol{\lambda}_t^\top\, d\boldsymbol{z}_t\right],$$

where $r^f = (r_t^f)$ is the short-term real interest rate and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_t)$ is the market price of risk. Assume that the dynamics of the price of the consumption good can be written as

$$dF_t = F_t\left[\mu_{\varphi t}\, dt + \boldsymbol{\sigma}_{\varphi t}^\top\, d\boldsymbol{z}_t\right]. \tag{4.59}$$

We can interpret $\varphi_{t+dt} = dF_t/F_t$ as the realized inflation rate over the next instant, $\mu_{\varphi t} = \mathrm{E}_t[\varphi_{t+dt}]$ as the expected inflation rate, and $\boldsymbol{\sigma}_{\varphi t}$ as the sensitivity vector of the inflation rate.

Consider now a nominal bank account which over the next instant promises a risk-free monetary return represented by the nominal short-term interest rate $\tilde{r}_t^f$. If we let $\tilde{N}_t$ denote the time t dollar value of such an account, we have that

$$d\tilde{N}_t = \tilde{r}_t^f \tilde{N}_t\, dt.$$

The real price of this account is $N_t = \tilde{N}_t/F_t$, since this is the number of units of the consumption good that has the same value as the account. An application of Itô's Lemma implies a real price dynamics of

$$dN_t = N_t\left[\left(\tilde{r}_t^f - \mu_{\varphi t} + \|\boldsymbol{\sigma}_{\varphi t}\|^2\right) dt - \boldsymbol{\sigma}_{\varphi t}^\top\, d\boldsymbol{z}_t\right]. \tag{4.60}$$

Note that the real return on this instantaneously nominally risk-free asset, dN_t/N_t, is risky. Since the percentage sensitivity vector is given by $-\boldsymbol{\sigma}_{\varphi t}$, the expected return is given by the real short rate plus $-\boldsymbol{\sigma}_{\varphi t}^\top \boldsymbol{\lambda}_t$. Comparing this with the drift term in the equation above, we have that

$$\tilde{r}_t^f - \mu_{\varphi t} + \|\boldsymbol{\sigma}_{\varphi t}\|^2 = r_t^f - \boldsymbol{\sigma}_{\varphi t}^\top \boldsymbol{\lambda}_t.$$

Consequently the nominal short-term interest rate is given by

$$\tilde{r}_t^f = r_t^f + \mu_{\varphi t} - \|\boldsymbol{\sigma}_{\varphi t}\|^2 - \boldsymbol{\sigma}_{\varphi t}^\top \boldsymbol{\lambda}_t, \tag{4.61}$$

i.e. the nominal short rate is equal to the real short rate plus the expected inflation rate minus the variance of the inflation rate minus a risk premium. The presence of the last two terms invalidates the Fisher relation, which says that the nominal interest rate is equal to the sum of the real interest rate and the expected inflation rate. The Fisher hypothesis will hold if and only if the inflation rate is instantaneously risk-free. In Chapter 10 we will discuss the link between real and nominal interest rates and yields in more detail.

Individuals should primarily be concerned about real values since, in the end, they should care about the number of goods they can consume. Therefore, most theoretical asset pricing models make predictions about expected real returns.

4.5 A preview of alternative formulations

The previous sections show that a state-price deflator is a good way to represent the market-wide pricing mechanism in a financial market. Paired with characteristics of any individual asset, the state-price deflator leads to the price of the asset. This section shows that we can capture the same information in other ways. The alternative representations can be preferable for some specific purposes and we will return to them in later chapters. Here we will only give a preview. For simplicity we keep the discussion in a one-period framework.

4.5.1 Risk-neutral probabilities

Suppose that a risk-free dividend can be constructed and that it provides a gross return of R^f. A probability measure $\mathbb{Q}$ is called a risk-neutral probability measure if the following conditions are satisfied:

(i) $\mathbb{P}$ and $\mathbb{Q}$ are equivalent, i.e. attach zero probability to the same events;

(ii) the random variable $d\mathbb{Q}/d\mathbb{P}$ (explained below) has finite variance;

(iii) the price of any asset $i = 1, \ldots, I$ is given by

$$P_i = \mathrm{E}^{\mathbb{Q}}\left[(R^f)^{-1} D_i\right] = (R^f)^{-1}\,\mathrm{E}^{\mathbb{Q}}\left[D_i\right], \tag{4.62}$$

i.e. the price of any asset equals the expected discounted dividend using the risk-free interest rate as the discount rate and the risk-neutral probabilities when computing the expectation.

The risk-free return is not random and can therefore be moved in and out of expectations as in the above equation. Given the return (or, equivalently, the price) of the risk-free asset, all the market-wide pricing information is captured by a risk-neutral probability measure.

In the case of a finite state space $\Omega = \{1, 2, \ldots, S\}$, a probability measure $\mathbb{Q}$ is fully characterized by the state probabilities $q_\omega = \mathbb{Q}(\omega)$. Since we have assumed that the real-world probability measure $\mathbb{P}$ is such that $p_\omega > 0$ for all ω, equivalence between $\mathbb{P}$ and $\mathbb{Q}$ demands that $q_\omega > 0$ for all ω. With finite Ω the pricing equation in (iii) can be written as $P_i = R_f^{-1} \sum_{\omega=1}^{S} q_\omega D_{i\omega}$.

Why is $\mathbb{Q}$ called a risk-neutral probability measure? Since the gross return on asset i is $R_i = D_i/P_i$, we can rewrite (4.62) as

$$\mathrm{E}^Q[R_i] = R^f, \tag{4.63}$$

i.e. all assets have an expected return equal to the risk-free return under the risk-neutral probability measure. If all investors were risk-neutral, they would rank assets according to their expected returns only and the market could only be in equilibrium if all assets had the same expected returns. The definition of a risk-neutral probability measure $\mathbb{Q}$ thus implies that asset prices in the real-world are just as they would have been in an economy in which all individuals are risk-neutral and the state probabilities are given by $\mathbb{Q}$. The price adjustments for risk are thus incorporated in the risk-neutral probabilities.

Next, let us explore the link between risk-neutral probability measures and state prices. First, assume a finite state space. Given a state-price vector $\boldsymbol{\psi}$ and the associated state-price deflator ζ, we can define

$$q_\omega = \frac{\psi_\omega}{\sum_{s=1}^{S} \psi_s} = R^f \psi_\omega = R^f p_\omega \zeta_\omega, \quad \omega = 1, \ldots, S.$$

All the q_ω's are strictly positive and sum to one so they define an equivalent probability measure. Furthermore, (4.16) implies that

$$P_i = \boldsymbol{\psi} \cdot \boldsymbol{D}_i = \sum_{\omega=1}^{S} \psi_\omega D_{i\omega} = \sum_{\omega=1}^{S} (R^f)^{-1} q_\omega D_{i\omega} = \mathrm{E}^{\mathbb{Q}} \left[(R^f)^{-1} D_i \right],$$

so $\mathbb{Q}$ is indeed a risk-neutral probability measure. Note that $q_\omega > p_\omega$ if and only if $\zeta_\omega > (R^f)^{-1} = \mathrm{E}[\zeta]$, i.e. if the value of the state-price deflator for state ω is higher than average.

The change of measure from the real-world probability measure $\mathbb{P}$ to the risk-neutral probability measure $\mathbb{Q}$ is given by the ratios $\xi_\omega \equiv q_\omega/p_\omega = R^f \zeta_\omega$. The change of measure is fully captured by the random variable ξ that takes on the value ξ_ω if state ω is realized. This random variable is called the Radon-Nikodym derivative for the change of measure and is often denoted by $d\mathbb{Q}/d\mathbb{P}$. Note that the $\mathbb{P}$-expectation of any Radon-Nikodym derivative $\xi = d\mathbb{Q}/d\mathbb{P}$ must be 1 to ensure that the new measure is a probability measure. This is satisfied by our risk-neutral probability measure since

$$\mathrm{E}^{\mathbb{P}} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \right] = \sum_{\omega=1}^{S} p_\omega \xi_\omega = \sum_{\omega=1}^{S} p_\omega R^f \zeta_\omega = R^f \sum_{\omega=1}^{S} p_\omega \zeta_\omega = 1.$$

When the state space is infinite, state-price deflators still make sense. Given a state-price deflator ζ, we can define a risk-neutral probability measure $\mathbb{Q}$ by the random variable

$$\xi = \frac{d\mathbb{Q}}{d\mathbb{P}} = R^f \zeta.$$

Conversely, given a risk-neutral probability measure $\mathbb{Q}$ and the risk-free gross return R^f, we can define a state-price deflator ζ by

$$\zeta = (R^f)^{-1} \frac{d\mathbb{Q}}{d\mathbb{P}}.$$

In the case of a finite state space, the risk-neutral probability measure is given by $\xi_\omega = q_\omega/p_\omega$, $\omega = 1, \ldots, S$, and we can construct a state-price vector $\boldsymbol{\psi}$ and a state-price deflator $\boldsymbol{\zeta}$ as

$$\psi_\omega = (R^f)^{-1} q_\omega, \qquad \zeta_\omega = (R^f)^{-1} \xi_\omega = (R^f)^{-1} \frac{q_\omega}{p_\omega}, \quad \omega = 1, \ldots, S.$$

We summarize the above observations in the following theorem:

Theorem 4.4 *Assume that a risk-free asset exists. Then there is a one-to-one correspondence between state-price deflators and risk-neutral probability measures.*

Combining this result with Theorems 4.1 and 4.2, we reach the next conclusion.

Theorem 4.5 *Assume that a risk-free asset exists. Prices admit no arbitrage if and only if a risk-neutral probability measure exists. The market is complete if and only if there is a unique risk-neutral probability measure. If the market is complete and arbitrage-free, the unique risk-neutral probability measure $\mathbb{Q}$ is characterized by $d\mathbb{Q}/d\mathbb{P} = R_f \zeta^*$, where ζ^* is given by* (4.41).

Risk-neutral probabilities are especially useful for the pricing of derivative assets. In Chapter 11 we will generalize the definition of risk-neutral probabilities to multi-period settings and we will also define other probability measures that are useful in derivative pricing.

4.5.2 Pricing factors

We will say that a (one-dimensional) random variable x is a pricing factor for the market if there exists some $\alpha, \eta \in \mathbb{R}$ so that

$$\mathrm{E}[R_i] = \alpha + \beta[R_i, x]\eta, \quad i = 1, \ldots, I, \tag{4.64}$$

where the factor-beta of asset i is given by

$$\beta[R_i, x] = \frac{\mathrm{Cov}[R_i, x]}{\mathrm{Var}[x]}. \tag{4.65}$$

The constant η is called the factor risk premium and α the zero-beta return. Due to the linearity of expectations and covariance, (4.64) will also hold for all portfolios of the I assets. Note that if a risk-free asset is traded in the market, it will have a zero factor-beta and, consequently, $\alpha = R^f$.

The relation (4.64) does not directly involve prices. But since the expected gross return is $\mathrm{E}[R_i] = \mathrm{E}[D_i]/P_i$, we have $P_i = \mathrm{E}[D_i]/\mathrm{E}[R_i]$ and hence the equivalent relation

$$P_i = \frac{\mathrm{E}[D_i]}{\alpha + \beta[R_i, x]\eta}. \tag{4.66}$$

The price is equal to the expected dividend discounted by a risk-adjusted rate. You may find this relation unsatisfactory since the price implicitly enters the right-hand side through the return-beta $\beta[R_i, x]$. However, we can define a dividend-beta by $\beta[D_i, x] = \mathrm{Cov}[D_i, x]/\mathrm{Var}[x]$ and inserting $D_i = R_i P_i$ we see that $\beta[D_i, x] = P_i \beta[R_i, x]$. Equation (4.64) now implies that

$$\frac{\mathrm{E}[D_i]}{P_i} = \alpha + \frac{1}{P_i}\beta[D_i, x]\eta$$

so that

$$P_i = \frac{\mathrm{E}[D_i] - \beta[D_i, x]\eta}{\alpha}. \tag{4.67}$$

Think of the numerator as a certainty equivalent of the risky dividend. The current price is the certainty equivalent discounted by the zero-beta return, which is the risk-free return if this exists.

What is the link between pricing factors and state-price deflators? It follows from (4.10) that any state-price deflator ζ itself is a pricing factor. That equation does not require positivity of the state-price deflator, only the pricing condition. Therefore any random variable x that satisfies $P_i = \mathrm{E}[xD_i]$ for all assets works as a pricing factor. More generally, if x is a random variable and a, b are constants so that $P_i = \mathrm{E}[(a+bx)D_i]$ for all assets i, then x is a pricing factor. In particular, whenever we have a state-price deflator of the form $\zeta = a + bx$, we can use x as a pricing factor.

Conversely, if we have a pricing factor x for which the associated zero-beta return α is non-zero, we can find constants a, b so that $\zeta = a + bx$ satisfies the pricing condition $P_i = \mathrm{E}[\zeta D_i]$ for $i = 1, \ldots, I$. In order to see this let η denote the factor risk premium associated with the pricing factor x and define

$$b = -\frac{\eta}{\alpha \operatorname{Var}[x]}, \quad a = \frac{1}{\alpha} - b \operatorname{E}[x].$$

Then $\zeta = a + bx$ works since

$$\begin{aligned}
\mathrm{E}[\zeta R_i] &= a \operatorname{E}[R_i] + b \operatorname{E}[R_i x] \\
&= a \operatorname{E}[R_i] + b \left(\operatorname{Cov}[R_i, x] + \operatorname{E}[R_i] \operatorname{E}[x]\right) \\
&= (a + b \operatorname{E}[x]) \operatorname{E}[R_i] + b \operatorname{Cov}[R_i, x] \\
&= \frac{1}{\alpha} \left(\operatorname{E}[R_i] - \frac{\operatorname{Cov}[R_i, x]}{\operatorname{Var}[x]} \eta\right) \\
&= \frac{1}{\alpha} \left(\operatorname{E}[R_i] - \beta[R_i, x] \eta\right) \\
&= 1
\end{aligned}$$

for any $i = 1, \ldots, I$. Inserting a and b, we get

$$\zeta = a + bx = \frac{1}{\alpha} \left(1 - \frac{\eta}{\operatorname{Var}[x]} \left(x - \operatorname{E}[x]\right)\right).$$

Any pricing factor x gives us a candidate $a + bx$ for a state-price deflator but it will only be a true state-price deflator if it is strictly positive. The fact that we can find a pricing factor for a given market does not imply that the market is arbitrage-free.

Can the pricing factor be the return on some portfolio? No problem! Suppose x is a pricing factor. Look for a portfolio $\boldsymbol{\theta}$ which will generate the dividend as close as possible to x in the sense that it minimizes $\operatorname{Var}[D^{\boldsymbol{\theta}} - x]$. Since

$$\begin{aligned}
\operatorname{Var}\left[D^{\boldsymbol{\theta}} - x\right] &= \operatorname{Var}\left[\boldsymbol{D}^\top \boldsymbol{\theta} - x\right] = \operatorname{Var}\left[\boldsymbol{D}^\top \boldsymbol{\theta}\right] + \operatorname{Var}[x] - 2 \operatorname{Cov}\left[\boldsymbol{D}^\top \boldsymbol{\theta}, x\right] \\
&= \boldsymbol{\theta}^\top \operatorname{Var}[\boldsymbol{D}] \boldsymbol{\theta} + \operatorname{Var}[x] - 2 \boldsymbol{\theta}^\top \operatorname{Cov}[\boldsymbol{D}, x],
\end{aligned}$$

the minimum is obtained for

$$\boldsymbol{\theta} = \left(\operatorname{Var}[\boldsymbol{D}]\right)^{-1} \operatorname{Cov}[\boldsymbol{D}, x].$$

This portfolio is called the factor-mimicking portfolio. Using (3.6) and (3.7), the gross return on this portfolio is

$$R^x = \frac{\boldsymbol{\theta}^\top \operatorname{diag}(\boldsymbol{P}) \boldsymbol{R}}{\boldsymbol{\theta}^\top \operatorname{diag}(\boldsymbol{P}) \mathbf{1}} = \frac{\operatorname{Cov}[\boldsymbol{D}, x]^\top (\operatorname{Var}[\boldsymbol{D}])^{-1} \operatorname{diag}(\boldsymbol{P}) \boldsymbol{R}}{\operatorname{Cov}[\boldsymbol{D}, x]^\top (\operatorname{Var}[\boldsymbol{D}])^{-1} \operatorname{diag}(\boldsymbol{P}) \mathbf{1}} = \frac{\operatorname{Cov}[\boldsymbol{R}, x]^\top (\operatorname{Var}[\boldsymbol{R}])^{-1} \boldsymbol{R}}{\operatorname{Cov}[\boldsymbol{R}, x]^\top (\operatorname{Var}[\boldsymbol{R}])^{-1} \mathbf{1}}. \tag{4.68}$$

The vector of covariances of the returns on the basic assets and the return on the factor-mimicking portfolio is

$$\mathrm{Cov}[\boldsymbol{R}, R^x] = \frac{\mathrm{Cov}[\boldsymbol{R}, x]}{\mathrm{Cov}[\boldsymbol{R}, x]^\top (\mathrm{Var}[\boldsymbol{R}])^{-1}\mathbf{1}}$$

and therefore the beta of asset i with respect to R^x is

$$\begin{aligned}\beta[R_i, R^x] &= \frac{\mathrm{Cov}[R_i, R^x]}{\mathrm{Var}[R^x]} = \frac{\mathrm{Cov}[R_i, x]}{\mathrm{Var}[R^x]\,\mathrm{Cov}[\boldsymbol{R}, x]^\top (\mathrm{Var}[\boldsymbol{R}])^{-1}\mathbf{1}} \\ &= \beta[R_i, x]\frac{\mathrm{Var}[x]}{\mathrm{Var}[R^x]\,\mathrm{Cov}[\boldsymbol{R}, x]^\top (\mathrm{Var}[\boldsymbol{R}])^{-1}\mathbf{1}}.\end{aligned}$$

Consequently, if x is a pricing factor with zero-beta return α and factor risk premium η, then the corresponding factor-mimicking return R^x is a pricing factor with zero-beta return and factor risk premium

$$\hat{\eta} = \frac{\eta\,\mathrm{Var}[R^x]\,\mathrm{Cov}[\boldsymbol{R}, x]^\top (\mathrm{Var}[\boldsymbol{R}])^{-1}\mathbf{1}}{\mathrm{Var}[x]}.$$

In that sense it is not restrictive to look for pricing factors only in the set of returns.

Note that when the factor x itself is a return, then it must satisfy

$$\mathrm{E}[x] = \alpha + \beta[x, x]\eta = \alpha + \eta \quad \Rightarrow \quad \eta = \mathrm{E}[x] - \alpha$$

so that

$$\mathrm{E}[R_i] = \alpha + \beta[R_i, x]\,(\mathrm{E}[x] - \alpha)\,. \tag{4.69}$$

Now it is clear that the standard CAPM simply says that the return on the market portfolio is a pricing factor.

We will discuss factor models in detail in Chapter 9. There we will also allow for multi-dimensional pricing factors.

4.5.3 Mean-variance efficient returns

A portfolio is said to be mean-variance efficient if there is no other portfolio with the same expected return and a lower return variance. The return on a mean-variance efficient portfolio is called a mean-variance efficient return. The mean-variance frontier is the curve in a $(\sigma[R], \mathrm{E}[R])$-plane traced out by all the mean-variance efficient returns.

The analysis of mean-variance efficient portfolios was introduced by Markowitz (1952, 1959) as a tool for investors in making portfolio decisions. Nevertheless, mean-variance efficient portfolios are also relevant for asset pricing purposes due to the following theorem:

Theorem 4.6 *A return is a pricing factor if and only if it is a mean-variance efficient return different from the minimum-variance return.*

Combining this with results from the previous subsection, we can conclude that (almost) any mean-variance return R^{mv} give rise to a (candidate) state-price deflator of the form $\zeta = a + bR^{\mathrm{mv}}$. And the standard CAPM can be reformulated as "the return on the market portfolio is mean-variance efficient."

We will not provide a proof of the theorem here but return to the issue in Chapter 9.

4.6 Concluding remarks

This chapter has introduced state-price deflators as a way representing the general pricing mechanism of a financial market. Important properties of state-price deflators were discussed. Examples have illustrated the valuation of assets with a given state-price deflator. But what determines the state-price deflator? Intuitively, state prices reflect the value market participants attach to an extra payment in a given state at a given point in time. This must be related to their marginal utility of consumption. To follow this idea, we must consider the optimal consumption choice of individuals. This is the topic of the next two chapters.

4.7 Exercises

EXERCISE 4.1 Imagine a one-period economy where the state-price deflator ζ is lognormally distributed with $\mathrm{E}[\ln\zeta] = \mu_\zeta$ and $\mathrm{Var}[\ln\zeta] = \sigma_\zeta^2$. What is the maximal Sharpe ratio of a risky asset? (Look at Equation (4.14).) What determines the sign of an asset's Sharpe ratio?

EXERCISE 4.2 Consider a one-period, three-state economy with two assets traded. Asset 1 has a price of 0.9 and pays a dividend of 1 no matter what state is realized. Asset 2 has a price of 2 and pays 1, 2, and 4 in state 1, 2, and 3, respectively. The real-world probabilities of the states are 0.2, 0.6, and 0.2, respectively. Assume absence of arbitrage.

(a) Find the state-price vector $\boldsymbol{\psi}^*$ and the associated state-price deflator. What portfolio generates a dividend of $\boldsymbol{\psi}^*$?

(b) Find the state-price deflator ζ^* and the associated state-price vector. What portfolio generates a dividend of ζ^*?

(c) Find a portfolio with dividends 4, 3, and 1 in states 1, 2, and 3, respectively. What is the price of the portfolio?

EXERCISE 4.3 Give a proof of Theorem 4.3 both for the one-period framework and the continuous-time framework.

EXERCISE 4.4 Show Lemma 4.1.

EXERCISE 4.5 Assume a one-period framework with no redundant assets. Show that ζ^* can be rewritten as in (4.55).

EXERCISE 4.6 Consider a one-period economy where the assets are correctly priced by a state-price deflator M. A nutty professor believes that the assets are priced according to a model in which Y is a state-price deflator, where Y is a random variable with $E[Y] = E[M]$. Refer to this model as the Y-model.

(a) Show that the Y-model prices a risk-free asset correctly.

(b) Argue that the expected return on an arbitrary asset i according to the Y-model is given by

$$\mathrm{E}_Y[R_i] \equiv \frac{1}{\mathrm{E}[M]} - \frac{1}{\mathrm{E}[M]}\,\mathrm{Cov}[Y, R_i].$$

(c) Show that

$$\frac{|\,\mathrm{E}[R_i] - \mathrm{E}_Y[R_i]\,|}{\sigma[R_i]} \leq \frac{\sigma[Y - M]}{\mathrm{E}[M]}$$

so that the mispricing of the Y-model (in terms of expected returns) is limited.

(d) What can you say about the returns for which the left-hand side in the above inequality will be largest?

(e) Under which condition on Y will the Y-model price all assets correctly?

EXERCISE 4.7 Consider a one-period economy where two basic financial assets are traded without portfolio constraints or transaction costs. There are three equally likely end-of-period states of the economy and the prices and state-contingent dividends of the two assets are given in the following table:

	state-contingent dividend			price
	state 1	state 2	state 3	
Asset 1	1	1	0	0.5
Asset 2	2	2	2	1.8

The economy is known to be arbitrage-free.

(a) Characterize the set of state-contingent dividends that can be attained by combining the two assets.

(b) Characterize the set of state-price vectors $\boldsymbol{\psi} = (\psi_1, \psi_2, \psi_3)$ consistent with the prices and dividends of the two basic assets.

(c) Find the state-price vector $\boldsymbol{\psi}^*$ that belongs to the set of attainable dividend vectors.

(d) Characterize the set of state-price deflators $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \zeta_3)$ consistent with the prices and dividends of the two basic assets.

(e) What is the risk-free (gross) return R^f over the period?

(f) What prices of the state-contingent dividend vector (1,1,5) are consistent with absence of arbitrage?

(g) What prices of the state-contingent dividend vector (1,2,5) are consistent with absence of arbitrage?

(h) Show that the state-price deflator $\boldsymbol{\zeta}^*$ that belongs to the set of attainable dividend vectors is given by the vector $(0.75, 0.75, 1.2)$, i.e. it has the value 0.75 in states 1 and 2 and the value 1.2 in state 3.

EXERCISE 4.8 Consider a two-period economy where the resolution of uncertainty can be represented by the tree in Figure 2.2 in Chapter 2. Assume that three assets are traded. Their dividend processes are illustrated in Figure 4.1, where a triple (D_1, D_2, D_3) near a node means that asset $i = 1, 2, 3$ pays a dividend of D_i if/when that node is reached. For example, if the economy at time 2 is in the scenario F_{22}, assets 1 and 2 will pay a dividend of 1 and asset 3 will pay a dividend of 2. In Figure 4.1, the numbers near the lines connecting nodes denote values of the next-period state-price deflator, i.e. $\zeta_1/\zeta_0 = \zeta_1$ over the first period and ζ_2/ζ_1 over the second period. For example, the value of ζ_2/ζ_1 given that the economy is in scenario F_{11} at time 1 is 1 if the economy moves to scenario F_{21} and 2/3 if the economy moves to scenario F_{22}. This characterizes completely the state-price deflator to be used in the computations below.

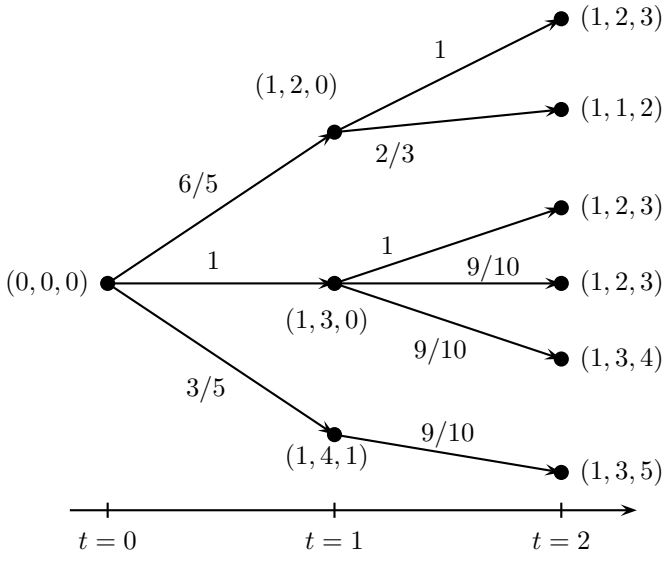


Figure 4.1: Dividends and state prices.

(a) Find the price processes of the three assets. (You should find that the time 0 price of asset 1 is 1.802.)

(b) Find the short-term (one-period) interest rate process. What are the one-period and the two-period zero-coupon yields at time 0?

(c) Is the market complete?

EXERCISE 4.9 Consider a discrete-time economy where a state-price deflator $\zeta = (\zeta_t)$ satisfies

$$\ln\left(\frac{\zeta_{t+1}}{\zeta_t}\right) \sim N(\mu_\zeta, \sigma_\zeta^2)$$

for each $t = 0, 1, \ldots, T-1$. An asset pays a dividend process $D = (D_t)$ satisfying

$$\frac{D_{t+1}}{D_t} = a\left(\frac{\zeta_{t+1}}{\zeta_t}\right)^b + \varepsilon_{t+1}, \quad t = 0, 1, \ldots, T-1,$$

where a and b are constants, and $\varepsilon_1, \ldots, \varepsilon_T$ are independent with $\mathrm{E}_t[\varepsilon_{t+1}] = 0$ and $\mathrm{E}_t[\varepsilon_{t+1}\zeta_{t+1}] = 0$ for all t. What is the price of the asset at any time t (in terms of $a, b, \mu_\zeta, \sigma_\zeta^2$)?

EXERCISE 4.10 Consider a continuous-time economy in which the state-price deflator follows a geometric Brownian motion:

$$d\zeta_t = -\zeta_t \left[r^f\, dt + \boldsymbol{\lambda}^\top d\boldsymbol{z}_t\right],$$

where r^f and $\boldsymbol{\lambda}$ are constant.

(a) What is the price B_t^s of a zero-coupon bond maturing at time s with a face value of 1?

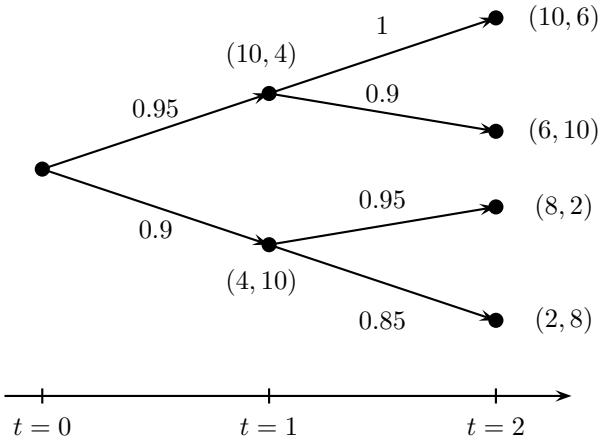
Define the continuously compounded yield y_t^s for maturity s via the equation $B_t^s = e^{-y_t^s(s-t)}$.

(b) Compute y_t^s. What can you say about the yield curve $s \mapsto y_t^s$?

EXERCISE 4.11 Show that $\zeta_t^* = V_0^*/V_t^*$, where V^* and ζ^* are given by (4.46) and (4.45).

EXERCISE 4.12 In a one-period framework show that if x is a pricing factor and k_1, k_2 are constants with $k_2 \neq 0$, then $y = k_1 + k_2 x$ is also a pricing factor.

EXERCISE 4.13 Consider a two-period arbitrage-free economy where the resolution of uncertainty is illustrated in the following binomial tree.



Each branch in the tree has a conditional probability of $\frac{1}{2}$. Assets in the economy are priced by a state-price deflator $\zeta = (\zeta_t)_{t\in\{0,1,2\}}$. The numbers along the branches show the possible values of the state-price deflator over that period, i.e. ζ_1/ζ_0 over the first period and ζ_2/ζ_1 over the second period. The pair of numbers written at each node shows the dividend payments of asset 1 and asset 2, respectively, if that node is reached. For example, if the up-branch is realized in both periods, then asset 1 will pay a dividend of 10 and asset 2 a dividend of 6 at time 2.

(a) For each of the two assets compute the following quantities in both the up-node and the down-node at time 1: (i) the conditional expectation of the dividend received at time 2, (ii) the ex-dividend price, and (iii) the expected net rate of return over the second period.

(b) For each of the two assets compute the following quantities at time 0: (i) the expectation of the dividend received at time 1, (ii) the price, and (iii) the expected net rate of return over the first period.

(c) Compare the prices of the two assets. Compare the expected returns of the two assets. Explain the differences.

(d) Is it always possible in this economy to construct a portfolio with a risk-free dividend over the next period? If so, find the one-period risk-free return at time 0 and in each of the two nodes at time 1.

(e) Is the market complete? Explain!

Chapter 5

Modeling the preferences of individuals

5.1 Introduction

In order to say anything concrete about the asset supply and demand of individuals we have to formalize the decision problem faced by individuals. We assume that individuals have preferences for consumption and must choose between different consumption plans, i.e. plans for how much to consume at different points in time and in different states of the world. The financial market allows individuals to reallocate consumption over time and over states and hence obtain a consumption plan different from their endowment.

Although an individual will typically obtain utility from consumption at many different dates (or in many different periods), we will first address the simpler case with consumption at only one future point in time. In such a setting a "consumption plan" is simply a random variable representing the consumption at that date. Even in one-period models individuals should be allowed to consume both at the beginning of the period and at the end of the period, but we will first ignore the influence of current consumption on the well-being of the individual. We do that both since current consumption is certain and we want to focus on how preferences for uncertain consumption can be represented, but also to simplify the notation and analysis somewhat. Since we have in mind a one-period economy, we basically have to model preferences for end-of-period consumption.

Sections 5.2–5.4 discuss how to represent individual preferences in a tractable way. We will demonstrate that under some fundamental assumptions ("axioms") on individual behavior, the preferences can be modeled by a utility index which to each consumption plan assigns a real number with higher numbers to the more preferred plans. Under an additional axiom we can represent the preferences in terms of expected utility, which is even simpler to work with and used in most models of financial economics. Section 5.5 defines and discusses the important concept of risk aversion. Section 5.6 introduces the utility functions that are typically applied in models of financial economics and provides a short discussion of which utility functions and levels of risk aversions that seem to be reasonable for representing the decisions of individuals. In Section 5.7 we discuss extensions to preferences for consumption at more than one point in time.

There is a large literature on how to model the preferences of individuals for uncertain outcomes and the presentation here is by no means exhaustive. The literature dates back at least to Swiss mathematician Daniel Bernoulli in 1738 (see English translation in Bernoulli (1954)), but was put on a firm formal setting by von Neumann and Morgenstern (1944). For some recent textbook presentations on a similar level as the one given here, see Huang and Litzenberger (1988, Ch. 1), Kreps (1990, Ch. 3), Gollier (2001, Chs. 1-3), and Danthine and Donaldson (2002, Ch. 2).

5.2 Consumption plans and preference relations

It seems fair to assume that whenever the individual compares two different consumption plans, she will be able either to say that she prefers one of them to the other or to say that she is indifferent between the two consumption plans. Moreover, she should make such pairwise comparisons in a consistent way. For example, if she prefers plan 1 to plan 2 and plan 2 to plan 3, she should prefer plan 1 to plan 3. If these properties hold, we can formally represent the preferences of the individual by a so-called *preference relation*. A preference relation itself is not very tractable so we are looking for simpler ways of representing preferences. First, we will find conditions under which it makes sense to represent preferences by a so-called *utility index* which attaches a real number to each consumption plan. If and only if plan 1 has a higher utility index than plan 2, the individual prefers plan 1 to plan 2. Attaching numbers to each possible consumption plan is also not easy so we look for an even simpler representation. We show that under an additional condition we can represent preferences in an even simpler way in terms of the expected value of a *utility function*. A utility function is a function defined on the set of possible levels of consumption. Since consumption is random it then makes sense to talk about the expected utility of a consumption plan. The individual will prefer consumption plan 1 to plan 2 if and only if the expected utility from consumption plan 1 is higher than the expected utility from consumption plan 2. This representation of preferences turns out to be very tractable and is applied in the vast majority of asset pricing models.

Our main analysis is formulated under some simplifying assumptions that are not necessarily appropriate. At the end of this section we will briefly discuss how to generalize the analysis and also discuss the appropriateness of the axioms on individual behavior that need to be imposed in order to obtain the expected utility representation.

We assume that there is uncertainty about how the variables affecting the well-being of an individual (e.g. asset returns) turn out. We model the uncertainty by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. In most of the chapter we will assume that the state space is finite, $\Omega = \{1, 2, \ldots, S\}$, so that there are S possible states of which exactly one will be realized. For simplicity, think of this as a model of one-period economy with S possible states at the end of the period. The set $\mathcal{F}$ of events that can be assigned a probability is the collection of all subsets of Ω. The probability measure $\mathbb{P}$ is defined by the individual state probabilities $p_\omega = \mathbb{P}(\omega)$, $\omega = 1, 2, \ldots, S$. We assume that all $p_\omega > 0$ and, of course, we have that $p_1 + \ldots p_S = 1$. We take the state probabilities as exogenously given and known to the individuals.

Individuals care about their consumption. It seems reasonable to assume that when an individual chooses between two different actions (e.g. portfolio choices), she only cares about the consumption plans generated by these choices. For example, she will be indifferent between two choices that

state ω	1	2	3
state prob. p_ω	0.2	0.3	0.5
cons. plan 1, $c^{(1)}$	3	2	4
cons. plan 2, $c^{(2)}$	3	1	5
cons. plan 3, $c^{(3)}$	4	4	1
cons. plan 4, $c^{(4)}$	1	1	4

Table 5.1: The possible state-contingent consumption plans in the example.

cons. level z	1	2	3	4	5
cons. plan 1, $\pi_{c^{(1)}}$	0	0.3	0.2	0.5	0
cons. plan 2, $\pi_{c^{(2)}}$	0.3	0	0.2	0	0.5
cons. plan 3, $\pi_{c^{(3)}}$	0.5	0	0	0.5	0
cons. plan 4, $\pi_{c^{(4)}}$	0.5	0	0	0.5	0

Table 5.2: The probability distributions corresponding to the state-contingent consumption plans shown in Table 5.1.

generate exactly the same consumption plans, i.e. the same consumption levels in all states. In order to simplify the following analysis, we will assume a bit more, namely that the individual only cares about the probability distribution of consumption generated by each portfolio. This is effectively an assumption of *state-independent preferences*.

We can represent a consumption plan by a random variable c on $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that there is only one consumption good and since consumption should be non-negative, c is valued in $\mathbb{R}_+ = [0, \infty)$. As long as we are assuming a finite state space $\Omega = \{1, 2, \ldots, S\}$ we can equivalently represent the consumption plan by a vector $(c_1, \ldots, c_S)$, where $c_\omega \in [0, \infty)$ denotes the consumption level if state ω is realized, i.e. $c_\omega \equiv c(\omega)$. Let $\mathcal{C}$ denote the set of consumption plans that the individual has to choose among. Let $Z \subseteq \mathbb{R}_+$ denote the set of all the possible levels of the consumption plans that are considered, i.e. no matter which of these consumption plans we take, its value will be in Z no matter which state is realized. Each consumption plan $c \in \mathcal{C}$ is associated with a probability distribution π_c, which is the function $\pi_c : Z \to [0, 1]$, given by

$$\pi_c(z) = \sum_{\omega \in \Omega:\, c_\omega = z} p_\omega,$$

i.e. the sum of the probabilities of those states in which the consumption level equals z.

As an example consider an economy with three possible states and four possible state-contingent consumption plans as illustrated in Table 5.1. These four consumption plans may be the product of four different portfolio choices. The set of possible end-of-period consumption levels is $Z = \{1, 2, 3, 4, 5\}$. Each consumption plan generates a probability distribution on the set Z. The probability distributions corresponding to these consumption plans are as shown in Table 5.2. We see that although the consumption plans $c^{(3)}$ and $c^{(4)}$ are different they generate identical probability distributions. By assumption individuals will be indifferent between these two consumption plans.

Given these assumptions the individual will effectively choose between probability distributions

on the set of possible consumption levels Z. We assume for simplicity that Z is a finite set, but the results can be generalized to the case of infinite Z at the cost of further mathematical complexity. We denote by $\mathcal{P}(Z)$ the set of all probability distributions on Z that are generated by consumption plans in $\mathcal{C}$. A probability distribution π on the finite set Z is simply a function $\pi : Z \to [0,1]$ with the properties that $\sum_{z\in Z} \pi(z) = 1$ and $\pi(A \cup B) = \pi(A) + \pi(B)$ whenever $A \cap B = \emptyset$.

We assume that the preferences of the individual can be represented by a preference relation $\succeq$ on $\mathcal{P}(Z)$, which is a binary relation satisfying the following two conditions:

(i) if $\pi_1 \succeq \pi_2$ and $\pi_2 \succeq \pi_3$, then $\pi_1 \succeq \pi_3$ [transitivity]

(ii) $\forall \pi_1, \pi_2 \in \mathcal{P}(Z)$: either $\pi_1 \succeq \pi_2$ or $\pi_2 \succeq \pi_1$ [completeness]

Here, $\pi_1 \succeq \pi_2$ is to be read as "π_1 is preferred to π_2". We write $\pi_1 \not\succeq \pi_2$ if π_1 is not preferred to π_2. If both $\pi_1 \succeq \pi_2$ and $\pi_2 \succeq \pi_1$, we write $\pi_1 \sim \pi_2$ and say that the individual is indifferent between π_1 and π_2. If $\pi_1 \succeq \pi_2$, but $\pi_2 \not\succeq \pi_1$, we say that π_1 is strictly preferred to π_2 and write $\pi_1 \succ \pi_2$.

Note that if $\pi_1, \pi_2 \in \mathcal{P}(Z)$ and $\alpha \in [0,1]$, then $\alpha\pi_1 + (1-\alpha)\pi_2 \in \mathcal{P}(Z)$. The mixed distribution $\alpha\pi_1 + (1-\alpha)\pi_2$ assigns the probability $(\alpha\pi_1 + (1-\alpha)\pi_2)(z) = \alpha\pi_1(z) + (1-\alpha)\pi_2(z)$ to the consumption level z. When can think of the mixed distribution $\alpha\pi_1 + (1-\alpha)\pi_2$ as the outcome of a two-stage "gamble." The first stage is to flip a coin which with probability α shows head and with probability $1-\alpha$ shows tails. If head comes out, the second stage is the "consumption gamble" corresponding to the probability distribution π_1. If tails is the outcome of the first stage, the second stage is the consumption gamble corresponding to π_2. When we assume that preferences are represented by a preference relation on the set $\mathcal{P}(Z)$ of probability distributions, we have implicitly assumed that the individual evaluates the two-stage gamble (or any multi-stage gamble) by the combined probability distribution, i.e. the ultimate consequences of the gamble. This is sometimes referred to as *consequentialism*.

Let z be some element of Z, i.e. some possible consumption level. By $\mathbf{1}_z$ we will denote the probability distribution that assigns a probability of one to z and a zero probability to all other elements in Z. Since we have assumed that the set Z of possible consumption levels only has a finite number of elements, it must have a maximum element, say z^u, and a minimum element, say z^l. Since the elements represent consumption levels, it is certainly natural that individuals prefer higher elements than lower. We will therefore assume that the probability distribution $\mathbf{1}_{z^u}$ is preferred to any other probability distribution. Conversely, any probability distribution is preferred to the probability distribution $\mathbf{1}_{z^l}$. We assume that $\mathbf{1}_{z^u}$ is strictly preferred to $\mathbf{1}_{z^l}$ so that the individual is not indifferent between all probability distributions. For any $\pi \in \mathcal{P}(Z)$ we thus have that,

$$\mathbf{1}_{z^u} \succ \pi \succ \mathbf{1}_{z^l} \quad \text{or} \quad \mathbf{1}_{z^u} \sim \pi \succ \mathbf{1}_{z^l} \quad \text{or} \quad \mathbf{1}_{z^u} \succ \pi \sim \mathbf{1}_{z^l}.$$

5.3 Utility indices

A utility index for a given preference relation $\succeq$ is a function $\mathcal{U} : \mathcal{P}(Z) \to \mathbb{R}$ that to each probability distribution over consumption levels attaches a real-valued number such that

$$\pi_1 \succeq \pi_2 \quad \Leftrightarrow \quad \mathcal{U}(\pi_1) \geq \mathcal{U}(\pi_2).$$

Note that a utility index is only unique up to a strictly increasing transformation. If $\mathcal{U}$ is a utility index and $f : \mathbb{R} \to \mathbb{R}$ is any strictly increasing function, then the composite function $\mathcal{V} = f \circ \mathcal{U}$, defined by $\mathcal{V}(\pi) = f(\mathcal{U}(\pi))$, is also a utility index for the same preference relation.

We will show below that a utility index exists under the following two axiomatic assumptions on the preference relation $\succeq$:

Axiom 5.1 (Monotonicity) *Suppose that $\pi_1, \pi_2 \in \mathcal{P}(Z)$ with $\pi_1 \succ \pi_2$ and let $a, b \in [0,1]$. The preference relation $\succeq$ has the property that*

$$a > b \quad \Leftrightarrow \quad a\pi_1 + (1-a)\pi_2 \succ b\pi_1 + (1-b)\pi_2.$$

This is certainly a very natural assumption on preferences. If you consider a weighted average of two probability distributions, you will prefer a high weight on the best of the two distributions.

Axiom 5.2 (Archimedean) *The preference relation $\succeq$ has the property that for any three probability distributions $\pi_1, \pi_2, \pi_3 \in \mathcal{P}(Z)$ with $\pi_1 \succ \pi_2 \succ \pi_3$, numbers $a, b \in (0,1)$ exist such that*

$$a\pi_1 + (1-a)\pi_3 \succ \pi_2 \succ b\pi_1 + (1-b)\pi_3.$$

The axiom basically says that no matter how good a probability distribution π_1 is, it is so that for any $\pi_2 \succ \pi_3$ we can find some mixed distribution of π_1 and π_3 to which π_2 is preferred. We just have to put a sufficiently low weight on π_1 in the mixed distribution. Similarly, no matter how bad a probability distribution π_3 is, it is so that for any $\pi_1 \succ \pi_2$ we can find some mixed distribution of π_1 and π_3 that is preferred to π_2. We just have to put a sufficiently low weight on π_3 in the mixed distribution.

We shall say that a preference relation has the *continuity property* if for any three probability distributions $\pi_1, \pi_2, \pi_3 \in \mathcal{P}(Z)$ with $\pi_1 \succ \pi_2 \succ \pi_3$, a unique number $\alpha \in (0,1)$ exists such that

$$\pi_2 \sim \alpha\pi_1 + (1-\alpha)\pi_3.$$

We can easily extend this to the case where either $\pi_1 \sim \pi_2$ or $\pi_2 \sim \pi_3$. For $\pi_1 \sim \pi_2 \succ \pi_3$, $\pi_2 \sim 1\pi_1 + (1-1)\pi_3$ corresponding to $\alpha = 1$. For $\pi_1 \succ \pi_2 \sim \pi_3$, $\pi_2 \sim 0\pi_1 + (1-0)\pi_3$ corresponding to $\alpha = 0$. In words the continuity property means that for any three probability distributions there is a unique combination of the best and the worst distribution so that the individual is indifferent between the third "middle" distribution and this combination of the other two. This appears to be closely related to the Archimedean Axiom and, in fact, the next lemma shows that the Monotonicity Axiom and the Archimedean Axiom imply continuity of preferences.

Lemma 5.1 *Let $\succeq$ be a preference relation satisfying the Monotonicity Axiom and the Archimedean Axiom. Then it has the continuity property.*

Proof: Given $\pi_1 \succ \pi_2 \succ \pi_3$. Define the number α by

$$\alpha = \sup\{k \in [0,1] \mid \pi_2 \succ k\pi_1 + (1-k)\pi_3\}.$$

By the Monotonicity Axiom we have that $\pi_2 \succ k\pi_1 + (1-k)\pi_3$ for all $k < \alpha$ and that $k\pi_1 + (1-k)\pi_3 \succeq \pi_2$ for all $k > \alpha$. We want to show that $\pi_2 \sim \alpha\pi_1 + (1-\alpha)\pi_3$. Note that by the

Archimedean Axiom, there is some $k > 0$ such that $\pi_2 \succ k\pi_1 + (1-k)\pi_3$ and some $k < 1$ such that $k\pi_1 + (1-k)\pi_3 \succ \pi_2$. Consequently, α is in the open interval $(0,1)$.

Suppose that $\pi_2 \succ \alpha\pi_1 + (1-\alpha)\pi_3$. Then according to the Archimedean Axiom we can find a number $b \in (0,1)$ such that $\pi_2 \succ b\pi_1 + (1-b)\{\alpha\pi_1 + (1-\alpha)\pi_3\}$. The mixed distribution on the right-hand side has a total weight of $k = b + (1-b)\alpha = \alpha + (1-\alpha)b > \alpha$ on π_1. Hence we have found some $k > \alpha$ for which $\pi_2 \succ k\pi_1 + (1-k)\pi_3$. This contradicts the definition of α. Consequently, we must have that $\pi_2 \not\succ \alpha\pi_1 + (1-\alpha)\pi_3$.

Now suppose that $\alpha\pi_1 + (1-\alpha)\pi_3 \succ \pi_2$. Then we know from the Archimedean Axiom that a number $a \in (0,1)$ exists such that $a\{\alpha\pi_1 + (1-\alpha)\pi_3\} + (1-a)\pi_3 \succ \pi_2$. The mixed distribution on the left-hand side has a total weight of $a\alpha < \alpha$ on π_1. Hence we have found some $k < \alpha$ for which $k\pi_1 + (1-k)\pi_3 \succ \pi_2$. This contradicts the definition of α. We can therefore also conclude that $\alpha\pi_1 + (1-\alpha)\pi_3 \not\succ \pi_2$. In sum, we have $\pi_2 \sim \alpha\pi_1 + (1-\alpha)\pi_3$. □

The next result states that a preference relation which satisfies the Monotonicity Axiom and has the continuity property can always be represented by a utility index. In particular this is true when $\succeq$ satisfies the Monotonicity Axiom and the Archimedean Axiom.

Theorem 5.1 *Let $\succeq$ be a preference relation which satisfies the Monotonicity Axiom and has the continuity property. Then it can be represented by a utility index $\mathcal{U}$, i.e. a function $\mathcal{U} : \mathcal{P}(Z) \to \mathbb{R}$ with the property that*

$$\pi_1 \succeq \pi_2 \quad \Leftrightarrow \quad \mathcal{U}(\pi_1) \geq \mathcal{U}(\pi_2).$$

Proof: Recall that we have assumed a best probability distribution $\mathbf{1}_{z^u}$ and a worst probability distribution $\mathbf{1}_{z^l}$ in the sense that

$$\mathbf{1}_{z^u} \succ \pi \succ \mathbf{1}_{z^l} \quad \text{or} \quad \mathbf{1}_{z^u} \sim \pi \succ \mathbf{1}_{z^l} \quad \text{or} \quad \mathbf{1}_{z^u} \succ \pi \sim \mathbf{1}_{z^l}$$

for any $\pi \in \mathcal{P}(Z)$. For any $\pi \in \mathcal{P}(Z)$ we know from the continuity property that a unique number $\alpha_\pi \in [0,1]$ exists such that

$$\pi \sim \alpha_\pi \mathbf{1}_{z^u} + (1-\alpha_\pi)\mathbf{1}_{z^l}.$$

If $\mathbf{1}_{z^u} \sim \pi \succ \mathbf{1}_{z^l}$, $\alpha_\pi = 1$. If $\mathbf{1}_{z^u} \succ \pi \sim \mathbf{1}_{z^l}$, $\alpha_\pi = 0$. If $\mathbf{1}_{z^u} \succ \pi \succ \mathbf{1}_{z^l}$, $\alpha_\pi \in (0,1)$.

We define the function $\mathcal{U} : \mathcal{P}(Z) \to \mathbb{R}$ by $\mathcal{U}(\pi) = \alpha_\pi$. By the Monotonicity Axiom we know that $\mathcal{U}(\pi_1) \geq \mathcal{U}(\pi_2)$ if and only if

$$\mathcal{U}(\pi_1)\mathbf{1}_{z^u} + (1 - \mathcal{U}(\pi_1))\,\mathbf{1}_{z^l} \succeq \mathcal{U}(\pi_2)\mathbf{1}_{z^u} + (1 - \mathcal{U}(\pi_2))\,\mathbf{1}_{z^l},$$

and hence if and only if $\pi_1 \succeq \pi_2$. It follows that $\mathcal{U}$ is a utility index. □

5.4 Expected utility representation of preferences

Utility indices are functions of probability distributions on the set of possible consumption levels. With many states of the world and many assets to trade in, the set of such probability distributions will be very, very large. This will significantly complicate the analysis of optimal choice using

utility indices to represent preferences. To simplify the analysis financial economists traditionally put more structure on the preferences so that they can be represented in terms of expected utility.

We say that a preference relation $\succeq$ on $\mathcal{P}(Z)$ has an expected utility representation if there exists a function $u : Z \to \mathbb{R}$ such that

$$\pi_1 \succeq \pi_2 \quad \Leftrightarrow \quad \sum_{z\in Z} \pi_1(z)u(z) \geq \sum_{z\in Z} \pi_2(z)u(z). \tag{5.1}$$

Here $\sum_{z\in Z} \pi(z)u(z)$ is the expected utility of end-of-period consumption given the consumption probability distribution π. The function u is called a von Neumann-Morgenstern utility function or simply a utility function. Note that u is defined on the set Z of consumption levels, which in general has a simpler structure than the set of probability distributions on Z. Given a utility function u, we can obviously define a utility index by $\mathcal{U}(\pi) = \sum_{z\in Z} \pi(z)u(z)$.

5.4.1 Conditions for expected utility

When can we use an expected utility representation of a preference relation? The next lemma is a first step.

Lemma 5.2 *A preference relation $\succeq$ has an expected utility representation if and only if it can be represented by a linear utility index $\mathcal{U}$ in the sense that*

$$\mathcal{U}\left(a\pi_1 + (1-a)\pi_2\right) = a\mathcal{U}(\pi_1) + (1-a)\mathcal{U}(\pi_2)$$

for any $\pi_1, \pi_2 \in \mathcal{P}(Z)$ and any $a \in [0,1]$.

Proof: Suppose that $\succeq$ has an expected utility representation with utility function u. Define $\mathcal{U} : \mathcal{P}(Z) \to \mathbb{R}$ by $\mathcal{U}(\pi) = \sum_{z\in Z} \pi(z)u(z)$. Then clearly $\mathcal{U}$ is a utility index representing $\succeq$ and $\mathcal{U}$ is linear since

$$\begin{aligned}
\mathcal{U}\left(a\pi_1 + (1-a)\pi_2\right) &= \sum_{z\in Z} \left(a\pi_1(z) + (1-a)\pi_2(z)\right) u(z) \\
&= a\sum_{z\in Z} \pi_1(z)u(z) + (1-a)\sum_{z\in Z} \pi_2(z)u(z) \\
&= a\mathcal{U}(\pi_1) + (1-a)\mathcal{U}(\pi_2).
\end{aligned}$$

Conversely, suppose that $\mathcal{U}$ is a linear utility index representing $\succeq$. Define a function $u : Z \to \mathbb{R}$ by $u(z) = \mathcal{U}(\mathbf{1}_z)$. For any $\pi \in \mathcal{P}(Z)$ we have

$$\pi \sim \sum_{z\in Z} \pi(z)\mathbf{1}_z.$$

Therefore,

$$\mathcal{U}(\pi) = \mathcal{U}\left(\sum_{z\in Z} \pi(z)\mathbf{1}_z\right) = \sum_{z\in Z} \pi(z)\mathcal{U}(\mathbf{1}_z) = \sum_{z\in Z} \pi(z)u(z).$$

Since $\mathcal{U}$ is a utility index, we have $\pi_1 \succeq \pi_2 \Leftrightarrow \mathcal{U}(\pi_1) \geq \mathcal{U}(\pi_2)$, which the computation above shows is equivalent to $\sum_{z\in Z} \pi_1(z)u(z) \geq \sum_{z\in Z} \pi_2(z)u(z)$. Consequently, u gives an expected utility

z	1	2	3	4
π_1	0	0.2	0.6	0.2
π_2	0	0.4	0.2	0.4
π_3	1	0	0	0
π_4	0.5	0.1	0.3	0.1
π_5	0.5	0.2	0.1	0.2

Table 5.3: The probability distributions used in the illustration of the Substitution Axiom.

representation of $\succeq$. □

The question then is under what assumptions the preference relation $\succeq$ can be represented by a linear utility index. As shown by von Neumann and Morgenstern (1944) we need an additional axiom, the so-called Substitution Axiom.

Axiom 5.3 (Substitution) *For all $\pi_1, \pi_2, \pi_3 \in \mathcal{P}(Z)$ and all $a \in (0, 1]$, we have*

$$\pi_1 \succ \pi_2 \quad \Leftrightarrow \quad a\pi_1 + (1-a)\pi_3 \succ a\pi_2 + (1-a)\pi_3$$

and

$$\pi_1 \sim \pi_2 \quad \Leftrightarrow \quad a\pi_1 + (1-a)\pi_3 \sim a\pi_2 + (1-a)\pi_3.$$

The Substitution Axiom is sometimes called the Independence Axiom or the Axiom of the Irrelevance of the Common Alternative. Basically, it says that when the individual is to compare two probability distributions, she need only consider the parts of the two distributions which are different from each other. As an example, suppose the possible consumption levels are $Z = \{1, 2, 3, 4\}$ and consider the probability distributions on Z given in Table 5.3. Suppose you want to compare the distributions π_4 and π_5. They only differ in the probabilities they associate with consumption levels 2, 3, and 4 so it should only be necessary to focus on these parts. More formally observe that

$$\pi_4 \sim 0.5\pi_1 + 0.5\pi_3 \quad \text{and} \quad \pi_5 \sim 0.5\pi_2 + 0.5\pi_3.$$

π_1 is the conditional distribution of π_4 given that the consumption level is different from 1 and π_2 is the conditional distribution of π_5 given that the consumption level is different from 1. The Substitution Axiom then says that

$$\pi_4 \succ \pi_5 \quad \Leftrightarrow \quad \pi_1 \succ \pi_2.$$

The next lemma shows that the Substitution Axiom is more restrictive than the Monotonicity Axiom.

Lemma 5.3 *If a preference relation $\succeq$ satisfies the Substitution Axiom, it will also satisfy the Monotonicity Axiom.*

Proof: Given $\pi_1, \pi_2 \in \mathcal{P}(Z)$ with $\pi_1 \succ \pi_2$ and numbers $a, b \in [0, 1]$. We have to show that

$$a > b \quad \Leftrightarrow \quad a\pi_1 + (1-a)\pi_2 \succ b\pi_1 + (1-b)\pi_2.$$

Note that if $a = 0$, we cannot have $a > b$, and if $a\pi_1 + (1-a)\pi_2 \succ b\pi_1 + (1-b)\pi_2$ we cannot have $a = 0$. We can therefore safely assume that $a > 0$.

First assume that $a > b$. Observe that it follows from the Substitution Axiom that

$$a\pi_1 + (1-a)\pi_2 \succ a\pi_2 + (1-a)\pi_2$$

and hence that $a\pi_1 + (1-a)\pi_2 \succ \pi_2$. Also from the Substitution Axiom we have that for any $\pi_3 \succ \pi_2$, we have

$$\pi_3 \sim \left(1 - \frac{b}{a}\right)\pi_3 + \frac{b}{a}\pi_3 \succ \left(1 - \frac{b}{a}\right)\pi_2 + \frac{b}{a}\pi_3.$$

Due to our observation above, we can use this with $\pi_3 = a\pi_1 + (1-a)\pi_2$. Then we get

$$\begin{aligned} a\pi_1 + (1-a)\pi_2 &\succ \frac{b}{a}\{a\pi_1 + (1-a)\pi_2\} + \left(1 - \frac{b}{a}\right)\pi_2 \\ &\sim b\pi_1 + (1-b)\pi_2, \end{aligned}$$

as was to be shown.

Conversely, assuming that

$$a\pi_1 + (1-a)\pi_2 \succ b\pi_1 + (1-b)\pi_2,$$

we must argue that $a > b$. The above inequality cannot be true if $a = b$ since the two combined distributions are then identical. If b was greater than a, we could follow the steps above with a and b swapped and end up concluding that $b\pi_1 + (1-b)\pi_2 \succ a\pi_1 + (1-a)\pi_2$, which would contradict our assumption. Hence, we cannot have neither $a = b$ nor $a < b$ but must have $a > b$. □

Next we state the main result:

Theorem 5.2 *Assume that Z is finite and that $\succeq$ is a preference relation on $\mathcal{P}(Z)$. Then $\succeq$ can be represented by a linear utility index if and only if $\succeq$ satisfies the Archimedean Axiom and the Substitution Axiom.*

Proof: First suppose the preference relation $\succeq$ satisfies the Archimedean Axiom and the Substitution Axiom. Define a utility index $\mathcal{U} : \mathcal{P}(Z) \to \mathbb{R}$ exactly as in the proof of Theorem 5.1, i.e. $\mathcal{U}(\pi) = \alpha_\pi$, where $\alpha_\pi \in [0,1]$ is the unique number such that

$$\pi \sim \alpha_\pi \mathbf{1}_{z^u} + (1 - \alpha_\pi)\mathbf{1}_{z^l}.$$

We want to show that, as a consequence of the Substitution Axiom, $\mathcal{U}$ is indeed linear. For that purpose, pick any two probability distributions $\pi_1, \pi_2 \in \mathcal{P}(Z)$ and any number $a \in [0,1]$. We want to show that $\mathcal{U}(a\pi_1 + (1-a)\pi_2) = a\mathcal{U}(\pi_1) + (1-a)\mathcal{U}(\pi_2)$. We can do that by showing that

$$a\pi_1 + (1-a)\pi_2 \sim (a\mathcal{U}(\pi_1) + (1-a)\mathcal{U}(\pi_2))\,\mathbf{1}_{z^u} + (1 - \{a\mathcal{U}(\pi_1) + (1-a)\mathcal{U}(\pi_2)\})\,\mathbf{1}_{z^l}.$$

This follows from the Substitution Axiom:

$$\begin{aligned} a\pi_1 + (1-a)\pi_2 &\sim a\{\mathcal{U}(\pi_1)\mathbf{1}_{z^u} + (1 - \mathcal{U}(\pi_1))\,\mathbf{1}_{z^l}\} + (1-a)\{\mathcal{U}(\pi_2)\mathbf{1}_{z^u} + (1 - \mathcal{U}(\pi_2))\,\mathbf{1}_{z^l}\} \\ &\sim (a\mathcal{U}(\pi_1) + (1-a)\mathcal{U}(\pi_2))\,\mathbf{1}_{z^u} + (1 - \{a\mathcal{U}(\pi_1) + (1-a)\mathcal{U}(\pi_2)\})\,\mathbf{1}_{z^l}. \end{aligned}$$

Now let us show the converse, i.e. if $\succeq$ can be represented by a linear utility index $\mathcal{U}$, then it must satisfy the Archimedean Axiom and the Substitution Axiom. In order to show the Archimedean Axiom, we pick $\pi_1 \succ \pi_2 \succ \pi_3$, which means that $\mathcal{U}(\pi_1) > \mathcal{U}(\pi_2) > \mathcal{U}(\pi_3)$, and must find numbers $a, b \in (0,1)$ such that

$$a\pi_1 + (1-a)\pi_3 \succ \pi_2 \succ b\pi_1 + (1-b)\pi_3,$$

i.e. that

$$\mathcal{U}\left(a\pi_1 + (1-a)\pi_3\right) > \mathcal{U}(\pi_2) > \mathcal{U}\left(b\pi_1 + (1-b)\pi_3\right).$$

Define the number a by

$$a = 1 - \frac{1}{2}\frac{\mathcal{U}(\pi_1) - \mathcal{U}(\pi_2)}{\mathcal{U}(\pi_1) - \mathcal{U}(\pi_3)}.$$

Then $a \in (0,1)$ and by linearity of $\mathcal{U}$ we get

$$\begin{aligned}
\mathcal{U}\left(a\pi_1 + (1-a)\pi_3\right) &= a\mathcal{U}(\pi_1) + (1-a)\mathcal{U}(\pi_3) \\
&= \mathcal{U}(\pi_1) + (1-a)\left(\mathcal{U}(\pi_3) - \mathcal{U}(\pi_1)\right) \\
&= \mathcal{U}(\pi_1) - \frac{1}{2}\left(\mathcal{U}(\pi_1) - \mathcal{U}(\pi_2)\right) \\
&= \frac{1}{2}\left(\mathcal{U}(\pi_1) + \mathcal{U}(\pi_2)\right) \\
&> \mathcal{U}(\pi_2).
\end{aligned}$$

Similarly for b.

In order to show the Substitution Axiom, we take $\pi_1, \pi_2, \pi_3 \in \mathcal{P}(Z)$ and any number $a \in (0,1]$. We must show that $\pi_1 \succ \pi_2$ if and only if $a\pi_1 + (1-a)\pi_3 \succ a\pi_2 + (1-a)\pi_3$, i.e.

$$\mathcal{U}(\pi_1) > \mathcal{U}(\pi_2) \quad \Leftrightarrow \quad \mathcal{U}\left(a\pi_1 + (1-a)\pi_3\right) > \mathcal{U}\left(a\pi_2 + (1-a)\pi_3\right).$$

This follows immediately by linearity of $\mathcal{U}$:

$$\begin{aligned}
\mathcal{U}\left(a\pi_1 + (1-a)\pi_3\right) &= a\mathcal{U}(\pi_1) + \mathcal{U}\left((1-a)\pi_3\right) \\
&> a\mathcal{U}(\pi_2) + \mathcal{U}\left((1-a)\pi_3\right) \\
&= \mathcal{U}\left(a\pi_2 + (1-a)\pi_3\right)
\end{aligned}$$

with the inequality holding if and only if $\mathcal{U}(\pi_1) > \mathcal{U}(\pi_2)$. Similarly, we can show that $\pi_1 \sim \pi_2$ if and only if $a\pi_1 + (1-a)\pi_3 \sim a\pi_2 + (1-a)\pi_3$. □

The next theorem shows which utility functions that represent the same preference relation. The proof is left for the reader as Exercise 5.1.

Theorem 5.3 *A utility function for a given preference relation is only determined up to a strictly increasing affine transformation, i.e. if u is a utility function for $\succeq$, then v will be so if and only if there exist constants $a > 0$ and b such that $v(z) = au(z) + b$ for all $z \in Z$.*

If one utility function is an affine function of another, we will say that they are equivalent. Note that an easy consequence of this theorem is that it does not really matter whether the utility is positive or negative. At first, you might find negative utility strange but we can always add a sufficiently large positive constant without affecting the ranking of different consumption plans.

Suppose $\mathcal{U}$ is a utility index with an associated utility function u. If f is any strictly increasing transformation, then $V = f \circ \mathcal{U}$ is also a utility index for the same preferences, but $f \circ u$ is only the utility function for V if f is affine.

The expected utility associated with a probability distribution π on Z is $\sum_{z \in Z} \pi(z)u(z)$. Recall that the probability distributions we consider correspond to consumption plans. Given a consumption plan, i.e. a random variable c, the associated probability distribution is defined by the probabilities

$$\pi(z) = \mathbb{P}\left(\{\omega \in \Omega | c(\omega) = z\}\right) = \sum_{\omega \in \Omega : c(\omega) = z} p_\omega.$$

The expected utility associated with the consumption plan c is therefore

$$\mathrm{E}[u(c)] = \sum_{\omega \in \Omega} p_\omega u(c(\omega)) = \sum_{z \in Z} \sum_{\omega \in \Omega : c(\omega) = z} p_\omega u(z) = \sum_{z \in Z} \pi(z)u(z).$$

Of course, if c is a risk-free consumption plan in the sense that a z exists such that $c(\omega) = z$ for all ω, then the expected utility is $\mathrm{E}[u(c)] = u(z)$. With a slight abuse of notation we will just write this as $u(c)$.

5.4.2 Some technical issues

Infinite Z. What if Z is infinite, e.g. $Z = \mathbb{R}_+ \equiv [0, \infty)$? It can be shown that in this case a preference relation has an expected utility representation if the Archimedean Axiom, the Substitution Axiom, an additional axiom ("the sure thing principle"), and "some technical conditions" are satisfied. Fishburn (1970) gives the details.

Expected utility in this case: $\mathrm{E}[u(c)] = \int_Z u(z)\pi(z)\, dz$, where π is a probability density function derived from the consumption plan c.

Boundedness of expected utility. Suppose u is unbounded from above and $\mathbb{R}_+ \subseteq Z$. Then there exists $(z_n)_{n=1}^\infty \subseteq Z$ with $z_n \to \infty$ and $u(z_n) \geq 2^n$. Expected utility of consumption plan π_1 with $\pi_1(z_n) = 1/2^n$:

$$\sum_{n=1}^{\infty} u(z_n)\pi_1(z_n) \geq \sum_{n=1}^{\infty} 2^n \frac{1}{2^n} = \infty.$$

If π_2, π_3 are such that $\pi_1 \succ \pi_2 \succ \pi_3$, then the expected utility of π_2 and π_3 must be finite. But for no $b \in (0, 1)$ do we have

$$\pi_2 \succ b\pi_1 + (1 - b)\pi_3 \qquad [\text{expected utility} = \infty].$$

- no problem if Z is finite
- no problem if $\mathbb{R}_+ \subseteq Z$, u is concave, and consumption plans have finite expectations:

  u concave $\Rightarrow$ u is differentiable in some point b and

  $$u(z) \leq u(b) + u'(b)(z - b), \quad \forall z \in Z.$$

  If the consumption plan c has finite expectations, then

  $$\mathrm{E}[u(c)] \leq \mathrm{E}[u(b) + u'(b)(c - b)] = u(b) + u'(b)\left(\mathrm{E}[c] - b\right) < \infty.$$

z	0	1	5
π_1	0	1	0
π_2	0.01	0.89	0.1
π_3	0.9	0	0.1
π_4	0.89	0.11	0

Table 5.4: The probability distributions used in the illustration of the Allais Paradox.

Subjective probability. We have taken the probabilities of the states of nature as exogenously given, i.e. as *objective* probabilities. However, in real life individuals often have to form their own probabilities about many events, i.e. they form *subjective* probabilities. Although the analysis is a bit more complicated, Savage (1954) and Anscombe and Aumann (1963) show that the results we developed above carry over to the case of subjective probabilities. For an introduction to this analysis, see Kreps (1990, Ch. 3).

5.4.3 Are the axioms reasonable?

The validity of the Substitution Axiom, which is necessary for obtaining the expected utility representation, has been intensively discussed in the literature. Some researchers have conducted experiments in which the decisions made by the participating individuals conflict with the Substitution Axiom.

The most famous challenge is the so-called Allais Paradox named after Allais (1953). Here is one example of the paradox. Suppose $Z = \{0, 1, 5\}$. Consider the consumption plans in Table 5.4. The Substitution Axiom implies that $\pi_1 \succ \pi_2 \Rightarrow \pi_4 \succ \pi_3$. This can be seen from the following:

$$0.11(\$1) + 0.89\boxed{(\$1)} \sim \pi_1 \succ \pi_2 \sim 0.11\left(\frac{1}{11}(\$0) + \frac{10}{11}(\$5)\right) + 0.89\boxed{(\$1)} \quad \Rightarrow$$

$$\underbrace{0.11(\$1) + 0.89\boxed{(\$0)}}_{\pi_4 \sim} \succ 0.11\left(\frac{1}{11}(\$0) + \frac{10}{11}(\$5)\right) + 0.89\boxed{(\$0)} \sim \underbrace{0.9(\$0) + 0.1(\$5)}_{\pi_3 \sim}$$

Nevertheless individuals preferring π_1 to π_2 often choose π_3 over π_4. Apparently people tend to over-weight small probability events, e.g. (\$0) in π_2.

Other "problems":

- the "framing" of possible choices, i.e. the way you get the alternatives presented, seem to affect decisions
- models assume individuals have unlimited rationality

5.5 Risk aversion

In this section we focus on the attitudes towards risk reflected by the preferences of an individual. We assume that the preferences can be represented by a utility function u and that u is strictly increasing so that the individual is "greedy," i.e. prefers high consumption to low consumption. We assume that the utility function is defined on some interval Z of $\mathbb{R}$, e.g. $Z = \mathbb{R}_+ \equiv [0, \infty)$.

5.5.1 Risk attitudes

Fix a consumption level $c \in Z$. Consider a random variable ε with $\mathrm{E}[\varepsilon] = 0$. We can think of $c + \varepsilon$ as a random variable representing a consumption plan with consumption $c + \varepsilon(\omega)$ if state ω is realized. Note that $\mathrm{E}[c+\varepsilon] = c$. Such a random variable ε is called a fair gamble or a zero-mean risk.

An individual is said to be (strictly) **risk-averse** if she for all $c \in Z$ and all fair gambles ε (strictly) prefers the sure consumption level c to $c + \varepsilon$. In other words, a risk-averse individual rejects all fair gambles. Similarly, an individual is said to be (strictly) **risk-loving** if she for all $c \in Z$ (strictly) prefers $c + \varepsilon$ to c, and said to be **risk-neutral** if she for all $c \in Z$ is indifferent between accepting any fair gamble or not. Of course, individuals may be neither risk-averse, risk-neutral, or risk-loving, for example if they reject fair gambles around some values of c and accept fair gambles around other values of c. Individuals may be locally risk-averse, locally risk-neutral, and locally risk-loving. Since it is generally believed that individuals are risk-averse, we focus on preferences exhibiting that feature.

We can think of any consumption plan c as the sum of its expected value $\mathrm{E}[c]$ and a fair gamble $\varepsilon = c - \mathrm{E}[c]$. It follows that an individual is risk-averse if she prefers the sure consumption $\mathrm{E}[c]$ to the random consumption c, i.e. if $u(\mathrm{E}[c]) \geq \mathrm{E}[u(c)]$. By Jensen's Inequality, this is true exactly when u is a concave function and the strict inequality holds if u is strictly concave and c is a non-degenerate random variable, i.e. it does not have the same value in all states. Recall that $u : Z \to \mathbb{R}$ concave means that for all $z_1, z_2 \in Z$ and all $a \in (0, 1)$ we have

$$u\left(az_1 + (1-a)z_2\right) \geq au(z_1) + (1-a)u(z_2).$$

If the strict inequality holds in all cases, the function is said to be strictly concave. By the above argument, we have the following theorem:

Theorem 5.4 *An individual with a utility function u is (strictly) risk-averse if and only if u is (strictly) concave.*

Similarly, an individual is (strictly) risk-loving if and only if the utility function is (strictly) convex. An individual is risk-neutral if and only if the utility function is affine.

5.5.2 Quantitative measures of risk aversion

We will focus on utility functions that are continuous and twice differentiable on the interior of Z. By our assumption of greedy individuals, we then have $u' > 0$, and the concavity of the utility function for risk-averse investors is then equivalent to $u'' \leq 0$.

The **certainty equivalent** of the random consumption plan c is defined as the $c^* \in Z$ such that

$$u(c^*) = \mathrm{E}[u(c)],$$

i.e. the individual is just as satisfied getting the consumption level c^* for sure as getting the random consumption c. With $Z \subseteq \mathbb{R}$, c^* uniquely exists due to our assumptions that u is continuous and strictly increasing. From the definition of the certainty equivalent it is clear that an individual will rank consumption plans according to their certainty equivalents.

For a risk-averse individual we have the certainty equivalent c^* of a consumption plan is smaller than the expected consumption level $\mathrm{E}[c]$. The **risk premium** associated with the consumption plan c is defined as $\lambda(c) = \mathrm{E}[c] - c^*$ so that

$$\mathrm{E}[u(c)] = u(c^*) = u(\mathrm{E}[c] - \lambda(c)).$$

The risk premium is the consumption the individual is willing to give up in order to eliminate the uncertainty.

The degree of risk aversion is associated with u'', but a good measure of risk aversion should be invariant to strictly positive, affine transformations. This is satisfied by the Arrow-Pratt measures of risk aversion defined as follows. The **Absolute Risk Aversion** is given by

$$\mathrm{ARA}(c) = -\frac{u''(c)}{u'(c)}. \tag{5.2}$$

The **Relative Risk Aversion** is given by

$$\mathrm{RRA}(c) = -\frac{cu''(c)}{u'(c)} = c\,\mathrm{ARA}(c). \tag{5.3}$$

We can link the Arrow-Pratt measures to the risk premium in the following way. Let $\bar{c} \in Z$ denote some fixed consumption level and let ε be a fair gamble. The resulting consumption plan is then $c = \bar{c} + \varepsilon$. Denote the corresponding risk premium by $\lambda(\bar{c}, \varepsilon)$ so that

$$\mathrm{E}[u(\bar{c} + \varepsilon)] = u(c^*) = u\left(\bar{c} - \lambda(\bar{c}, \varepsilon)\right). \tag{5.4}$$

We can approximate the left-hand side of (5.4) by

$$\mathrm{E}[u(\bar{c} + \varepsilon)] \approx \mathrm{E}\left[u(\bar{c}) + \varepsilon u'(\bar{c}) + \frac{1}{2}\varepsilon^2 u''(\bar{c})\right] = u(\bar{c}) + \frac{1}{2}\,\mathrm{Var}[\varepsilon]u''(\bar{c}),$$

using $\mathrm{E}[\varepsilon] = 0$ and $\mathrm{Var}[\varepsilon] = \mathrm{E}[\varepsilon^2] - \mathrm{E}[\varepsilon]^2 = \mathrm{E}[\varepsilon^2]$, and we can approximate the right-hand side of (5.4) by

$$u\left(\bar{c} - \lambda(\bar{c}, \varepsilon)\right) \approx u(\bar{c}) - \lambda(\bar{c}, \varepsilon)u'(\bar{c}).$$

Hence we can write the risk premium as

$$\lambda(\bar{c}, \varepsilon) \approx -\frac{1}{2}\,\mathrm{Var}[\varepsilon]\frac{u''(\bar{c})}{u'(\bar{c})} = \frac{1}{2}\,\mathrm{Var}[\varepsilon]\,\mathrm{ARA}(\bar{c}).$$

Of course, the approximation is more accurate for "small" gambles. Thus the risk premium for a small fair gamble around $\bar{c}$ is roughly proportional to the absolute risk aversion at $\bar{c}$. We see that the absolute risk aversion $\mathrm{ARA}(\bar{c})$ is constant if and only if $\lambda(\bar{c}, \varepsilon)$ is independent of $\bar{c}$.

Loosely speaking, the absolute risk aversion $\mathrm{ARA}(c)$ measures the aversion to a fair gamble of a given dollar amount around c, such as a gamble where there is an equal probability of winning or loosing 1000 dollars. Since we expect that a wealthy investor will be less averse to that gamble than a poor investor, the absolute risk aversion is expected to be a decreasing function of wealth. Note that

$$\mathrm{ARA}'(c) = -\frac{u'''(c)u'(c) - u''(c)^2}{u'(c)^2} = \left(\frac{u''(c)}{u'(c)}\right)^2 - \frac{u'''(c)}{u'(c)} < 0 \quad \Rightarrow \quad u'''(c) > 0,$$

that is, a positive third-order derivative of u is necessary for the utility function u to exhibit decreasing absolute risk aversion.

Now consider a "multiplicative" fair gamble around $\bar{c}$ in the sense that the resulting consumption plan is $c = \bar{c}\,(1+\varepsilon) = \bar{c} + \bar{c}\varepsilon$, where $\mathrm{E}[\varepsilon] = 0$. The risk premium is then

$$\lambda(\bar{c}, \bar{c}\varepsilon) \approx \frac{1}{2}\,\mathrm{Var}[\bar{c}\varepsilon]\,\mathrm{ARA}(\bar{c}) = \frac{1}{2}\bar{c}^2\,\mathrm{Var}[\varepsilon]\,\mathrm{ARA}(\bar{c}) = \frac{1}{2}\bar{c}\,\mathrm{Var}[\varepsilon]\,\mathrm{RRA}(\bar{c})$$

implying that

$$\frac{\lambda(\bar{c}, \bar{c}\varepsilon)}{\bar{c}} \approx \frac{1}{2}\,\mathrm{Var}[\varepsilon]\,\mathrm{RRA}(\bar{c}). \tag{5.5}$$

The fraction of consumption you require to engage in the multiplicative risk is thus (roughly) proportional to the relative risk aversion at $\bar{c}$. Note that utility functions with constant or decreasing (or even modestly increasing) relative risk aversion will display decreasing absolute risk aversion.

Some authors use terms like risk tolerance and risk cautiousness. The **absolute risk tolerance** at c is simply the reciprocal of the absolute risk aversion, i.e.

$$\mathrm{ART}(c) = \frac{1}{\mathrm{ARA}(c)} = -\frac{u'(c)}{u''(c)}.$$

Similarly, the relative risk tolerance is the reciprocal of the relative risk aversion. The **risk cautiousness** at c is defined as the rate of change in the absolute risk tolerance, i.e. $\mathrm{ART}'(c)$.

5.5.3 Comparison of risk aversion between individuals

An individual with utility function u is said to be *more risk-averse* than an individual with utility function v if for any consumption plan c and any fixed $\bar{c} \in Z$ with $\mathrm{E}[u(c)] \geq u(\bar{c})$, we have $\mathrm{E}[v(c)] \geq v(\bar{c})$. So the v-individual will accept all gambles that the u-individual will accept – and possibly some more. Pratt (1964) has shown the following theorem:

Theorem 5.5 *Suppose u and v are twice continuously differentiable and strictly increasing. Then the following conditions are equivalent:*

(a) u is more risk-averse than v,

(b) $\mathrm{ARA}_u(c) \geq \mathrm{ARA}_v(c)$ for all $c \in Z$,

(c) a strictly increasing and concave function f exists such that $u = f \circ v$.

Proof: First let us show (a) ⇒ (b): Suppose u is more risk-averse than v, but that $\mathrm{ARA}_u(\hat{c}) < \mathrm{ARA}_v(\hat{c})$ for some $\hat{c} \in Z$. Since ARA_u and ARA_v are continuous, we must then have that $\mathrm{ARA}_u(c) < \mathrm{ARA}_v(c)$ for all c in an interval around $\hat{c}$. Then we can surely find a small gamble around $\hat{c}$, which the u-individual will accept, but the v-individual will reject. This contradicts the assumption in (a).

Next, we show (b) ⇒ (c): Since v is strictly increasing, it has an inverse v^{-1} and we can define a function f by $f(x) = u\left(v^{-1}(x)\right)$. Then clearly $f(v(c)) = u(c)$ so that $u = f \circ v$. The first-order derivative of f is

$$f'(x) = \frac{u'\left(v^{-1}(x)\right)}{v'\left(v^{-1}(x)\right)},$$

which is positive since u and v are strictly increasing. Hence, f is strictly increasing. The second-order derivative is

$$\begin{aligned} f''(x) &= \frac{u''\left(v^{-1}(x)\right) - \left\{v''\left(v^{-1}(x)\right) u'\left(v^{-1}(x)\right)\right\} / v'\left(v^{-1}(x)\right)}{v'\left(v^{-1}(x)\right)^2} \\ &= \frac{u'\left(v^{-1}(x)\right)}{v'\left(v^{-1}(x)\right)^2} \left(\mathrm{ARA}_v\left(v^{-1}(x)\right) - \mathrm{ARA}_u\left(v^{-1}(x)\right)\right). \end{aligned}$$

From (b), it follows that $f''(x) < 0$, hence f is concave.

Finally, we show that (c) $\Rightarrow$ (a): assume that for some consumption plan c and some $\bar{c} \in Z$, we have $\mathrm{E}[u(c)] \geq u(\bar{c})$ but $\mathrm{E}[v(c)] < v(\bar{c})$. We want to arrive at a contradiction.

$$\begin{aligned} f\left(v(\bar{c})\right) = u(\bar{c}) &\leq \mathrm{E}[u(c)] = \mathrm{E}[f(v(c))] \\ &< f\left(\mathrm{E}[v(c)]\right) \\ &< f\left(v(\bar{c})\right), \end{aligned}$$

where we use the concavity of f and Jensen's Inequality to go from the first to the second line, and we use that f is strictly increasing to go from the second to the third line. Now the contradiction is clear. □

5.6 Utility functions in models and in reality

5.6.1 Frequently applied utility functions

CRRA utility. (Also known as power utility or isoelastic utility.) Utility functions $u(c)$ in this class are defined for $c \geq 0$:

$$u(c) = \frac{c^{1-\gamma}}{1-\gamma}, \tag{5.6}$$

where $\gamma > 0$ and $\gamma \neq 1$. Since

$$u'(c) = c^{-\gamma} \quad \text{and} \quad u''(c) = -\gamma c^{-\gamma-1},$$

the absolute and relative risk aversions are given by

$$\mathrm{ARA}(c) = -\frac{u''(c)}{u'(c)} = \frac{\gamma}{c}, \qquad \mathrm{RRA}(c) = c\,\mathrm{ARA}(c) = \gamma.$$

The relative risk aversion is constant across consumption levels c, hence the name CRRA (Constant Relative Risk Aversion) utility. Note that $u'(0+) \equiv \lim_{c\to 0} u'(c) = \infty$ with the consequence that an optimal solution will have the property that consumption/wealth c will be strictly above 0 with probability one. Hence, we can ignore the very appropriate non-negativity constraint on consumption since the constraint will never be binding. Furthermore, $u'(\infty) \equiv \lim_{c\to\infty} u'(c) = 0$.

Some authors assume a utility function of the form $u(c) = c^{1-\gamma}$, which only makes sense for $\gamma \in (0, 1)$. However, empirical studies indicate that most investors have a relative risk aversion above 1, cf. the discussion below. The absolute risk tolerance is linear in c:

$$\mathrm{ART}(c) = \frac{1}{\mathrm{ARA}(c)} = \frac{c}{\gamma}.$$

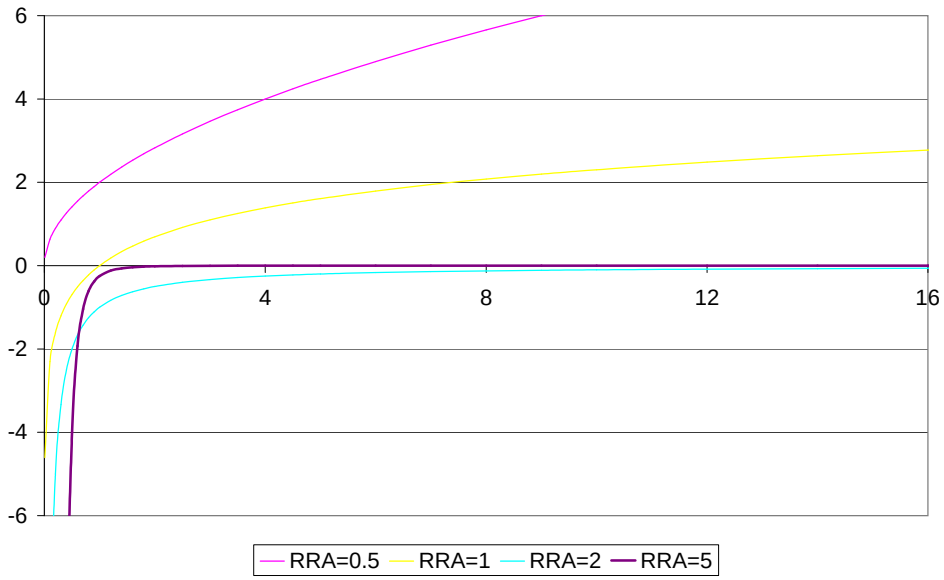


Figure 5.1: Some CRRA utility functions.

Except for a constant, the utility function

$$u(c) = \frac{c^{1-\gamma} - 1}{1 - \gamma}$$

is identical to the utility function specified in (5.6). The two utility functions are therefore equivalent in the sense that they generate the same optimal choices. The advantage in using the latter definition is that this function has a well-defined limit as $\gamma \to 1$. From l'Hôspital's rule we have that

$$\lim_{\gamma \to 1} \frac{c^{1-\gamma} - 1}{1 - \gamma} = \lim_{\gamma \to 1} \frac{-c^{1-\gamma} \ln c}{-1} = \ln c,$$

which is the important special case of **logarithmic utility**. When we consider CRRA utility, we will assume the simpler version (5.6), but we will use the fact that we can obtain the optimal strategies of a log-utility investor as the limit of the optimal strategies of the general CRRA investor as $\gamma \to 1$.

Some CRRA utility functions are illustrated in Figure 5.1.

HARA utility. (Also known as extended power utility.) The absolute risk aversion for CRRA utility is hyperbolic in c. More generally a utility function is said to be a HARA (Hyperbolic Absolute Risk Aversion) utility function if

$$\mathrm{ARA}(c) = -\frac{u''(c)}{u'(c)} = \frac{1}{\alpha c + \beta}$$

for some constants α, β such that $\alpha c + \beta > 0$ for all relevant c. HARA utility functions are sometimes referred to as affine (or linear) risk tolerance utility functions since the absolute risk tolerance is

$$\mathrm{ART}(c) = \frac{1}{\mathrm{ARA}(c)} = \alpha c + \beta.$$

The risk cautiousness is $\mathrm{ART}'(c) = \alpha$.

How do the HARA utility functions look like? First, let us take the case $\alpha = 0$, which implies that the absolute risk aversion is constant (so-called CARA utility) and β must be positive.

$$\frac{d(\ln u'(c))}{dc} = \frac{u''(c)}{u'(c)} = -\frac{1}{\beta}$$

implies that

$$\ln u'(c) = -\frac{c}{\beta} + k_1 \quad \Rightarrow \quad u'(c) = e^{k_1} e^{-c/\beta}$$

for some constant k_1. Hence,

$$u(c) = -\frac{1}{\beta} e^{k_1} e^{-c/\beta} + k_2$$

for some other constant k_2. Applying the fact that increasing affine transformations do not change decisions, the basic representative of this class of utility functions is the **negative exponential utility** function

$$u(c) = -e^{-ac}, \quad c \in \mathbb{R}, \tag{5.7}$$

where the parameter $a = 1/\beta$ is the absolute risk aversion. Constant absolute risk aversion is certainly not very reasonable. Nevertheless, the negative exponential utility function is sometimes used for computational purposes in connection with normally distributed returns, e.g. in one-period models.

Next, consider the case $\alpha \neq 0$. Applying the same procedure as above we find

$$\frac{d(\ln u'(c))}{dc} = \frac{u''(c)}{u'(c)} = -\frac{1}{c+\beta} \quad \Rightarrow \quad \ln u'(c) = -\frac{1}{\alpha} \ln(\alpha c + \beta) + k_1$$

so that

$$u'(c) = e^{k_1} \exp\left\{ -\frac{1}{\alpha} \ln(\alpha c + \beta) \right\} = e^{k_1} (\alpha c + \beta)^{-1/\alpha}. \tag{5.8}$$

For $\alpha = 1$ this implies that

$$u(c) = e^{k_1} \ln(c + \beta) + k_2.$$

The basic representative of such utility functions is the **extended log utility** function

$$u(c) = \ln(c - \bar{c}), \quad c > \bar{c}, \tag{5.9}$$

where we have replaced β by $-\bar{c}$. For $\alpha \neq 1$, Equation (5.8) implies that

$$u(c) = \frac{1}{\alpha} e^{k_1} \frac{1}{1 - \frac{1}{\alpha}} (\alpha c + \beta)^{1 - 1/\alpha} + k_2.$$

For $\alpha < 0$, we can write the basic representative is

$$u(c) = -(\bar{c} - c)^{1-\gamma}, \quad c < \bar{c}, \tag{5.10}$$

where $\gamma = 1/\alpha < 0$. We can think of $\bar{c}$ as a satiation level and call this subclass **satiation HARA utility** functions. The absolute risk aversion is

$$\mathrm{ARA}(c) = \frac{-\gamma}{\bar{c} - c},$$

which is increasing in c, conflicting with intuition and empirical studies. Some older financial models used the quadratic utility function, which is the special case with $\gamma = -1$ so that $u(c) = -(\bar{c} - c)^2$. An equivalent utility function is $u(c) = c - ac^2$.

For $\alpha > 0$ (and $\alpha \neq 1$), the basic representative is

$$u(c) = \frac{(c-\bar{c})^{1-\gamma}}{1-\gamma}, \quad c > \bar{c}, \tag{5.11}$$

where $\gamma = 1/\alpha > 0$. The limit as $\gamma \to 1$ of the equivalent utility function $\frac{(c-\bar{c})^{1-\gamma}-1}{1-\gamma}$ is equal to the extended log utility function $u(c) = \ln(c-\bar{c})$. We can think of $\bar{c}$ as a subsistence level of wealth or consumption (which makes sense only if $\bar{c} \geq 0$) and refer to this subclass as **subsistence HARA utility** functions. The absolute and relative risk aversions are

$$\mathrm{ARA}(c) = \frac{\gamma}{c-\bar{c}}, \quad \mathrm{RRA}(c) = \frac{\gamma c}{c-\bar{c}} = \frac{\gamma}{1-(\bar{c}/c)},$$

which are both decreasing in c. The relative risk aversion approaches ∞ for $c \to \bar{c}$ and decreases to the constant γ for $c \to \infty$. Clearly, for $\bar{c} = 0$, we are back to the CRRA utility functions so that these also belong to the HARA family.

Mean-variance preferences. For some problems it is convenient to assume that the expected utility associated with an uncertain consumption plan only depends on the expected value and the variance of the consumption plan. This is certainly true if the consumption plan is a normally distributed random variable since its probability distribution is fully characterized by the mean and variance. However, it is generally not appropriate to use a normal distribution for consumption (or wealth or asset returns).

For a quadratic utility function, $u(c) = c - ac^2$, the expected utility is

$$\mathrm{E}[u(c)] = \mathrm{E}\left[c - ac^2\right] = \mathrm{E}[c] - a\,\mathrm{E}\left[c^2\right] = \mathrm{E}[c] - a\left(\mathrm{Var}[c] + \mathrm{E}[c]^2\right),$$

which is indeed a function of the expected value and the variance of the consumption plan. Alas, the quadratic utility function is inappropriate for several reasons. Most importantly, it exhibits increasing absolute risk aversion.

For a general utility function the expected utility of a consumption plan will depend on all moments. This can be seen by the Taylor expansion of $u(c)$ around the expected consumption, $\mathrm{E}[c]$:

$$u(c) = u(\mathrm{E}[c]) + u'(\mathrm{E}[c])(c-\mathrm{E}[c]) + \frac{1}{2}u''(\mathrm{E}[c])(c-\mathrm{E}[c])^2 + \sum_{n=3}^{\infty}\frac{1}{n!}u^{(n)}(\mathrm{E}[c])(c-\mathrm{E}[c])^n,$$

where $u^{(n)}$ is the n'th derivative of u. Taking expectations, we get

$$\mathrm{E}[u(c)] = u(\mathrm{E}[c]) + \frac{1}{2}u''(\mathrm{E}[c])\,\mathrm{Var}[c] + \sum_{n=3}^{\infty}\frac{1}{n!}u^{(n)}(\mathrm{E}[c])\,\mathrm{E}\left[(c-\mathrm{E}[c])^n\right].$$

Here $\mathrm{E}\left[(c-\mathrm{E}[c])^n\right]$ is the central moment of order n. The variance is the central moment of order 2. Obviously, a greedy investor (which just means that u is increasing) will prefer higher expected consumption to lower for fixed central moments of order 2 and higher. Moreover, a risk-averse investor (so that $u'' < 0$) will prefer lower variance of consumption to higher for fixed expected consumption and fixed central moments of order 3 and higher. But when the central moments of order 3 and higher are not the same for all alternatives, we cannot just evaluate them on the basis of their expectation and variance. With quadratic utility, the derivatives of u of order 3 and higher are zero so there it works. In general, mean-variance preferences can only serve as an approximation of the true utility function.

5.6.2 What do we know about individuals' risk aversion?

From our discussion of risk aversion and various utility functions we expect that individuals are risk averse and exhibit decreasing absolute risk aversion. But can this be supported by empirical evidence? Do individuals have constant relative risk aversion? And what is a reasonable level of risk aversion for individuals?

You can get an idea of the risk attitudes of an individual by observing how they choose between risky alternatives. Some researchers have studied this by setting up "laboratory experiments" in which they present some risky alternatives to a group of individuals and simply see what they prefer. Some of these experiments suggest that expected utility theory is frequently violated, see e.g. Grether and Plott (1979). However, laboratory experiments are problematic for several reasons. You cannot be sure that individuals will make the same choice in what they know is an experiment as they would in real life. It is also hard to formulate alternatives that resemble the rather complex real-life decisions. It seems more fruitful to study actual data on how individuals have acted confronted with real-life decision problems under uncertainty. A number of studies do that.

Friend and Blume (1975) analyze data on household asset holdings. They conclude that the data is consistent with individuals having roughly constant relative risk aversion and that the coefficients of relative risk aversion are "on average well in excess of one and probably in excess of two" (quote from page 900 in their paper). Pindyck (1988) finds support of a relative risk aversion between 3 and 4 in a structural model of the reaction of stock prices to fundamental variables.

Other studies are based on insurance data. Using U.S. data on so-called property/liability insurance, Szpiro (1986) finds support of CRRA utility with a relative risk aversion coefficient between 1.2 and 1.8. Cicchetti and Dubin (1994) work with data from the U.S. on whether individuals purchased an insurance against the risk of trouble with their home telephone line. They conclude that the data is consistent with expected utility theory and that a subsistence HARA utility function performs better than log utility or negative exponential utility.

Ogaki and Zhang (2001) study data on individual food consumption from Pakistan and India and conclude that relative risk aversion is decreasing for poor individuals, which is consistent with a subsistence HARA utility function.

It is an empirical fact that even though consumption and wealth have increased tremendously over the years, the magnitude of real rates of return has not changed dramatically. As indicated by (5.5) relative risk premia are approximately proportional to the relative risk aversion. As we shall see in later chapters, basic asset pricing theory implies that relative risk premia on financial assets (in terms of expected real return in excess of the real risk-free return) will be proportional to the "average" relative risk aversion in the economy. If the "average" relative risk aversion was significantly decreasing (increasing) in the level of consumption or wealth, we should have seen decreasing (increasing) real returns on risky assets in the past. The data seems to be consistent with individuals having "on average" close to CRRA utility.

To get a feeling of what a given risk aversion really means, suppose you are confronted with two consumption plans. One plan is a sure consumption of $\bar{c}$, the other plan gives you $(1-\alpha)\bar{c}$ with probability 0.5 and $(1+\alpha)\bar{c}$ with probability 0.5. If you have a CRRA utility function

$\gamma = \text{RRA}$	$\alpha = 1\%$	$\alpha = 10\%$	$\alpha = 50\%$
0.5	0.00%	0.25%	6.70%
1	0.01%	0.50%	13.40%
2	0.01%	1.00%	25.00%
5	0.02%	2.43%	40.72%
10	0.05%	4.42%	46.00%
20	0.10%	6.76%	48.14%
50	0.24%	8.72%	49.29%
100	0.43%	9.37%	49.65%

Table 5.5: Relative risk premia for a fair gamble of the fraction α of your consumption.

$u(c) = c^{1-\gamma}/(1-\gamma)$, the certainty equivalent c^* of the risky plan is determined by

$$\frac{1}{1-\gamma}(c^*)^{1-\gamma} = \frac{1}{2}\frac{1}{1-\gamma}((1-\alpha)\bar{c})^{1-\gamma} + \frac{1}{2}\frac{1}{1-\gamma}((1+\alpha)\bar{c})^{1-\gamma},$$

which implies that

$$c^* = \left(\frac{1}{2}\right)^{1/(1-\gamma)}\left[(1-\alpha)^{1-\gamma} + (1+\alpha)^{1-\gamma}\right]^{1/(1-\gamma)}\bar{c}.$$

The risk premium $\lambda(\bar{c}, \alpha)$ is

$$\lambda(\bar{c}, \alpha) = \bar{c} - c^* = \left(1 - \left(\frac{1}{2}\right)^{1/(1-\gamma)}\left[(1-\alpha)^{1-\gamma} + (1+\alpha)^{1-\gamma}\right]^{1/(1-\gamma)}\right)\bar{c}.$$

Both the certainty equivalent and the risk premium are thus proportional to the consumption level $\bar{c}$. The relative risk premium $\lambda(\bar{c}, \alpha)/\bar{c}$ is simply one minus the relative certainty equivalent $c^*/\bar{c}$. These equations assume $\gamma \neq 1$. In Exercise 5.5 you are asked to find the certainty equivalent and risk premium for log-utility corresponding to $\gamma = 1$.

Table 5.5 shows the relative risk premium for various values of the relative risk aversion coefficient γ and various values of α, the "size" of the risk. For example, an individual with $\gamma = 5$ is willing to sacrifice 2.43% of the safe consumption in order to avoid a fair gamble of 10% of that consumption level. Of course, even extremely risk averse individuals will not sacrifice more than they can loose but in some cases it is pretty close. Looking at these numbers, it is hard to believe in γ-values outside, say, $[1, 10]$. In Exercise 5.6 you are asked to compare the exact relative risk premia shown in the table with the approximate risk premia given by (5.5).

5.7 Preferences for multi-date consumption plans

Above we implicitly considered preferences for consumption at one given future point in time. We need to generalize the ideas and results to settings with consumption at several dates. In one-period models individuals can consume both at time 0 (beginning-of-period) and at time 1 (end-of-period). In multi-period models individuals can consume either at each date in the discrete time set $\mathcal{T} = \{0, 1, 2, \ldots, T\}$ or at each date in the continuous time set $\mathcal{T} = [0, T]$. In any case a consumption plan is a stochastic process $c = (c_t)_{t\in\mathcal{T}}$ where each c_t is a random variable representing the state-dependent level of consumption at time t.

Consider the discrete-time case and, for each t, let $Z_t \subseteq \mathbb{R}$ denote the set of all possible consumption levels at date t and define $Z = Z_0 \times Z_1 \times \cdots \times Z_T \subseteq \mathbb{R}^{T+1}$, then any consumption plan c can again be represented by a probability distribution π on the set Z. For finite Z, we can again apply Theorem 5.1 so that under the relevant axioms, we can represent preferences by a utility index $\mathcal{U}$, which to each consumption plan $(c_t)_{t\in\mathcal{T}} = (c_0, c_1, \ldots, c_T)$ attaches a real number $\mathcal{U}(c_0, c_1, \ldots, c_T)$ with higher numbers to the more preferred consumption plans. If we further impose the Substitution Axiom, Theorem 5.2 ensures an expected utility representation, i.e. the existence of a utility function $U : Z \to \mathbb{R}$ so that consumption plans are ranked according to their expected utility, i.e.

$$\mathcal{U}(c_0, c_1, \ldots, c_T) = \mathrm{E}\left[U(c_0, c_1, \ldots, c_T)\right] \equiv \sum_{\omega\in\Omega} p_\omega U\left(c_0, c_1(\omega), \ldots, c_T(\omega)\right).$$

We can call U a **multi-date utility function** since it depends on the consumption levels at all dates. Again this result can be extended to the case of an infinite Z, e.g. $Z = \mathbb{R}_+^{T+1}$, but also to continuous-time settings where U will then be a function of the entire consumption process $c = (c_t)_{t\in[0,T]}$.

Often **time-additivity** is assumed so that the utility the individual gets from consumption in one period does not directly depend on what she consumed in earlier periods or what she plan to consume in later periods. For the discrete-time case, this means that

$$U(c_0, c_1, \ldots, c_T) = \sum_{t=0}^{T} u_t(c_t)$$

where each u_t is a valid "single-date" utility function. Still, when the individual has to choose her current consumption rate, she will take her prospects for future consumption into account. The continuous-time analogue is

$$U((c_t)_{t\in[0,T]}) = \int_0^T u_t(c_t)\, dt.$$

In addition it is typically assumed that $u_t(c_t) = e^{-\delta t}u(c_t)$ for all t. This is to say that the direct utility the individual gets from a given consumption level is basically the same for all dates, but the individual prefers to consume any given number of goods sooner than later. This is modeled by the subjective time preference rate δ, which we assume to be constant over time and independent of the consumption level. More impatient individuals have higher δ's. In sum, the life-time utility is typically assumed to be given by

$$U(c_0, c_1, \ldots, c_T) = \sum_{t=0}^{T} e^{-\delta t}u(c_t)$$

in discrete-time models and

$$U((c_t)_{t\in[0,T]}) = \int_0^T e^{-\delta t}u(c_t)\, dt$$

in continuous-time models. In both cases, u is a "single-date" utility function such as those discussed in Section 5.6.

Time-additivity is mostly assumed for tractability. However, it is important to realize that the time-additive specification does not follow from the basic axioms of choice under uncertainty, but is in fact a strong assumption, which most economists agree is not very realistic. One problem is that time-additive preferences induce a close link between the reluctance to substitute consumption across different states of the economy (which is measured by risk aversion) and the

willingness to substitute consumption over time (which can be measured by the so-called elasticity of intertemporal substitution). Solving intertemporal utility maximization problems of individuals with time-additive CRRA utility, it turns out that an individual with a high relative risk aversion will also choose a very smooth consumption process, i.e. she will have a low elasticity of intertemporal substitution. There is nothing in the basic theory of choice that links the risk aversion and the elasticity of intertemporal substitution together. For one thing, risk aversion makes sense even in an atemporal (i.e. one-date) setting where intertemporal substitution is meaningless and, conversely, intertemporal substitution makes sense in a multi-period setting without uncertainty in which risk aversion is meaningless. The close link between the two concepts in the multi-period model with uncertainty is an unfortunate consequence of the assumption of time-additive expected utility.

According to Browning (1991), non-additive preferences were already discussed in the 1890 book "Principles of Economics" by Alfred Marshall. See Browning's paper for further references to the critique on intertemporally separable preferences. Let us consider some alternatives that are more general and still tractable.

The key idea of **habit formation** is to let the utility associated with the choice of consumption at a given date depend on past choices of consumption. In a discrete-time setting the utility index of a given consumption process c is now given as $\mathrm{E}[\sum_{t=0}^{T} e^{-\delta t} u(c_t, h_t)]$, where h_t is a measure of the standard of living or the habit level of consumption, e.g. a weighted average of past consumption rates such as

$$h_t = h_0 e^{-\beta t} + \alpha \sum_{s=1}^{t-1} e^{-\beta(t-s)} c_s,$$

where h_0, α, and β are non-negative constants. It is assumed that u is decreasing in h so that high past consumption generates a desire for high current consumption, i.e. preferences display intertemporal complementarity. In particular, models where $u(c, h)$ is assumed to be of the power-linear form,

$$u(c, h) = \frac{1}{1-\gamma}(c-h)^{1-\gamma}, \quad \gamma > 0, c \geq h,$$

turn out to be computationally tractable. This is closely related to the subsistence HARA utility, but with habit formation the "subsistence level" h is endogenously determined by past consumption. The corresponding absolute and relative risk aversions are

$$\mathrm{ARA}(c, h) \equiv -\frac{u_{cc}(c, h)}{u_c(c, h)} = \frac{\gamma}{c-h}, \quad \mathrm{RRA}(c, h) \equiv -c\frac{u_{cc}(c, h)}{u_c(c, h)} = \frac{\gamma c}{c-h},$$

where u_c and u_{cc} are the first- and second-order derivatives of u with respect to c. In particular, the relative risk aversion is decreasing in c. Note that the habit formation preferences are still consistent with expected utility.

A related line of extension of the basic preferences is to allow the preferences of an individual to depend on some external factors, i.e. factors that are not fully determined by choices made by the individual. One example that has received some attention is where the utility which some individual attaches to her consumption plan depends on the consumption plans of other individuals or maybe the aggregate consumption in the economy. This is often referred to as "keeping up with the Jones'es." If you see your neighbors consume at high rates, you want to consume at a high rate too. Utility is state-dependent. Models of this type are sometimes said to have an

external habit, whereas the habit formation discussed above is then referred to as *internal* habit. If we denote the external factor by X_t, a time-additive life-time expected utility representation is $\mathrm{E}[\sum_{t=0}^{T} e^{-\delta t} u(c_t, X_t)]$, and a tractable version is $u(c, X) = \frac{1}{1-\gamma} (c - X)^{1-\gamma}$ very similar to the subsistence CRRA or the specific habit formation utility given above. In this case, however, "subsistence" level is determined by external factors. Another tractable specification is $u(c, X) = \frac{1}{1-\gamma}(c/X)^{1-\gamma}$.

Another preference specification gaining popularity is the so-called **recursive preferences** or Epstein-Zin preferences, suggested and discussed by, e.g., Kreps and Porteus (1978), Epstein and Zin (1989, 1991), and Weil (1989). The original motivation of this representation of preferences is that it allows individuals to have preferences for the timing of resolution of uncertainty, which is not consistent with the standard multi-date expected utility theory and violates the set of behavioral axioms. With recursive preferences the utility index (in this literature sometimes called the "felicity") $\mathcal{U}_t$ at some point in time t (capturing the preferences for consumption at time t and all later dates) depends both on consumption and that date and expectations of next period's utility index $\mathcal{U}_{t+1}$. The most tractable, non-trivial specification is usually written as

$$\mathcal{U}_t = \left[\left(1 - e^{-\delta}\right) c_t^{(1-\gamma)/\theta} + e^{-\delta} \left(\mathrm{E}_t \left[\mathcal{U}_{t+1}^{1-\gamma} \right] \right)^{1/\theta} \right]^{\theta/(1-\gamma)}, \quad \theta \equiv \frac{1-\gamma}{1-\frac{1}{\psi}}. \tag{5.12}$$

Here γ has the interpretation of the relative risk aversion, as before, and $\psi = \theta/(\theta + \gamma - 1)$ has the interpretation of the intertemporal elasticity of substitution. It is often easier to work with the "normalized" utility index $\tilde{\mathcal{U}}_t = \frac{1}{1-\gamma} \mathcal{U}_t^{1-\gamma}$, which represents the same preferences and satisfies

$$\tilde{\mathcal{U}}_t = \frac{1}{1-\gamma} \left[\left(1 - e^{-\delta}\right) c_t^{(1-\gamma)/\theta} + e^{-\delta} \left((1-\gamma) \, \mathrm{E}_t \left[\tilde{\mathcal{U}}_{t+1} \right] \right)^{1/\theta} \right]^{\theta}. \tag{5.13}$$

When $\gamma = 1/\psi$, we have $\theta = 1$, and therefore

$$\tilde{\mathcal{U}}_t = \frac{1}{1-\gamma} \left[\left(1 - e^{-\delta}\right) c_t^{1-\gamma} + e^{-\delta} (1-\gamma) \, \mathrm{E}_t \left[\tilde{\mathcal{U}}_{t+1} \right] \right].$$

There is a similar expression for $\tilde{\mathcal{U}}_{t+1}$ in terms of c_{t+1} and $\mathrm{E}_{t+1}\left[\tilde{\mathcal{U}}_{t+2}\right]$, which we can substitute into the above equation. If we keep doing that and assume that $\tilde{\mathcal{U}}_{T+1} = 0$, we obtain

$$\tilde{\mathcal{U}}_t = \left(1 - e^{-\delta}\right) \mathrm{E}_t \left[\sum_{s=t}^{T} e^{-\delta(s-t)} \frac{1}{1-\gamma} c_{t+s}^{1-\gamma} \right]. \tag{5.14}$$

Since the positive constant in front of the expectation does not affect the ordering of alternatives, we see that time-additive power utility is the special case of recursive preferences where $\gamma = 1/\psi$, i.e. the relative risk aversion equals the inverse of the elasticity of intertemporal substitution. The continuous-time equivalent of recursive utility is called stochastic differential utility and studied by, e.g., Duffie and Epstein (1992b). Note that generally recursive preferences are not consistent with expected utility since $\mathcal{U}_t$ depends non-linearly on the probabilities of future consumption levels.

For studying some problems it is useful or even necessary to distinguish between different consumption goods. Until now we have implicitly assumed a single consumption good which is perishable in the sense that it cannot be stored. However, individuals spend large amounts on durable goods such as houses and cars. These goods provide utility to the individual beyond the period of purchase and can potentially be resold at a later date so that it also acts as an investment.

Another important good is leisure. Individuals have preferences both for consumption of physical goods and for leisure. A tractable two-good utility function is the Cobb-Douglas function:

$$u(c_1, c_2) = \frac{1}{1-\gamma}\left(c_1^{\psi} c_2^{1-\psi}\right)^{1-\gamma},$$

where $\psi \in [0, 1]$ determines the relative weighting of the two goods.

5.8 Exercises

EXERCISE 5.1 Give a proof of Theorem 5.3.

EXERCISE 5.2 (Adapted from Problem 3.3 in Kreps (1990).) Consider the following two probability distributions of consumption. π_1 gives 5, 15, and 30 (dollars) with probabilities 1/3, 5/9, and 1/9, respectively. π_2 gives 10 and 20 with probabilities 2/3 and 1/3, respectively.

(a) Show that we can think of π_1 as a two-step gamble, where the first gamble is identical to π_2. If the outcome of the first gamble is 10, then the second gamble gives you an additional 5 (total 15) with probability 1/2 and an additional -5 (total 5) also with probability 1/2. If the outcome of the first gamble is 20, then the second gamble gives you an additional 10 (total 30) with probability 1/3 and an additional -5 (total 15) with probability 2/3.

(b) Observe that the second gamble has mean zero and that π_1 is equal to π_2 plus mean-zero noise. Conclude that any risk-averse expected utility maximizer will prefer π_2 to π_1.

EXERCISE 5.3 (Adapted from Chapter 3 in Kreps (1990).) Imagine a greedy, risk-averse, expected utility maximizing consumer whose end-of-period income level is subject to some uncertainty. The income will be Y with probability $\bar{p}$ and $Y' < Y$ with probability $1 - \bar{p}$. Think of $\Delta = Y - Y'$ as some loss the consumer might incur due an accident. An insurance company is willing to insure against this loss by paying Δ to the consumer if she sustains the loss. In return, the company wants an upfront premium of δ. The consumer may choose partial coverage in the sense that if she pays a premium of $a\delta$, she will receive $a\Delta$ if she sustains the loss. Let u denote the von Neumann-Morgenstern utility function of the consumer. Assume for simplicity that the premium is paid at the end of the period.

(a) Show that the first order condition for the choice of a is

$$\bar{p}\delta u'(Y - a\delta) = (1 - \bar{p})(\Delta - \delta)u'(Y - (1 - a)\Delta - a\delta).$$

(b) Show that if the insurance is actuarially fair in the sense that the expected payout $(1 - \bar{p})\Delta$ equals the premium δ, then the consumer will purchase full insurance, i.e. $a = 1$ is optimal.

(c) Show that if the insurance is actuarially unfair, meaning $(1 - \bar{p})\Delta < \delta$, then the consumer will purchase partial insurance, i.e. the optimal a is less than 1.

EXERCISE 5.4 Consider a one-period choice problem with four equally likely states of the world at the end of the period. The consumer maximizes expected utility of end-of-period wealth. The current wealth must be invested in a single financial asset today. The consumer has three assets to choose from. All three assets have a current price equal to the current wealth of the consumer. The assets have the following end-of-period values:

state	1	2	3	4
probability	0.25	0.25	0.25	0.25
asset 1	100	100	100	100
asset 2	81	100	100	144
asset 3	36	100	100	225

(a) What asset would a risk-neutral individual choose?

(b) What asset would a power utility investor, $u(W) = \frac{1}{1-\gamma}W^{1-\gamma}$ choose if $\gamma = 0.5$? If $\gamma = 2$? If $\gamma = 5$?

Now assume a power utility with $\gamma = 0.5$.

(c) Suppose the individual could obtain a perfect signal about the future state before she makes her asset choice. There are thus four possible signals, which we can represent by $s_1 = \{1\}$, $s_2 = \{2\}$, $s_3 = \{3\}$, and $s_4 = \{4\}$. What is the optimal asset choice for each signal? What is her expected utility before she receives the signal, assuming that the signals have equal probability?

(d) Now suppose that the individual can receive a less-than-perfect signal telling her whether the state is in $s_1 = \{1, 4\}$ or in $s_2 = \{2, 3\}$. The two possible signals are equally likely. What is the expected utility of the investor before she receives the signal?

EXERCISE 5.5 Consider an individual with log utility, $u(c) = \ln c$. What is her certainty equivalent and risk premium for the consumption plan which with probability 0.5 gives her $(1-\alpha)\bar{c}$ and with probability 0.5 gives her $(1+\alpha)\bar{c}$? Confirm that your results are consistent with numbers for $\gamma = 1$ shown in Table 5.5.

EXERCISE 5.6 Use Equation (5.5) to compute approximate relative risk premia for the consumption gamble underlying Table 5.5 and compare with the exact numbers given in the table.

Chapter 6

Individual optimality

6.1 Introduction

Chapter 4 discussed how the general pricing mechanism of a financial market can be represented by a state-price deflator. Given a state-price deflator we can price all state-contingent dividends. Conversely, given the market prices of state-contingent dividends we can extract one or several state-price deflators. Market prices and hence the state-price deflator(s) are determined by the supply and demand of the individuals in the economy. Therefore, we have to study the portfolio decisions of individuals. This is the topic of the present chapter. In the next chapter we will then discuss market equilibrium.

Section 6.2 studies the individual's maximization problem with various preference specifications in the one-period setting. Sections 6.3 and 6.4 extend the analysis to the discrete-time and the continuous-time framework, respectively. The main result of these sections is that the (marginal utility of) optimal consumption of any individual induces a valid state-price deflator, which gives us a link between individual optimality and asset prices. This is the cornerstone of the consumption-based asset pricing models studied in Chapter 8. Section 6.5 introduces the dynamic programming approach to the solution of multi-period utility maximization problems. In particular, we derive the so-called envelope condition that links marginal utility of consumption to marginal utility of optimal investments. In this way the state-price deflator is related to the optimally invested wealth of the individual plus some state variables capturing other information affecting the decisions of the individual. This will be useful for the factor pricing models studied in Chapter 9.

6.2 The one-period framework

In the one-period framework the individual consumes at time 0 (the beginning of the period) and at time 1 (the end of the period). We denote time 0 consumption by c_0 and the state-dependent time 1 consumption by the random variable c. The individual has some initial wealth $e_0 \geq 0$ at time 0 and may receive a non-negative state-dependent endowment (income) at time 1 represented by the random variable e. The individual picks a portfolio $\boldsymbol{\theta}$ at time 0 with a time 0 price of $P^{\boldsymbol{\theta}} = \boldsymbol{\theta}^\top \boldsymbol{P} = \sum_{i=1}^{I} \theta_i P_i$, assuming the Law of One Price, and a time 1 random dividend of

$D^{\boldsymbol{\theta}} = \boldsymbol{\theta}^\top \boldsymbol{D} = \sum_{i=1}^{I} \theta_i D_i$. The budget constraints are therefore

$$c_\omega \leq e_\omega + D^{\boldsymbol{\theta}}_\omega = e_\omega + \sum_{i=1}^{I} D_{i\omega}\theta_i, \quad \text{for all } \omega \in \Omega, \tag{6.1}$$

$$c_0 \leq e_0 - P^{\boldsymbol{\theta}} = e_0 - \sum_{i=1}^{I} \theta_i P_i. \tag{6.2}$$

The individual can choose the consumption plan and the portfolio. Since we will always assume that individuals prefer more consumption to less, it is clear that the budget constraints will hold as equalities. Therefore we can think of the individual choosing only the portfolio and then the consumption plan follows from the budget constraints above.

Consumption has to be non-negative both at time 0 and in all states at time 1 so we should add such constraints when looking for the optimal strategy. However, we assume throughout that the individual has infinite marginal utility at zero consumption so that the non-negativity constraints are automatically satisfied and can be ignored. We assume that the preferences are concave so that first-order conditions provide the optimal choice. When solving the problem we will also assume that the individual acts as a price taker so that prices are unaffected by her portfolio choice. We assume that prices admit no arbitrage. If there was an arbitrage, it would be possible to obtain infinite consumption. We do not impose any constraints on the portfolios the individual may choose among.

The following subsections characterize the optimal consumption plan for various preference specifications.

6.2.1 Time-additive expected utility

With time-additive expected utility there is a "single-date" utility function $u : \mathbb{R}_+ \to \mathbb{R}$ such that the objective of the individual is

$$\max_{\boldsymbol{\theta}} u(c_0) + \mathrm{E}\left[e^{-\delta} u(c)\right], \tag{6.3}$$

where δ is a subjective time preference rate. Substituting in the budget constraints, we get

$$\max_{\boldsymbol{\theta}} u\left(e_0 - \sum_{i=1}^{I} \theta_i P_i\right) + \mathrm{E}\left[e^{-\delta} u\left(e + \sum_{i=1}^{I} \theta_i D_i\right)\right].$$

The first-order condition with respect to θ_i is

$$-P_i u'\left(e_0 - \sum_{i=1}^{I} \theta_i P_i\right) + \mathrm{E}\left[e^{-\delta} D_i u'\left(e + \sum_{i=1}^{I} \theta_i D_i\right)\right] = 0$$

or

$$P_i u'(c_0) = \mathrm{E}\left[e^{-\delta} D_i u'(c)\right], \tag{6.4}$$

where c_0 and c denotes the optimal consumption plan, i.e. the consumption plan generated by the optimal portfolio $\boldsymbol{\theta}$. We can rewrite the above equation as

$$P_i = \mathrm{E}\left[e^{-\delta} \frac{u'(c)}{u'(c_0)} D_i\right]. \tag{6.5}$$

This equation links prices to the optimal consumption plan of an individual investor.

Comparing (6.5) and the pricing condition in the definition of a state-price deflator it is clear that

$$\zeta = e^{-\delta} \frac{u'(c)}{u'(c_0)} \tag{6.6}$$

defines a state-price deflator. It is positive since marginal utilities are positive. This state-price deflator is the marginal rate of substitution of the individual capturing the willingness of the individual to substitute a bit of time 0 consumption for some time 1 consumption.

The optimality condition (6.5) can also be justified by a variational argument, which goes as follows. Assume that (c_0, c) denotes the optimal consumption plan for the individual. Then any deviation from this plan will give the individual a lower utility. One deviation is obtained by investing in $\varepsilon > 0$ additional units of asset i at the beginning of the period. This leaves an initial consumption of $c_0 - \varepsilon P_i$. On the other hand, the end-of-period consumption in state ω becomes $c_\omega + \varepsilon D_{i\omega}$. We know that

$$u(c_0 - \varepsilon P_i) + e^{-\delta} \operatorname{E}\left[u(c + \varepsilon D_i)\right] \leq u(c_0) + e^{-\delta} \operatorname{E}\left[u(c)\right].$$

Subtracting the right-hand side from the left-hand side and dividing by ε yields

$$\frac{u(c_0 - \varepsilon P_i) - u(c_0)}{\varepsilon} + e^{-\delta} \operatorname{E}\left[\frac{u(c + \varepsilon D_i) - u(c)}{\varepsilon}\right] \leq 0.$$

Letting ε go to zero, the fractions on the left-hand side approaches derivatives and we obtain

$$-P_i u'(c_0) + e^{-\delta} \operatorname{E}\left[u'(c) D_i\right] \leq 0,$$

which implies that

$$P_i \geq \operatorname{E}\left[e^{-\delta} \frac{u'(c)}{u'(c_0)} D_i\right].$$

On the other hand, if we consider selling $\varepsilon > 0$ units of asset i at the beginning of the period, the same reasoning can be used to show that

$$P_i \leq \operatorname{E}\left[e^{-\delta} \frac{u'(c)}{u'(c_0)} D_i\right].$$

Hence, the relation must hold as an equality, just as in (6.5).

Example 6.1 For the case of CRRA utility, $u(c) = \frac{1}{1-\gamma} c^{1-\gamma}$, we have $u'(c) = c^{-\gamma}$. Therefore the optimal consumption plan satisfies

$$P_i = \operatorname{E}\left[e^{-\delta} \left(\frac{c}{c_0}\right)^{-\gamma} D_i\right], \tag{6.7}$$

and the state-price deflator derived from the individual's problem is

$$\zeta = e^{-\delta} \left(\frac{c}{c_0}\right)^{-\gamma}. \tag{6.8}$$

□

6.2.2 Non-additive expected utility

Next consider non-additive expected utility where the objective is to maximize $\mathrm{E}[U(c_0, c)]$ for some $U : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$. Again, substituting in the budget constraint, the problem is

$$\max_{\boldsymbol{\theta}} \mathrm{E}\left[U\left(e_0 - \sum_{i=1}^{I} \theta_i P_i, e + \sum_{i=1}^{I} \theta_i D_i\right)\right]. \tag{6.9}$$

The first-order condition with respect to θ_i is

$$-P_i \,\mathrm{E}\left[\frac{\partial U}{\partial c_0}\left(e_0 - \sum_{i=1}^{I} \theta_i P_i, e + \sum_{i=1}^{I} \theta_i D_i\right)\right] + \mathrm{E}\left[D_i \frac{\partial U}{\partial c}\left(e_0 - \sum_{i=1}^{I} \theta_i P_i, e + \sum_{i=1}^{I} \theta_i D_i\right)\right] = 0$$

which implies that

$$P_i = \mathrm{E}\left[\frac{\frac{\partial U}{\partial c}(c_0, c)}{\mathrm{E}\left[\frac{\partial U}{\partial c_0}(c_0, c)\right]} D_i\right], \tag{6.10}$$

so that the corresponding state-price deflator is

$$\zeta = \frac{\frac{\partial U}{\partial c}(c_0, c)}{\mathrm{E}\left[\frac{\partial U}{\partial c_0}(c_0, c)\right]}. \tag{6.11}$$

This could be supported by a variational argument as in the case of time-additive expected utility. Again note that these equations hold for the optimal consumption plan.

Example 6.2 Consider the very simple habit-style utility function

$$U(c_0, c) = \frac{1}{1-\gamma} c_0^{1-\gamma} + \frac{1}{1-\gamma} e^{-\delta} \left(c - \alpha c_0\right)^{1-\gamma}.$$

The subsistence level for time 1 consumption is some fraction α of time 0 consumption. In this case the relevant marginal utilities are

$$\begin{aligned} \frac{\partial U}{\partial c_0}(c_0, c) &= c_0^{-\gamma} - \beta e^{-\delta} \left(c - \beta c_0\right)^{-\gamma}, \\ \frac{\partial U}{\partial c}(c_0, c) &= e^{-\delta} \left(c - \beta c_0\right)^{-\gamma}. \end{aligned}$$

The first-order condition therefore implies that

$$P_i = \mathrm{E}\left[\frac{e^{-\delta} \left(c - \beta c_0\right)^{-\gamma}}{\mathrm{E}\left[c_0^{-\gamma} - \beta e^{-\delta} \left(c - \beta c_0\right)^{-\gamma}\right]} D_i\right], \tag{6.12}$$

and the associated state-price deflator is

$$\zeta = \frac{e^{-\delta} \left(c - \beta c_0\right)^{-\gamma}}{\mathrm{E}\left[c_0^{-\gamma} - \beta e^{-\delta} \left(c - \beta c_0\right)^{-\gamma}\right]}. \tag{6.13}$$

This simple example indicates that (internal) habit formation leads to pricing expressions that are considerably more complicated than time-additive utility. □

6.2.3 A general utility index

A general utility index $\mathcal{U}$ is tractable for a finite state space where we can write the objective as

$$\max_{\boldsymbol{\theta}} \mathcal{U}(c_0, c_1, \ldots, c_S),$$

where c_ω is consumption in state ω, $\omega = 1, \ldots, S$. From the budget constraint $c_\omega = e_\omega + \sum_{i=1}^{I} D_{i\omega}\theta_i$ so the first-order condition implies that

$$P_i = \sum_{\omega=1}^{S} \frac{\frac{\partial \mathcal{U}}{\partial c_\omega}(c_0, c_1, \ldots, c_S)}{\frac{\partial \mathcal{U}}{\partial c_0}(c_0, c_1, \ldots, c_S)} D_{i\omega}. \tag{6.14}$$

This defines a state-price vector $\boldsymbol{\psi}$ by

$$\psi_\omega = \frac{\frac{\partial \mathcal{U}}{\partial c_\omega}(c_0, c_1, \ldots, c_S)}{\frac{\partial \mathcal{U}}{\partial c_0}(c_0, c_1, \ldots, c_S)}.$$

6.2.4 A two-step procedure in a complete market

In a complete market we can separate the consumption and the portfolio decision as follows:

1. find the optimal consumption plan given the budget constraints,
2. find the portfolio financing the optimal consumption plan; such a portfolio will exist when the market is complete.

Let us show this in the case of a finite state space. Suppose the market is complete and let $\boldsymbol{\psi}$ denote the unique state-price vector. The individual can obtain any dividend vector $\boldsymbol{D}$ at the cost of $\boldsymbol{\psi} \cdot \boldsymbol{D}$. Hence the individual can first solve the problem

$$\max_{c_0, \boldsymbol{c}, \boldsymbol{D}} \ \mathcal{U}(c_0, \boldsymbol{c}) \tag{6.15}$$

$$\text{s.t. } \boldsymbol{c} \leq \boldsymbol{e} + \boldsymbol{D}, \tag{6.16}$$

$$c_0 \leq e_0 - \boldsymbol{\psi} \cdot \boldsymbol{D}, \tag{6.17}$$

$$c_0, \boldsymbol{c} \geq 0,$$

for the optimal consumption plan. Here (6.16) is a vector inequality, which means that the inequality should hold element by element, i.e.

$$c_\omega \leq e_\omega + D_\omega, \quad \omega = 1, \ldots, S.$$

In fact we can eliminate the dividends from the problem. Multiplying (6.16) by $\boldsymbol{\psi}$, we get

$$\boldsymbol{\psi} \cdot \boldsymbol{c} \leq \boldsymbol{\psi} \cdot \boldsymbol{e} + \boldsymbol{\psi} \cdot \boldsymbol{D}.$$

Adding this to (6.17), we see that any feasible consumption plan $(c_0, \boldsymbol{c})$ must satisfy

$$c_0 + \boldsymbol{\psi} \cdot \boldsymbol{c} \leq e_0 + \boldsymbol{\psi} \cdot \boldsymbol{e}. \tag{6.18}$$

This is natural, since the left-hand side is the present value of the consumption plan and the right-hand side is the present value of the endowment, which is well-defined since market completeness

ensures that some portfolio will provide a dividend identical to the endowment. Conversely, suppose a consumption plan $(c_0, \boldsymbol{c})$ satisfies (6.18). Then it will also satisfy the conditions (6.16) and (6.17) with $\boldsymbol{D} = \boldsymbol{c} - \boldsymbol{e}$. Thus we can find the utility maximizing consumption plan by solving

$$\begin{aligned} \max_{c_0, \boldsymbol{c}} \ & \mathcal{U}(c_0, \boldsymbol{c}) \\ \text{s.t.} \ & c_0 + \boldsymbol{\psi} \cdot \boldsymbol{c} \leq e_0 + \boldsymbol{\psi} \cdot \boldsymbol{e} \\ & c_0, \boldsymbol{c} \geq 0, \end{aligned} \tag{6.19}$$

and we will still assume that the non-negativity constraint will be automatically satisfied. The Lagrangian for the problem is therefore

$$\mathcal{L} = \mathcal{U}(c_0, \boldsymbol{c}) + \nu \left(e_0 + \boldsymbol{\psi} \cdot \boldsymbol{e} - c_0 - \boldsymbol{\psi} \cdot \boldsymbol{c} \right),$$

where ν is the Lagrange multiplier. The first-order conditions are

$$\frac{\partial \mathcal{U}}{\partial c_0}(c_0, \boldsymbol{c}) = \nu, \quad \frac{\partial \mathcal{U}}{\partial \boldsymbol{c}}(c_0, \boldsymbol{c}) = \nu \boldsymbol{\psi}.$$

In particular, the optimal consumption plan satisfies

$$\frac{\frac{\partial \mathcal{U}}{\partial \boldsymbol{c}}(c_0, \boldsymbol{c})}{\frac{\partial \mathcal{U}}{\partial c_0}(c_0, \boldsymbol{c})} = \boldsymbol{\psi}.$$

Given the chosen consumption plan $\boldsymbol{c}$ and the future income $\boldsymbol{e}$, we can back out the portfolio by solving $\underline{\underline{D}}^\top \boldsymbol{\theta} = \boldsymbol{e} - \boldsymbol{c}$ for $\boldsymbol{\theta}$. In a complete market, such a portfolio can always be found.

With an infinite state space and an expected utility representation $\mathcal{U}(c_0, c) = \mathrm{E}[U(c_0, c)]$, we can formulate the complete markets problem as

$$\begin{aligned} \max_{c_0, c} \ & \mathrm{E}[U(c_0, c)] \\ \text{s.t.} \ & c_0 + \mathrm{E}[\zeta c] \leq e_0 + \mathrm{E}[\zeta e] \\ & c_0, c \geq 0, \end{aligned} \tag{6.20}$$

where ζ is the unique state-price deflator. The Lagrangian is

$$\mathcal{L} = \mathrm{E}[U(c_0, c)] + \nu \left(e_0 - c_0 + \mathrm{E}[\zeta(e - c)] \right) = \nu \left(e_0 - c_0 \right) + \mathrm{E}\left[U(c_0, c) + \nu \zeta (e - c) \right].$$

We maximize the expectation $\mathrm{E}\left[U(c_0, c) + \nu \zeta (e - c) \right]$ by maximizing state-by-state, i.e. maximizing $U(c_0, c(\omega)) + \nu \zeta(\omega) \left(e(\omega) - c(\omega) \right)$ for each state ω. The first-order conditions with respect to $c(\omega)$ implies

$$\frac{\partial U}{\partial c}(c_0, c(\omega)) = \nu \zeta(\omega)$$

and the first-order condition with respect to c_0 implies that

$$\mathrm{E}\left[\frac{\partial U}{\partial c_0}(c_0, c) \right] = \nu$$

and, hence,

$$\frac{\frac{\partial U}{\partial c}(c_0, c)}{\mathrm{E}\left[\frac{\partial U}{\partial c_0}(c_0, c) \right]} = \zeta.$$

In particular, with time-additive expected utility $U(c_0, c) = u(c_0) + e^{-\delta}u(c)$, we get

$$e^{-\delta}\frac{u'(c)}{u'(c_0)} = \zeta$$

as found earlier.

If the market is incomplete, the individual cannot implement any consumption plan but only those that can be financed by portfolios of traded assets. Therefore, we cannot apply the above technique.

6.2.5 Optimal portfolios and mean-variance analysis

We have not explicitly solved for the optimal portfolio. Although it is certainly relevant to study the portfolio decisions of individuals in more detail, it is not necessary for asset pricing purposes. The most popular one-period model for portfolio choice is the mean-variance model introduced by Markowitz (1952, 1959). If the individual has mean-variance preferences, cf. Section 5.6, her optimal portfolio will be a mean-variance efficient portfolio corresponding to a point on the upward-sloping branch of the mean-variance frontier. Mean-variance analysis does not by itself say anything about exactly which portfolio a given individual should choose but if we assume a given mean-variance utility function, the optimal portfolio can be derived. Note, however, that the conditions justifying mean-variance portfolio choice are highly unrealistic: either returns must be normally distributed or individuals must have mean-variance preferences. Nevertheless, the mean-variance frontier remains an important concept in both portfolio choice and asset pricing (recall Theorem 4.6). Below we will use the traditional Lagrangian approach to characterize the mean-variance efficient portfolios. In Section 9.4 we will offer an alternative "orthogonal" characterization and use that to show the link between mean-variance returns, pricing factors, and state-price deflators.

We assume as before that I assets are traded and let $\boldsymbol{R} = (R_1, \ldots, R_I)^\top$ denote the vector of gross returns on these assets. Let $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{R}]$ denote the vector of expected gross returns and let $\underline{\underline{\Sigma}} = \mathrm{Var}[\boldsymbol{R}]$ denote the $I \times I$ variance-covariance matrix of gross returns. In the characterization of mean-variance efficient portfolios, we are only interested in their returns so we can represent portfolios by portfolio weight vectors, i.e. vectors $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_I)^\top$ with $\boldsymbol{\pi} \cdot \mathbf{1} = 1$, where π_i is the fraction of total portfolio value invested in asset i. The gross return on a portfolio $\boldsymbol{\pi}$ is $R^{\boldsymbol{\pi}} = \boldsymbol{\pi} \cdot \boldsymbol{R} = \sum_{i=1}^{I} \pi_i R_i$, cf. (3.7). The expectation and the variance of the return on a portfolio $\boldsymbol{\pi}$ are

$$\begin{aligned}
\mathrm{E}\left[R^{\boldsymbol{\pi}}\right] &= \mathrm{E}\left[\sum_{i=1}^{I} \pi_i R_i\right] = \sum_{i=1}^{I} \pi_i \,\mathrm{E}[R_i] = \sum_{i=1}^{I} \pi_i \mu_i = \boldsymbol{\pi} \cdot \boldsymbol{\mu} = \boldsymbol{\pi}^\top \boldsymbol{\mu}, \\
\mathrm{Var}\left[R^{\boldsymbol{\pi}}\right] &= \mathrm{Var}\left[\sum_{i=1}^{I} \pi_i R_i\right] = \sum_{i=1}^{I}\sum_{j=1}^{I} \pi_i \pi_j \,\mathrm{Cov}[R_i, R_j] = \boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}.
\end{aligned}$$

A portfolio $\boldsymbol{\pi}$ is then mean-variance efficient if there is an $m \in \mathbb{R}$ so that $\boldsymbol{\pi}$ solves

$$\min_{\boldsymbol{\pi}} \boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi} \quad \text{s.t.} \quad \boldsymbol{\pi}^\top \boldsymbol{\mu} = m, \quad \boldsymbol{\pi}^\top \mathbf{1} = 1, \tag{6.21}$$

i.e. $\boldsymbol{\pi}$ has the lowest return variance among all portfolios with expected return m.

Risky assets only

Assume that $\underline{\underline{\Sigma}}$ is positive definite, i.e. that $\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi} > 0$ for all $\boldsymbol{\pi}$, which means that the variance of the return on any portfolio is positive. Any portfolio of risky assets will be risky; there is no risk-free asset and no redundant assets. It follows that $\underline{\underline{\Sigma}}$ is non-singular and that the inverse $\underline{\underline{\Sigma}}^{-1}$ is also positive definite. We will allow for a risk-free asset later.

The Lagrangian associated with the constrained minimization problem (6.21) is

$$\mathcal{L} = \boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi} + \alpha \left(m - \boldsymbol{\pi}^\top \boldsymbol{\mu}\right) + \beta \left(1 - \boldsymbol{\pi}^\top \mathbf{1}\right),$$

where α and β are Lagrange multipliers. The first-order condition with respect to $\boldsymbol{\pi}$ is

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\pi}} = 2\underline{\underline{\Sigma}} \boldsymbol{\pi} - \alpha \boldsymbol{\mu} - \beta \mathbf{1} = 0,$$

which implies that

$$\boldsymbol{\pi} = \frac{1}{2} \alpha \underline{\underline{\Sigma}}^{-1} \boldsymbol{\mu} + \frac{1}{2} \beta \underline{\underline{\Sigma}}^{-1} \mathbf{1}. \tag{6.22}$$

The first-order conditions with respect to the multipliers simply give the two constraints to the minimization problem. Substituting the expression (6.22) for $\boldsymbol{\pi}$ into the two constraints, we obtain the equations

$$\begin{aligned} \alpha \boldsymbol{\mu}^\top \underline{\underline{\Sigma}}^{-1} \boldsymbol{\mu} + \beta \mathbf{1}^\top \underline{\underline{\Sigma}}^{-1} \boldsymbol{\mu} &= 2m, \\ \alpha \boldsymbol{\mu}^\top \underline{\underline{\Sigma}}^{-1} \mathbf{1} + \beta \mathbf{1}^\top \underline{\underline{\Sigma}}^{-1} \mathbf{1} &= 2. \end{aligned}$$

Defining the numbers A, B, C, D by

$$A = \boldsymbol{\mu}^\top \underline{\underline{\Sigma}}^{-1} \boldsymbol{\mu}, \qquad B = \boldsymbol{\mu}^\top \underline{\underline{\Sigma}}^{-1} \mathbf{1} = \mathbf{1}^\top \underline{\underline{\Sigma}}^{-1} \boldsymbol{\mu}, \qquad C = \mathbf{1}^\top \underline{\underline{\Sigma}}^{-1} \mathbf{1}, \qquad D = AC - B^2,$$

we can write the solution to the two equations in α and β as

$$\alpha = 2\frac{Cm - B}{D}, \qquad \beta = 2\frac{A - Bm}{D}.$$

Substituting this into (6.22) we obtain

$$\boldsymbol{\pi} = \boldsymbol{\pi}(m) \equiv \frac{Cm - B}{D} \underline{\underline{\Sigma}}^{-1} \boldsymbol{\mu} + \frac{A - Bm}{D} \underline{\underline{\Sigma}}^{-1} \mathbf{1}. \tag{6.23}$$

This is the mean-variance efficient portfolio with expected gross return m. Some tedious calculations show that the variance of the return on this portfolio is equal to

$$\sigma^2(m) \equiv \boldsymbol{\pi}(m)^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}(m) = \frac{Cm^2 - 2Bm + A}{D}. \tag{6.24}$$

This is to be verified in Exercise 6.4. Equation (6.24) shows that the combinations of variance and mean form a parabola in a (mean, variance)-diagram.

Traditionally the portfolios are depicted in a (standard deviation, mean)-diagram. The above relation can also be written as

$$\frac{\sigma^2(m)}{1/C} - \frac{(m - B/C)^2}{D/C^2} = 1,$$

from which it follows that the optimal combinations of standard deviation and mean form a hyperbola in the (standard deviation, mean)-diagram. This hyperbola is called the **mean-variance**

frontier of risky assets. The mean-variance efficient portfolios are sometimes called frontier portfolios.

Before we proceed let us clarify a point in the derivation above. We need to divide by D so D has to be non-zero. In fact, $D > 0$. To see this, first note that since $\underline{\underline{\Sigma}}$ and therefore $\underline{\underline{\Sigma}}^{-1}$ are positive definite, we have $A > 0$ and $C > 0$. Moreover,

$$AD = A(AC - B^2) = (B\boldsymbol{\mu} - A\mathbf{1})^\top \underline{\underline{\Sigma}}^{-1}(B\boldsymbol{\mu} - A\mathbf{1}) > 0,$$

again using that $\underline{\underline{\Sigma}}^{-1}$ is positive definite. Since $A > 0$, we must have $D > 0$.

The (global) **minimum-variance portfolio** is the portfolio that has the minimum variance among all portfolios. We can find this directly by solving the constrained minimization problem

$$\min_{\boldsymbol{\pi}} \boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi} \quad \text{s.t.} \quad \boldsymbol{\pi}^\top \mathbf{1} = 1 \tag{6.25}$$

where there is no constraint on the expected portfolio return. Alternatively, we can minimize the variance $\sigma^2(m)$ in (6.24) over all m. Taking the latter route, we find that the minimum variance is obtained when the mean return is $m_{\min} = B/C$ and the minimum variance is given by $\sigma^2_{\min} = \sigma^2(m_{\min}) = 1/C$. From (6.23) we get that the minimum-variance portfolio is

$$\boldsymbol{\pi}_{\min} = \frac{1}{C}\underline{\underline{\Sigma}}^{-1}\mathbf{1} = \frac{1}{\mathbf{1}^\top \underline{\underline{\Sigma}}^{-1}\mathbf{1}}\underline{\underline{\Sigma}}^{-1}\mathbf{1}. \tag{6.26}$$

It will be useful to consider the problem

$$\max_{\boldsymbol{\pi}} \frac{\boldsymbol{\pi}^\top \boldsymbol{\mu} - \alpha}{\left(\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}\right)^{1/2}} \quad \text{s.t.} \quad \boldsymbol{\pi}^\top \mathbf{1} = 1, \tag{6.27}$$

where α is some constant. The denominator in the objective is clearly the standard deviation of the return of the portfolio, while the numerator is the expected return in excess of α. This ratio is the slope of a straight line in the (standard deviation, mean)-diagram that goes through the point $\left((\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi})^{1/2}, \boldsymbol{\pi}^\top \boldsymbol{\mu}\right)$ corresponding to the portfolio $\boldsymbol{\pi}$ and intersects the mean-axis at α. Applying the constraint, the objective function can be rewritten as

$$f(\boldsymbol{\pi}) = \frac{\boldsymbol{\pi}^\top(\boldsymbol{\mu} - \alpha\mathbf{1})}{\left(\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}\right)^{1/2}} = \boldsymbol{\pi}^\top(\boldsymbol{\mu} - \alpha\mathbf{1}) \left(\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}\right)^{-1/2}.$$

The derivative is

$$\frac{\partial f}{\partial \boldsymbol{\pi}} = (\boldsymbol{\mu} - \alpha\mathbf{1}) \left(\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}\right)^{-1/2} - \left(\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}\right)^{-3/2} \boldsymbol{\pi}^\top(\boldsymbol{\mu} - \alpha\mathbf{1})\underline{\underline{\Sigma}}\boldsymbol{\pi}$$

and $\frac{\partial f}{\partial \boldsymbol{\pi}} = \mathbf{0}$ implies that

$$\frac{\boldsymbol{\pi}^\top(\boldsymbol{\mu} - \alpha\mathbf{1})}{\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}}\boldsymbol{\pi} = \underline{\underline{\Sigma}}^{-1}(\boldsymbol{\mu} - \alpha\mathbf{1}), \tag{6.28}$$

which we want to solve for $\boldsymbol{\pi}$. Note that the equation has a vector on each side. If two vectors are identical, they will also be identical after a division by the sum of the elements of the vector. The sum of the elements of the vector on the left-hand side of (6.28) is

$$\mathbf{1}^\top \left(\frac{\boldsymbol{\pi}^\top(\boldsymbol{\mu} - \alpha\mathbf{1})}{\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}}\boldsymbol{\pi}\right) = \frac{\boldsymbol{\pi}^\top(\boldsymbol{\mu} - \alpha\mathbf{1})}{\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}}\mathbf{1}^\top\boldsymbol{\pi} = \frac{\boldsymbol{\pi}^\top(\boldsymbol{\mu} - \alpha\mathbf{1})}{\boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi}},$$

where the last equality is due to the constraint. The sum of the elements of the vector on the right-hand side of (6.28) is simply $\mathbf{1}^\top \underline{\underline{\Sigma}}^{-1}(\boldsymbol{\mu} - \alpha\mathbf{1})$. Dividing each side of (6.28) with the sum of the elements we obtain the portfolio

$$\boldsymbol{\pi} = \frac{\underline{\underline{\Sigma}}^{-1}(\boldsymbol{\mu} - \alpha\mathbf{1})}{\mathbf{1}^\top \underline{\underline{\Sigma}}^{-1}(\boldsymbol{\mu} - \alpha\mathbf{1})}. \tag{6.29}$$

In particular, by letting $\alpha = 0$, we can see that the portfolio

$$\boldsymbol{\pi}_{\text{slope}} = \frac{1}{\mathbf{1}^\top \underline{\underline{\Sigma}}^{-1}\boldsymbol{\mu}} \underline{\underline{\Sigma}}^{-1}\boldsymbol{\mu} = \frac{1}{B}\underline{\underline{\Sigma}}^{-1}\boldsymbol{\mu} \tag{6.30}$$

is the portfolio that maximizes the slope of a straight line between the origin and a point on the mean-variance frontier in the (standard deviation, mean)-diagram. Let us call $\boldsymbol{\pi}_{\text{slope}}$ the **maximum-slope portfolio**. This portfolio has mean A/B and variance A/B^2.

From (6.23) we see that any mean-variance efficient portfolio can be written as a linear combination of the maximum-slope portfolio and the minimum-variance portfolio:

$$\boldsymbol{\pi}(m) = \frac{(Cm - B)B}{D}\boldsymbol{\pi}_{\text{slope}} + \frac{(A - Bm)C}{D}\boldsymbol{\pi}_{\min}. \tag{6.31}$$

Note that the two multipliers of the portfolios sum to one. This is a **two-fund separation** result. Any mean-variance efficient portfolio is a combination of two special portfolios or funds, namely the maximum slope portfolio and the minimum-variance portfolio. These two portfolios are said to generate the mean-variance frontier of risky assets. In fact, it can be shown that any other two frontier portfolios generate the entire frontier.

The following result is both interesting and useful. Let $R^{\boldsymbol{\pi}}$ denote the return on any mean-variance efficient portfolio different from the minimum-variance portfolio. Then there exists a unique mean-variance efficient portfolio with a return $R^{z(\boldsymbol{\pi})}$ such that $\text{Cov}[R^{\boldsymbol{\pi}}, R^{z(\boldsymbol{\pi})}] = 0$. The return $R^{z(\boldsymbol{\pi})}$ is called the zero-beta return for $R^{\boldsymbol{\pi}}$, which is consistent with the definition of betas in the section on pricing factors. Furthermore, one can show that

$$\text{E}[R^{z(\boldsymbol{\pi})}] = \frac{A - B\,\text{E}[R^{\boldsymbol{\pi}}]}{B - C\,\text{E}[R^{\boldsymbol{\pi}}]}$$

and that the tangent to the mean-variance frontier at the point corresponding to $R^{\boldsymbol{\pi}}$ will intersect the expected return axis exactly in $\text{E}[R^{z(\boldsymbol{\pi})}]$. These results are to be shown in Exercise 6.5.

Allowing for a risk-free asset

Now let us allow for a risk-free asset with a gross return of R^f. The risk-free asset corresponds to the point $(0, R^f)$ in the (standard deviation, mean)-diagram. Either the risk-free asset is one of the I basic assets or it can be constructed as a portfolio of the basic assets. Without loss of generality we can assume that the risk-free asset is the I'th basic asset. The remaining $M \equiv I - 1$ basic assets are risky. Let $\tilde{\boldsymbol{R}} = (R_1, \ldots, R_M)^\top$ denote the gross return vector of the risky assets with expectation $\tilde{\boldsymbol{\mu}} = \text{E}[\tilde{\boldsymbol{R}}]$ and variance $\underline{\underline{\tilde{\Sigma}}} = \text{Var}[\tilde{\boldsymbol{R}}]$. Now assume that $\underline{\underline{\tilde{\Sigma}}}$ is positive definite. We assume that the risk-free return is smaller than the expected return on the minimum-variance portfolio of the risky assets.

A portfolio of all I assets can be represented by an M-dimensional vector $\tilde{\boldsymbol{\pi}}$ of the portfolio weights invested in the risky assets, while the remaining fraction $\pi_f \equiv 1 - \tilde{\boldsymbol{\pi}}^\top\mathbf{1}$ is the invested in

the risk-free asset. A portfolio involving only the risky assets is an M-dimensional vector $\tilde{\boldsymbol{\pi}}$ with $\tilde{\boldsymbol{\pi}}^\top \mathbf{1} = 1$.

Fix for a moment a portfolio $\tilde{\boldsymbol{\pi}}$ of risky assets only. The gross return on this portfolio is $R^{\tilde{\boldsymbol{\pi}}} = \tilde{\boldsymbol{\pi}}^\top \tilde{\boldsymbol{R}}$. Suppose you invest a fraction π_f of some amount in the risk-free asset and the remaining fraction $1 - \pi_f$ in this particular risky portfolio. The gross return on this combination will be

$$R = \pi_f R^f + (1 - \pi_f) R^{\tilde{\boldsymbol{\pi}}}$$

with mean and variance given by

$$\mathrm{E}[R] = \pi_f R^f + (1 - \pi_f) \mathrm{E}\left[R^{\tilde{\boldsymbol{\pi}}}\right], \qquad \mathrm{Var}[R] = (1 - \pi_f)^2 \mathrm{Var}\left[R^{\tilde{\boldsymbol{\pi}}}\right]$$

If $\pi_f \leq 1$, the standard deviation of the return is $\sigma[R] = (1 - \pi_f)\sigma\left[R^{\tilde{\boldsymbol{\pi}}}\right]$ and we obtain

$$\mathrm{E}[R] = \pi_f R^f + \frac{\mathrm{E}\left[R^{\tilde{\boldsymbol{\pi}}}\right]}{\sigma\left[R^{\tilde{\boldsymbol{\pi}}}\right]} \sigma[R]$$

so varying π_f the set of points $\{(\sigma[R], \mathrm{E}[R]) \mid \pi_f \leq 1\}$ will form an upward-sloping straight line from $(0, R^f)$ through $(\sigma\left[R^{\tilde{\boldsymbol{\pi}}}\right], \mathrm{E}\left[R^{\tilde{\boldsymbol{\pi}}}\right])$. For $\pi_f > 1$, the standard deviation of the combined portfolio is $\sigma[R] = -(1 - \pi_f)\sigma\left[R^{\tilde{\boldsymbol{\pi}}}\right]$ and we get

$$\mathrm{E}[R] = \pi_f R^f - \frac{\mathrm{E}\left[R^{\tilde{\boldsymbol{\pi}}}\right]}{\sigma\left[R^{\tilde{\boldsymbol{\pi}}}\right]} \sigma[R],$$

which defines a downward-sloping straight line from $(0, R^f)$ and to the right.

Minimizing variance for a given expected return we will move as far to the "north-west" or to the "south-west" as possible in the (standard deviation, mean)-diagram. Therefore the mean-variance efficient portfolios will correspond to points on a line which is tangent to the mean-variance frontier of risky assets and goes through the point $(0, R^f)$. There are two such lines, an upward-sloping and a downward-sloping line. The point where the upward-sloping line is tangent to the frontier of risky assets corresponds to a portfolio which we refer to as the **tangency portfolio**. This is a portfolio of risky assets only. It is the portfolio that maximizes the Sharpe ratio over all risky portfolios. The Sharpe ratio of a portfolio is the ratio $(\mathrm{E}[R^{\tilde{\boldsymbol{\pi}}}] - R^f)/\sigma[R^{\tilde{\boldsymbol{\pi}}}]$ between the excess expected return of a portfolio and the standard deviation of the return. To determine the tangency portfolio we must solve the problem

$$\max_{\tilde{\boldsymbol{\pi}}} \frac{\tilde{\boldsymbol{\pi}}^\top \tilde{\boldsymbol{\mu}} - R^f}{\left(\tilde{\boldsymbol{\pi}}^\top \tilde{\underline{\underline{\Sigma}}} \tilde{\boldsymbol{\pi}}\right)^{1/2}} \quad \text{s.t.} \quad \tilde{\boldsymbol{\pi}}^\top \mathbf{1} = 1. \tag{6.32}$$

Except for the tildes above the symbols, this problem is identical to the problem (6.27) with $\alpha = R^f$ which we have already solved. We can therefore conclude that the tangency portfolio is given by

$$\tilde{\boldsymbol{\pi}}_{\tan} = \frac{\tilde{\underline{\underline{\Sigma}}}^{-1}\left(\tilde{\boldsymbol{\mu}} - R^f \mathbf{1}\right)}{\mathbf{1}^\top \tilde{\underline{\underline{\Sigma}}}^{-1}\left(\tilde{\boldsymbol{\mu}} - R^f \mathbf{1}\right)}. \tag{6.33}$$

The gross return on the tangency portfolio is

$$R_{\tan} = \tilde{\boldsymbol{\pi}}_{\tan}^\top \tilde{\boldsymbol{R}}$$

with expectation and standard deviation given by

$$\mu_{\tan} = \tilde{\boldsymbol{\mu}}^\top \tilde{\boldsymbol{\pi}}_{\tan} = \frac{\tilde{\boldsymbol{\mu}}^\top \tilde{\underline{\underline{\Sigma}}}^{-1} \left(\tilde{\boldsymbol{\mu}} - R^f \mathbf{1}\right)}{\mathbf{1}^\top \tilde{\underline{\underline{\Sigma}}}^{-1} \left(\tilde{\boldsymbol{\mu}} - R^f \mathbf{1}\right)}, \tag{6.34}$$

$$\sigma_{\tan} = \left(\tilde{\boldsymbol{\pi}}_{\tan}^\top \tilde{\underline{\underline{\Sigma}}} \tilde{\boldsymbol{\pi}}_{\tan}\right)^{1/2} = \frac{\left((\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})\right)^{1/2}}{\mathbf{1}^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})}. \tag{6.35}$$

The maximum Sharpe ratio, i.e. the slope of the line, is thus

$$\begin{aligned}
\frac{\mu_{\tan} - R^f}{\sigma_{\tan}} &= \frac{\frac{\tilde{\boldsymbol{\mu}}^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})}{\mathbf{1}^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})} - R^f}{\frac{\left((\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})\right)^{1/2}}{\mathbf{1}^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})}} = \frac{\tilde{\boldsymbol{\mu}}^\top \tilde{\underline{\underline{\Sigma}}}^{-1} \left(\tilde{\boldsymbol{\mu}} - R^f \mathbf{1}\right) - R^f [\mathbf{1}^\top \tilde{\underline{\underline{\Sigma}}}^{-1} \left(\tilde{\boldsymbol{\mu}} - R^f \mathbf{1}\right)]}{\left((\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})\right)^{1/2}} \\
&= \frac{(\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})}{\left((\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})\right)^{1/2}} = \left((\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})^\top \tilde{\underline{\underline{\Sigma}}}^{-1} (\tilde{\boldsymbol{\mu}} - R^f \mathbf{1})\right)^{1/2}.
\end{aligned}$$

The straight line from the point $(0, R^f)$ and to and through $(\sigma_{\tan}, \mu_{\tan})$ constitutes the upward-sloping part of the mean-variance frontier of all assets. Similarly, there is a downward-sloping part which starts out at $(0, R^f)$ and has a slope which equals minus the slope of the upward-sloping frontier. Again we have two-fund separation since all investors will combine just two funds, where one fund is simply the risk-free asset and the other is the tangency portfolio of only risky assets. A return R is mean-variance efficient if and only if

$$R = \alpha R^f + (1 - \alpha) R_{\tan}$$

for some α. If $\alpha \leq 1$, you will get a point on the upward-sloping part of the frontier. If $\alpha \geq 1$, you will get a point on the downward-sloping part. Of course, when a risk-free return is traded, it will be the minimum-variance return. The relation between the mean m and the standard deviation σ of the portfolios on the efficient frontier will be

$$\sigma = |m - R^f| \, \frac{\sigma[R_{\tan}]}{\mathrm{E}[R_{\tan}] - R^f}. \tag{6.36}$$

6.3 The discrete-time framework

In the discrete-time framework each individual has to choose a consumption process $c = (c_t)_{t \in \mathcal{T}}$, where $\mathcal{T} = \{0, 1, \ldots, T\}$ and c_t denotes the random, i.e. state-dependent, consumption at time t. The individual also has to choose a trading strategy $\boldsymbol{\theta} = (\boldsymbol{\theta}_t)_{t=0,1,\ldots,T-1}$ with $\boldsymbol{\theta}_t$ representing the portfolio held from time t until time $t+1$. Again, $\boldsymbol{\theta}_t$ may depend on the information available to the individual at time t so $\boldsymbol{\theta}$ is an adapted stochastic process. The individual has an endowment or income process $e = (e_t)_{t \in \mathcal{T}}$, where e_0 is the initial endowment (wealth) and e_t is the possibly state-dependent income received at time t.

6.3.1 Time-additive expected utility

Let us first focus on individuals with time-additive expected utility. Standing at time 0 the problem of an individual investor can therefore be written as

$$\begin{aligned}
&\max_{\boldsymbol{\theta}} u(c_0) + \sum_{t=1}^{T} e^{-\delta t}\, \mathrm{E}[u(c_t)] \\
&\text{s.t. } c_0 \le e_0 - \boldsymbol{\theta}_0 \cdot \boldsymbol{P}_0, \\
&\qquad c_t \le e_t + D_t^{\boldsymbol{\theta}}, \quad t = 1, \ldots, T, \\
&\qquad c_0, c_1, \ldots, c_T \ge 0.
\end{aligned}$$

Applying (3.8), we can also write the constraint on time t consumption as

$$c_t \le e_t + \boldsymbol{\theta}_{t-1} \cdot (\boldsymbol{P}_t + \boldsymbol{D}_t) - \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t.$$

As in the one-period case we will assume that the non-negativity constraint on consumption is automatically satisfied and that the budget constraints hold as equalities. Therefore the problem can be reformulated as

$$\max_{\boldsymbol{\theta}} u\left(e_0 - \boldsymbol{\theta}_0 \cdot \boldsymbol{P}_0\right) + \sum_{t=1}^{T} e^{-\delta t}\, \mathrm{E}\left[u\left(e_t + \boldsymbol{\theta}_{t-1} \cdot (\boldsymbol{P}_t + \boldsymbol{D}_t) - \boldsymbol{\theta}_t \cdot \boldsymbol{P}_t\right)\right]. \tag{6.37}$$

The only terms involving the initially chosen portfolio $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, \ldots, \theta_{I0})^\top$ will be

$$u\left(e_0 - \boldsymbol{\theta}_0 \cdot \boldsymbol{P}_0\right) + e^{-\delta}\, \mathrm{E}\left[u\left(e_1 + \boldsymbol{\theta}_0 \cdot (\boldsymbol{P}_1 + \boldsymbol{D}_1) - \boldsymbol{\theta}_1 \cdot \boldsymbol{P}_1\right)\right].$$

The first-order condition with respect to θ_{i0} implies that

$$P_{i0} = \mathrm{E}\left[e^{-\delta} \frac{u'(c_1)}{u'(c_0)} (P_{i1} + D_{i1})\right].$$

This can also be verified by a variational argument as in the one-period analysis. More generally, the first-order condition with respect to θ_{it} implies that

$$P_{it} = \mathrm{E}_t\left[\frac{e^{-\delta} u'(c_{t+1})}{u'(c_t)} (P_{i,t+1} + D_{i,t+1})\right]. \tag{6.38}$$

Note that c_t and c_{t+1} in these expressions are the optimal consumption rates of the individual. Not surprisingly, this condition is equivalent to the conclusion in the one-period framework. In particular, we can define a state-price deflator $\zeta = (\zeta_t)_{t \in \mathcal{T}}$ from the individual's optimal consumption process by $\zeta_0 = 1$ and

$$\frac{\zeta_{t+1}}{\zeta_t} = \frac{e^{-\delta} u'(c_{t+1})}{u'(c_t)},$$

which means that

$$\begin{aligned}
\zeta_t &= \frac{\zeta_t}{\zeta_{t-1}} \frac{\zeta_{t-1}}{\zeta_{t-2}} \cdots \frac{\zeta_1}{\zeta_0} \\
&= e^{-\delta} \frac{u'(c_t)}{u'(c_{t-1})} e^{-\delta} \frac{u'(c_{t-1})}{u'(c_{t-2})} \cdots e^{-\delta} \frac{u'(c_1)}{u'(c_0)} \\
&= e^{-\delta t} \frac{u'(c_t)}{u'(c_0)}.
\end{aligned} \tag{6.39}$$

This is the individual's marginal rate of substitution between time 0 consumption and time t consumption.

6.3.2 Habit formation utility

Next let us consider a non-additive specification of preferences and for concreteness we study a specification with habit formation. The objective of the individual is

$$\max_{\boldsymbol{\theta}=(\boldsymbol{\theta}_t)_{t=0,1,\ldots,T-1}} \mathrm{E}\left[\sum_{t=0}^{T} e^{-\delta t} u(c_t, h_t)\right],$$

where h_t is the habit level at time t. First assume that the habit level at time t is some fraction of the consumption level at time $t-1$,

$$h_t = \beta c_{t-1},$$

and let $h_0 = 0$. Apply the variational argument given earlier. Let $c_0, c_1, \ldots, c_T$ denote the optimal consumption process and let $h_0, h_1, \ldots, h_T$ denote the resulting process for the habit level. What happens if the individual purchases $\varepsilon > 0$ units extra of asset i at time t and sells those ε units again at time $t+1$? Consumption at time t and $t+1$ will change to $c_t - \varepsilon P_{it}$ and $c_{t+1} + \varepsilon(D_{i,t+1} + P_{i,t+1})$, respectively. The perturbation will also affect the habit level. With the assumed habit formation, only the habit level at time $t+1$ and time $t+2$ will be changed. The new habit levels will be $h_{t+1} - \beta\varepsilon P_{it}$ at time $t+1$ and $h_{t+2} + \beta\varepsilon(D_{i,t+1} + P_{i,t+1})$. Therefore the change in total utility from time t and onwards will be

$$\begin{aligned} &u\left(c_t - \varepsilon P_{it}, h_t\right) - u(c_t, h_t) + e^{-\delta}\,\mathrm{E}_t\left[u\left(c_{t+1} + \varepsilon(D_{i,t+1} + P_{i,t+1}), h_{t+1} - \beta\varepsilon P_{it}\right) - u\left(c_{t+1}, h_{t+1}\right)\right] \\ &+ e^{-2\delta}\,\mathrm{E}_t\left[u\left(c_{t+2}, h_{t+2} + \beta\varepsilon(D_{i,t+1} + P_{i,t+1})\right) - u(c_{t+2}, h_{t+2})\right] \leq 0 \end{aligned}$$

Dividing by ε and letting $\varepsilon \to 0$, we obtain

$$\begin{aligned} &- P_{it} u_c(c_t, h_t) + e^{-\delta}\,\mathrm{E}_t\left[u_c(c_{t+1}, h_{t+1})\left(D_{i,t+1} + P_{i,t+1}\right) - \beta P_{it} u_h(c_{t+1}, h_{t+1})\right] \\ &+ e^{-2\delta}\,\mathrm{E}_t\left[u_h(c_{t+2}, h_{t+2})\beta\left(D_{i,t+1} + P_{i,t+1}\right)\right] \leq 0. \end{aligned}$$

Here a subscript on u indicate the partial derivative of u with respect to that variable. Again the opposite inequality can be reached by a similar argument. Replacing the inequality sign with an equality sign and rearranging, we arrive at

$$\begin{aligned} P_{it}\left(u_c(c_t, h_t) + \beta e^{-\delta}\,\mathrm{E}_t\left[u_h(c_{t+1}, h_{t+1})\right]\right) & \\ &= e^{-\delta}\,\mathrm{E}_t\left[\left(u_c(c_{t+1}, h_{t+1}) + \beta e^{-\delta} u_h(c_{t+2}, h_{t+2})\right)\left(D_{i,t+1} + P_{i,t+1}\right)\right] \\ &= e^{-\delta}\,\mathrm{E}_t\left[\left(u_c(c_{t+1}, h_{t+1}) + \beta e^{-\delta}\,\mathrm{E}_{t+1}[u_h(c_{t+2}, h_{t+2})]\right)\left(D_{i,t+1} + P_{i,t+1}\right)\right], \end{aligned}$$

where the last equality is due to the Law of Iterated Expectations. Consequently,

$$P_{it} = \mathrm{E}_t\left[e^{-\delta}\frac{u_c(c_{t+1}, h_{t+1}) + \beta e^{-\delta}\,\mathrm{E}_{t+1}[u_h(c_{t+2}, h_{t+2})]}{u_c(c_t, h_t) + \beta e^{-\delta}\,\mathrm{E}_t\left[u_h(c_{t+1}, h_{t+1})\right]}\left(D_{i,t+1} + P_{i,t+1}\right)\right], \tag{6.40}$$

and a state-price deflator can be defined by

$$\zeta_t = e^{-\delta t}\frac{u_c(c_t, h_t) + \beta e^{-\delta}\,\mathrm{E}_t[u_h(c_{t+1}, h_{t+1})]}{u_c(c_0, h_0) + \beta e^{-\delta}\,\mathrm{E}\left[u_h(c_1, h_1)\right]}. \tag{6.41}$$

Note that if $\beta = 0$ we are back to the case of time-additive utility.

It is probably more realistic that the habit level at time t depends on all the previous consumption rates but such that the consumption in recent periods are more important for the habit level than consumption in the far past. This can be captured by a specification like

$$h_t = \sum_{s=0}^{t-1} \beta^{t-s} c_s,$$

where β is a constant between 0 and 1. A change in the consumption at time t will now affect the habit levels at all future dates $t+1, t+2, \ldots, T$. Following the same line of argumentation as above, it can be shown that this problem will generate the state-price deflator

$$\zeta_t = e^{-\delta t} \frac{u_c(c_t, h_t) + \sum_{s=1}^{T-t} \beta^s e^{-\delta s} \, \mathrm{E}_t[u_h(c_{t+s}, h_{t+s})]}{u_c(c_0, h_0) + \sum_{s=1}^{T} \beta^s e^{-\delta s} \, \mathrm{E}\left[u_h(c_s, h_s)\right]}. \tag{6.42}$$

6.4 The continuous-time framework

In a continuous-time setting an individual consumes according to a non-negative continuous-time process $c = (c_t)$. Suppose that her preferences are described by time-additive expected utility so that the objective is to maximize $\mathrm{E}[\int_0^T e^{-\delta t} u(c_t) \, dt]$.

We will again go through a variational argument giving a link between the optimal consumption process and asset prices. For simplicity assume that assets pay no intermediate dividends. Suppose $c = (c_t)$ is the optimal consumption process for some agent and consider the following deviation from this strategy: at time 0 increase the investment in asset i by ε units. The extra costs of εP_{i0} implies a reduced consumption now. Let us suppose that the individual finances this extra investment by cutting down the consumption rate in the time interval $[0, \Delta t]$ for some small positive Δt by $\varepsilon P_{i0}/\Delta t$. The extra ε units of asset i is resold at time $t < T$, yielding a revenue of εP_{it}. This finances an increase in the consumption rate over $[t, t + \Delta t]$ by $\varepsilon P_{it}/\Delta t$. The consumption rates outside the intervals $[0, \Delta t]$ and $[t, t + \Delta t]$ will be unaffected. Given the optimality of $c = (c_t)$, we must have that

$$\mathrm{E}\left[\int_0^{\Delta t} e^{-\delta s} \left(u\left(c_s - \frac{\varepsilon P_{i0}}{\Delta t}\right) - u(c_s)\right) ds + \int_t^{t+\Delta t} e^{-\delta s} \left(u\left(c_s + \frac{\varepsilon P_{it}}{\Delta t}\right) - u(c_s)\right) ds\right] \leq 0.$$

Dividing by ε and letting $\varepsilon \to 0$, we obtain

$$\mathrm{E}\left[-\frac{P_{i0}}{\Delta t} \int_0^{\Delta t} e^{-\delta s} u'(c_s) \, ds + \frac{P_{it}}{\Delta t} \int_t^{t+\Delta t} e^{-\delta s} u'(c_s) \, ds\right] \leq 0.$$

Letting $\Delta t \to 0$, we arrive at

$$\mathrm{E}\left[-P_{i0} u'(c_0) + P_{it} e^{-\delta t} u'(c_t)\right] \leq 0,$$

or, equivalently,

$$P_{i0} u'(c_0) \geq \mathrm{E}\left[e^{-\delta t} P_{it} u'(c_t)\right].$$

The reverse inequality can be shown similarly so that we have that $P_{i0} u'(c_0) = \mathrm{E}[e^{-\delta t} P_t u'(c_t)]$ or more generally

$$P_{it} = \mathrm{E}_t\left[e^{-\delta(t'-t)} \frac{u'(c_{t'})}{u'(c_t)} P_{it'}\right], \quad t \leq t' \leq T.$$

With intermediate dividends this relation is slightly more complicated:

$$P_{it} = \mathrm{E}_t\left[\int_t^{t'} \delta_{is} P_{is} e^{-\delta(s-t)} \frac{u'(c_s)}{u'(c_t)}\, ds + e^{-\delta(t'-t)} \frac{u'(c_{t'})}{u'(c_t)} P_{it'}\right]. \tag{6.43}$$

We see that

$$\zeta_t = e^{-\delta t} \frac{u'(c_t)}{u'(c_0)} \tag{6.44}$$

defines a state-price deflator, exactly as in the discrete-time case.

If the market is complete, we can easily reach (6.44) solving step one of the two-step procedure suggested in Section 6.2.4. The problem is

$$\max_{c=(c_t)} \mathrm{E}\left[\int_0^T e^{-\delta t} u(c_t)\, dt\right] \qquad \text{s.t. } \mathrm{E}\left[\int_0^T \zeta_t c_t\right] \le e_0 + \mathrm{E}\left[\int_0^T \zeta_t e_t\, dt\right],$$

where $\zeta = (\zeta_t)$ is the unique state-price deflator. The left-hand side of the constraint is the present value of the consumption process, the right-hand side is the sum of the initial wealth e_0 and the present value of the income process. The Lagrangian for this problem is

$$\begin{aligned}\mathcal{L} &= \mathrm{E}\left[\int_0^T e^{-\delta t} u(c_t)\, dt\right] + \alpha\left(e_0 + \mathrm{E}\left[\int_0^T \zeta_t e_t\, dt\right] - \mathrm{E}\left[\int_0^T \zeta_t c_t\right]\right) \\ &= \alpha\left(e_0 + \mathrm{E}\left[\int_0^T \zeta_t e_t\, dt\right]\right) + \mathrm{E}\left[\int_0^T \left(e^{-\delta t} u(c_t) - \alpha \zeta_t c_t\right)\, dt\right].\end{aligned}$$

If we for each t and each state maximize the integrand in the last integral above, we will surely maximize the Lagrangian. The first-order condition is $e^{-\delta t} u'(c_t) = \alpha \zeta_t$ and since $\zeta_0 = 1$, we must have (6.44).

Exercise 6.10 considers an individual with habit formation in a continuous-time setting.

6.5 Dynamic programming

Above we have linked the optimal consumption plan of an individual to asset prices. In Chapter 8 we will see how this leads to consumption-based asset pricing models. While this link is quite intuitive and theoretically elegant, empirical tests and practical applications of the model suffer from the fact that available data on individual or aggregate consumption are of poor quality. For that purpose it is tempting to link asset prices to other variables for which better data are available. One way to provide such a link is to explain optimal consumption in terms of other variables. If c_t is a function of some variable, say x_t, then the equations in this chapter proves a link between asset prices and x. To figure out what explains the consumption choice of an individual we have to dig deeper into the utility maximization problem.

We will consider both a discrete-time and a continuous-time framework. In both cases we will for simplicity assume time-additive expected utility. A central element of the analysis is the indirect utility function of the individual, which is defined as the maximum expected utility of current and future consumption. In the discrete-time case the indirect utility at time t is defined as

$$J_t = \sup_{(c_s, \boldsymbol{\theta}_s)_{s=t}^T} \mathrm{E}_t\left[\sum_{s=t}^T e^{-\delta(s-t)} u(c_s)\right]. \tag{6.45}$$

There is no portfolio chosen at the final date so it is understood that $\boldsymbol{\theta}_T = \mathbf{0}$. Consumption at the final date is equal to the time T value of the portfolio purchased at $T-1$. In continuous time the corresponding definition is

$$J_t = \sup_{(c_s, \boldsymbol{\theta}_s)_{s\in[t,T]}} \mathrm{E}_t \left[\int_t^T e^{-\delta(s-t)} u(c_s)\, ds \right]. \tag{6.46}$$

For tractability it is necessary to assume that the indirect utility is a function of a limited number of variables. Surely the indirect utility of a finitely-lived individual will depend on the length of the remaining life and therefore on calender time t. The indirect utility at a given time t will also depend on the wealth W_t of the individual at that point in time. Other variables containing information about current or future investment opportunities or current or future income may have to be added. Suppose that that extra information can be captured by a single variable x_t. In that case the indirect utility is of the form

$$J_t = J(W_t, x_t, t)$$

for some function J.

We will show below that both in discrete and in continuous time the optimal consumption strategy will satisfy the so-called envelope condition:

$$u'(c_t) = J_W(W_t, x_t, t), \tag{6.47}$$

where J_W is the partial derivative of J with respect to W. This is an intuitive optimality condition. The left-hand side is the marginal utility of an extra unit of consumption at time t. The right-hand side is the marginal utility from investing an extra unit at time t in the optimal way. In an optimum these marginal utilities have to be equal. If that was not the case the allocation of wealth between consumption and investment should be reconsidered. For example, if $u'(c_t) > J_W(W_t, x_t, t)$, the consumption c_t should be increased and the amount invested should be decreased. Using the envelope condition, the state-price deflator derived from the individual's optimization problem can be rewritten as

$$\zeta_t = e^{-\delta t} \frac{u'(c_t)}{u'(c_0)} = e^{-\delta t} \frac{J_W(W_t, x_t, t)}{J_W(W_0, x_0, 0)}, \tag{6.48}$$

which links state prices to the optimally invested wealth of the individual and the variable x_t. This will be useful in constructing factor models in Chapter 9.

We will derive (6.47) using the dynamic programming technique. Along the way we will also find interesting conclusions on the optimal trading strategy of the individual. We do not make specific assumptions on utility functions or the dynamics of asset prices but stick to a general setting. More details and a lot of specific models are discussed in Munk (2005a). The basic references for the discrete-time models are Samuelson (1969), Hakansson (1970), Fama (1970, 1976), and Ingersoll (1987, Ch. 11). The basic references for the continuous-time models are Merton (1969, 1971, 1973b).

6.5.1 The discrete-time framework

Assume that a risk-free and d risky assets are traded. Let $\boldsymbol{\theta}_t$ denote the d-vector of units invested in the risky assets at time t and let θ_{0t} denote the units of the risk-free asset. Assume for simplicity

that the assets do not pay intermediate dividends. Similar to (3.10) the change in the wealth of the individual between time t and time $t+1$ is

$$W_{t+1} - W_t = \sum_{i=0}^{d} \theta_{it} \left(P_{i,t+1} - P_{it} \right) + y_t - c_t, \tag{6.49}$$

where y_t denotes the income received at time t. After receiving income and consuming at time t, the funds invested will be $W_t + y_t - c_t$. Assuming this is non-negative, we can represent the portfolio in terms of the fractions of this total investment invested in the different assets, i.e.

$$\pi_{it} = \frac{\theta_{it} P_{it}}{W_t + y_t - c_t}, \quad i = 0, 1, \ldots, d.$$

Define the portfolio weight vector of the risky assets by $\boldsymbol{\pi}_t = (\pi_{1t}, \ldots, \pi_{dt})^\top$. By construction the fraction invested in the risk-free asset is then given by $\pi_{0t} = 1 - \sum_{i=1}^{d} \pi_{it} = 1 - \boldsymbol{\pi}_t^\top \mathbf{1}$. The wealth at the end of the period can then be restated as

$$W_{t+1} = \left(W_t + y_t - c_t \right) R_{t+1}^W, \tag{6.50}$$

where

$$R_{t+1}^W = 1 + r_t^f + \boldsymbol{\pi}_t^\top \left[\boldsymbol{r}_{t+1} - r_t^f \mathbf{1} \right] \tag{6.51}$$

is the gross rate of return on the portfolio, r_t^f is the risk-free net rate of return, and $\boldsymbol{r}_{t+1}$ is the d-vector of the net rates of return of the risky assets over the period. Note that the only random variable (seen from time t) on the right-hand side of the above expressions is the return vector $\boldsymbol{r}_{t+1}$.

In the definition of indirect utility in (6.45) the maximization is over both the current and all future consumption rates and portfolios. This is clearly a quite complicated maximization problem. We will now show that we can alternatively perform a sequence of simpler maximization problems. This result is based on the following manipulations:

$$\begin{aligned}
J_t &= \sup_{(c_s, \boldsymbol{\pi}_s)_{s=t}^T} \mathrm{E}_t \left[\sum_{s=t}^{T} e^{-\delta(s-t)} u(c_s) \right] \\
&= \sup_{(c_s, \boldsymbol{\pi}_s)_{s=t}^T} \mathrm{E}_t \left[u(c_t) + \sum_{s=t+1}^{T} e^{-\delta(s-t)} u(c_s) \right] \\
&= \sup_{(c_s, \boldsymbol{\pi}_s)_{s=t}^T} \mathrm{E}_t \left[u(c_t) + \mathrm{E}_{t+1} \left[\sum_{s=t+1}^{T} e^{-\delta(s-t)} u(c_s) \right] \right] \\
&= \sup_{(c_s, \boldsymbol{\pi}_s)_{s=t}^T} \mathrm{E}_t \left[u(c_t) + e^{-\delta} \mathrm{E}_{t+1} \left[\sum_{s=t+1}^{T} e^{-\delta(s-(t+1))} u(c_s) \right] \right] \\
&= \sup_{c_t, \boldsymbol{\pi}_t} \mathrm{E}_t \left[u(c_t) + e^{-\delta} \sup_{(c_s, \boldsymbol{\pi}_s)_{s=t+1}^T} \mathrm{E}_{t+1} \left[\sum_{s=t+1}^{T} e^{-\delta(s-(t+1))} u(c_s) \right] \right]
\end{aligned}$$

Here, the first equality is simply due to the definition of indirect utility, the second equality comes from separating out the first term of the sum, the third equality is valid according to the law of iterated expectations, the fourth equality comes from separating out the discount term $e^{-\delta}$, and the final equality is due to the fact the only the inner expectation depends on future consumption rates

and portfolios. Noting that the inner supremum is by definition the indirect utility at time $t+1$, we arrive at

$$J_t = \sup_{c_t, \boldsymbol{\pi}_t} \mathrm{E}_t \left[u(c_t) + e^{-\delta} J_{t+1} \right] = \sup_{c_t, \boldsymbol{\pi}_t} \left\{ u(c_t) + e^{-\delta} \mathrm{E}_t \left[J_{t+1} \right] \right\}. \tag{6.52}$$

This equation is called the **Bellman equation**, and the indirect utility J is said to have the **dynamic programming property**. The decision to be taken at time t is split up in two: (1) the consumption and portfolio decision for the current period and (2) the consumption and portfolio decisions for all future periods. We take the decision for the current period assuming that we will make optimal decisions in all future periods. Note that this does not imply that the decision for the current period is taken independently from future decisions. We take into account the effect that our current decision has on the maximum expected utility we can get from all future periods. The expectation $\mathrm{E}_t \left[J_{t+1} \right]$ will depend on our choice of c_t and $\boldsymbol{\pi}_t$.[1]

The dynamic programming property is the basis for a backward iterative solution procedure. First, note that $J_T = u(c_T) = u(W_T)$, and c_{T-1} and $\boldsymbol{\pi}_{T-1}$ are chosen to maximize

$$u(c_{T-1}) + e^{-\delta} \mathrm{E}_{T-1} \left[u(W_T) \right],$$

where

$$W_T = (W_{T-1} + y_{T-1} - c_{T-1}) R_T^W, \quad R_T^W = 1 + r_{T-1}^f + \boldsymbol{\pi}_{T-1}^\top \left[\boldsymbol{r}_T - r_{T-1}^f \mathbf{1} \right].$$

This is done for each possible scenario at time $T-1$ and gives us J_{T-1}. Then c_{T-2} and $\boldsymbol{\pi}_{T-2}$ are chosen to maximize

$$u(c_{T-2}) + e^{-\delta} \mathrm{E}_{T-2} \left[J_{T-1} \right],$$

and so on until time zero is reached. Since we have to perform a maximization for each scenario of the world at every point in time, we have to make assumptions on the possible scenarios at each point in time before we can implement the recursive procedure. The optimal decisions at any time are expected to depend on the wealth level of the agent at that date, but also on the value of other time-varying state variables that affect future returns on investment (e.g. the interest rate level) and future income levels. To be practically implementable only a few state variables can be incorporated. Also, these state variables must follow Markov processes so only the current values of the variables are relevant for the maximization at a given point in time.

Suppose that the relevant information is captured by a one-dimensional Markov process $x = (x_t)$ so that the indirect utility at any time $t = 0, 1, \ldots, T$ can be written as $J_t = J(W_t, x_t, t)$. Then the dynamic programming equation (6.52) becomes

$$J(W_t, x_t, t) = \sup_{c_t, \boldsymbol{\pi}_t} \left\{ u(c_t) + e^{-\delta} \mathrm{E}_t \left[J(W_{t+1}, x_{t+1}, t+1) \right] \right\} \tag{6.53}$$

with terminal condition $J(W_T, x_T, T) = u(W_T)$. Doing the maximization we have to remember that W_{t+1} will be affected by the choice of c_t and $\boldsymbol{\pi}_t$, cf. Equation (6.50). In particular, we see that

$$\frac{\partial W_{t+1}}{\partial c_t} = -R_{t+1}^W, \qquad \frac{\partial W_{t+1}}{\partial \boldsymbol{\pi}_t} = (W_t + y_t - c_t) \left(\boldsymbol{r}_{t+1} - r_t^f \mathbf{1} \right).$$

[1] Readers familiar with option pricing theory may note the similarity to the problem of determining the optimal exercise strategy of a Bermudan/American option. However, for that problem the decision to be taken is much simpler (exercise or not) than for the consumption/portfolio problem.

The first-order condition for the maximization with respect to c_t is

$$u'(c_t) + e^{-\delta}\,\mathrm{E}_t\left[J_W(W_{t+1}, x_{t+1}, t+1)\frac{\partial W_{t+1}}{\partial c_t}\right] = 0, \tag{6.54}$$

which implies that

$$u'(c_t) = e^{-\delta}\,\mathrm{E}_t\left[J_W(W_{t+1}, x_{t+1}, t+1)R^W_{t+1}\right]. \tag{6.55}$$

The first-order condition for the maximization with respect to $\boldsymbol{\pi}_t$ is

$$\mathrm{E}_t\left[J_W(W_{t+1}, x_{t+1}, t+1)\frac{\partial W_{t+1}}{\partial \boldsymbol{\pi}_t}\right] = 0, \tag{6.56}$$

which implies that

$$\mathrm{E}_t\left[J_W(W_{t+1}, x_{t+1}, t+1)\left(\boldsymbol{r}_{t+1} - r^f_t 1\right)\right] = 0. \tag{6.57}$$

While we cannot generally solve for the optimal decisions, we can show that the envelope condition (6.47) holds. First note that for the optimal choice $\hat{c}_t, \hat{\boldsymbol{\pi}}_t$ we have that

$$J(W_t, x_t, t) = u(\hat{c}_t) + e^{-\delta}\,\mathrm{E}_t\left[J(\hat{W}_{t+1}, x_{t+1}, t+1)\right],$$

where $\hat{W}_{t+1}$ is next period's wealth using $\hat{c}_t, \hat{\boldsymbol{\pi}}_t$. Taking derivatives with respect to W_t in this equation, and acknowledging that $\hat{c}_t$ and $\hat{\boldsymbol{\pi}}_t$ will in general depend on W_t, we get

$$J_W(W_t, x_t, t) = u'(\hat{c}_t)\frac{\partial \hat{c}_t}{\partial W_t} + e^{-\delta}\,\mathrm{E}_t\left[J_W(\hat{W}_{t+1}, x_{t+1}, t+1)\frac{\partial \hat{W}_{t+1}}{\partial W_t}\right],$$

where

$$\frac{\partial \hat{W}_{t+1}}{\partial W_t} = R^W_{t+1}\left(1 - \frac{\partial \hat{c}_t}{\partial W_t}\right) + (W_t + y_t - c_t)\left(\frac{\partial \hat{\boldsymbol{\pi}}_t}{\partial W_t}\right)^\top\left(\boldsymbol{r}_{t+1} - r^f_t \mathbf{1}\right).$$

Inserting this and rearranging terms, we get

$$\begin{aligned} J_W(W_t, x_t, t) = {} & e^{-\delta}\,\mathrm{E}_t\left[J_W(\hat{W}_{t+1}, x_{t+1}, t+1)R^W_{t+1}\right] \\ & + \left(u'(\hat{c}_t) - e^{-\delta}\,\mathrm{E}_t\left[J_W(\hat{W}_{t+1}, x_{t+1}, t+1)R^W_{t+1}\right]\right)\frac{\partial \hat{c}_t}{\partial W_t} \\ & + (W_t + y_t - c_t)\,e^{-\delta}\left(\frac{\partial \hat{\boldsymbol{\pi}}_t}{\partial W_t}\right)^\top \mathrm{E}_t\left[J_W(\hat{W}_{t+1}, x_{t+1}, t+1)\left(\boldsymbol{r}_{t+1} - r^f_t \mathbf{1}\right)\right]. \end{aligned}$$

On the right-hand side the last two terms are zero due to the first-order conditions and only the leading term remains, i.e.

$$J_W(W_t, x_t, t) = e^{-\delta}\,\mathrm{E}_t\left[J(\hat{W}_{t+1}, x_{t+1}, t+1)R^W_{t+1}\right].$$

Combining this with (6.55) we obtain the envelope condition (6.47).

6.5.2 The continuous-time framework

As in the discrete-time setting above assume that an instantaneously risk-free asset with a continuously compounded risk-free rate of r^f_t and d risky assets are traded. We assume for simplicity that the assets pay no intermediate dividends and write their price dynamics as

$$d\boldsymbol{P}_t = \mathrm{diag}(\boldsymbol{P}_t)\left[\boldsymbol{\mu}_t\,dt + \underline{\sigma}_t\,d\boldsymbol{z}_t\right], \tag{3.5}$$

where $\boldsymbol{z} = (z_1, \ldots, z_d)^\top$ is a d-dimensional standard Brownian motion. We can write this componentwise as

$$dP_{it} = P_{it}\left[\mu_{it}\,dt + \sum_{j=1}^{d}\sigma_{ijt}\,dz_{jt}\right], \quad i = 1, \ldots, d.$$

The instantaneous rate of return on asset i is given by dP_{it}/P_{it}. The d-vector $\boldsymbol{\mu}_t = (\mu_{1t}, \ldots, \mu_{dt})^\top$ contains the expected rates of return and the $(d\times d)$-matrix $\underline{\underline{\sigma}}_t = (\sigma_{ijt})_{i,j=1}^d$ measures the sensitivities of the risky asset prices with respect to exogenous shocks so that the $(d\times d)$-matrix $\underline{\underline{\Sigma}}_t = \underline{\underline{\sigma}}_t\underline{\underline{\sigma}}_t^\top$ contains the variance and covariance rates of instantaneous rates of return. We assume that $\underline{\underline{\sigma}}_t$ is non-singular and, hence, we can define the market price of risk associated with $\boldsymbol{z}$ as

$$\boldsymbol{\lambda}_t = \underline{\underline{\sigma}}_t^{-1}(\boldsymbol{\mu}_t - r_t^f\mathbf{1}),$$

so that

$$\boldsymbol{\mu}_t = r_t\mathbf{1} + \underline{\underline{\sigma}}_t\boldsymbol{\lambda}_t,$$

i.e. $\mu_{it} = r_t + \sum_{j=1}^d \sigma_{ijt}\lambda_{jt}$. We can now rewrite the price dynamics as

$$d\boldsymbol{P}_t = \operatorname{diag}(\boldsymbol{P}_t)\left[\left(r_t\mathbf{1} + \underline{\underline{\sigma}}_t\boldsymbol{\lambda}_t\right)\,dt + \underline{\underline{\sigma}}_t\,d\boldsymbol{z}_t\right].$$

We represent the trading strategy by the portfolio weight process $\boldsymbol{\pi} = (\boldsymbol{\pi}_t)$, where $\boldsymbol{\pi}_t$ is the d-vector of fractions of wealth invested in the d risky assets at time t. Again the weight of the risk-free asset is $\pi_{0t} = 1 - \boldsymbol{\pi}_t^\top\mathbf{1} = 1 - \sum_{i=1}^d \pi_{it}$. Analogous to (3.14), the wealth dynamics can be written as

$$dW_t = W_t\left[r_t^f + \boldsymbol{\pi}_t^\top\underline{\underline{\sigma}}_t\boldsymbol{\lambda}_t\right]\,dt + [y_t - c_t]\,dt + W_t\boldsymbol{\pi}_t^\top\underline{\underline{\sigma}}_t\,d\boldsymbol{z}_t. \tag{6.58}$$

For simplicity we assume in the following that the agent receives no labor income, i.e. $y_t \equiv 0$. We also assume that a single variable x_t captures the time t information about investment opportunities so that, in particular,

$$r_t^f = r^f(x_t), \quad \boldsymbol{\mu}_t = \boldsymbol{\mu}(x_t, t), \quad \underline{\underline{\sigma}}_t = \underline{\underline{\sigma}}(x_t, t),$$

where r^f, $\boldsymbol{\mu}$, and $\underline{\underline{\sigma}}$ now (also) denote sufficiently well-behaved functions. The market price of risk is also given by the state variable:

$$\boldsymbol{\lambda}(x_t) = \underline{\underline{\sigma}}(x_t, t)^{-1}\left(\boldsymbol{\mu}(x_t, t) - r^f(x_t)\mathbf{1}\right).$$

Note that we have assumed that the short-term interest rate r_t^f and the market price of risk vector $\boldsymbol{\lambda}_t$ do not depend on calendar time directly. The fluctuations in r_t^f and $\boldsymbol{\lambda}_t$ over time are presumably not due to the mere passage of time, but rather due to variations in some more fundamental economic variables. In contrast, the expected rates of returns and the price sensitivities of some assets will depend directly on time, e.g. the volatility and the expected rate of return on a bond will depend on the time-to-maturity of the bond and therefore on calendar time.

Now the wealth dynamics for a given portfolio and consumption strategy is

$$dW_t = W_t\left[r^f(x_t) + \boldsymbol{\pi}_t^\top\underline{\underline{\sigma}}(x_t, t)\boldsymbol{\lambda}(x_t)\right]\,dt - c_t\,dt + W_t\boldsymbol{\pi}_t^\top\underline{\underline{\sigma}}(x_t, t)\,d\boldsymbol{z}_t.$$

The state variable x is assumed to follow a one-dimensional diffusion process

$$dx_t = m(x_t)\,dt + \boldsymbol{v}(x_t)^\top\,d\boldsymbol{z}_t + \hat{v}(x_t)\,d\hat{z}_t, \tag{6.59}$$

where $\hat{z} = (\hat{z}_t)$ is a one-dimensional standard Brownian motion independent of $\boldsymbol{z} = (\boldsymbol{z}_t)$. Hence, if $\hat{v}(x_t) \neq 0$, there is an exogenous shock to the state variable that cannot be hedged by investments in the financial market. In other words, the financial market is incomplete. Conversely, if $\hat{v}(x_t)$ is identically equal to zero, the financial market is complete. We shall consider examples of both cases later. The d-vector $\boldsymbol{v}(x_t)$ represents the sensitivity of the state variable with respect to the exogenous shocks to market prices. Note that the d-vector $\underline{\underline{\sigma}}(x,t)\boldsymbol{v}(x)$ is the vector of instantaneous covariance rates between the returns on the risky assets and the state variable. Under the assumptions made above the indirect utility at time t is $J_t = J(W_t, x_t, t)$.

How do we implement the dynamic programming principle in continuous time? First consider a discrete-time approximation with time set $\{0, \Delta t, 2\Delta t, \ldots, T = N\Delta t\}$. The Bellman equation corresponding to this discrete-time utility maximization problem is

$$J(W,x,t) = \sup_{c_t \geq 0, \boldsymbol{\pi}_t \in \mathbb{R}^d} \left\{ u(c_t)\Delta t + e^{-\delta \Delta t} \, \mathrm{E}_t \left[J(W_{t+\Delta t}, x_{t+\Delta t}, t+\Delta t) \right] \right\}, \tag{6.60}$$

cf. (6.52). Here c_t and $\boldsymbol{\pi}_t$ are held fixed over the interval $[t, t+\Delta t)$. If we multiply by $e^{\delta \Delta t}$, subtract $J(W,x,t)$, and then divide by Δt, we get

$$\frac{e^{\delta \Delta t} - 1}{\Delta t} J(W,x,t) = \sup_{c_t \geq 0, \boldsymbol{\pi}_t \in \mathbb{R}^d} \left\{ e^{\delta \Delta t} u(c_t) + \frac{1}{\Delta t} \, \mathrm{E}_t \left[J(W_{t+\Delta t}, x_{t+\Delta t}, t+\Delta t) - J(W,x,t) \right] \right\}. \tag{6.61}$$

When we let $\Delta t \to 0$, we have that (by l'Hôspital's rule)

$$\frac{e^{\delta \Delta t} - 1}{\Delta t} \to \delta,$$

and that (by definition of the drift of a process)

$$\frac{1}{\Delta t} \, \mathrm{E}_t \left[J(W_{t+\Delta t}, x_{t+\Delta t}, t+\Delta t) - J(W,x,t) \right] \tag{6.62}$$

will approach the drift of J at time t, which according to Itô's Lemma is given by

$$\begin{aligned} &\frac{\partial J}{\partial t}(W,x,t) + J_W(W,x,t) \left(W \left[r(x) + \boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}(x,t) \boldsymbol{\lambda}(x) \right] - c_t \right) \\ &\quad + \frac{1}{2} J_{WW}(W,x,t) W^2 \boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}(x,t) \underline{\underline{\sigma}}(x,t)^\top \boldsymbol{\pi}_t + J_x(W,x,t) m(x) \\ &\quad + \frac{1}{2} J_{xx}(W,x,t) (\boldsymbol{v}(x)^\top \boldsymbol{v}(x) + \hat{v}(x)^2) + J_{Wx}(W,x,t) W \boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}(x,t) \boldsymbol{v}(x). \end{aligned}$$

The limit of (6.61) is therefore

$$\begin{aligned} \delta J(W,x,t) = \sup_{c \geq 0, \boldsymbol{\pi} \in \mathbb{R}^d} \Big\{ & u(c) + \frac{\partial J}{\partial t}(W,x,t) + J_W(W,x,t) \left(W \left[r(x) + \boldsymbol{\pi}^\top \underline{\underline{\sigma}}(x,t) \boldsymbol{\lambda}(x) \right] - c \right) \\ & + \frac{1}{2} J_{WW}(W,x,t) W^2 \boldsymbol{\pi}^\top \underline{\underline{\sigma}}(x,t) \underline{\underline{\sigma}}(x,t)^\top \boldsymbol{\pi} + J_x(W,x,t) m(x) \\ & + \frac{1}{2} J_{xx}(W,x,t) (\boldsymbol{v}(x)^\top \boldsymbol{v}(x) + \hat{v}(x)^2) \\ & + J_{Wx}(W,x,t) W \boldsymbol{\pi}^\top \underline{\underline{\sigma}}(x,t) \boldsymbol{v}(x) \Big\}. \end{aligned} \tag{6.63}$$

This is called the Hamilton-Jacobi-Bellman (HJB) equation corresponding to the dynamic optimization problem. Subscripts on J denote partial derivatives, however we will write the partial derivative with respect to time as $\partial J / \partial t$ to distinguish it from the value J_t of the indirect utility

process. The HJB equation involves the supremum over the feasible time t consumption rates and portfolios (*not* the supremum over the entire processes!) and is therefore a highly non-linear second-order partial differential equation.

From the analysis above we will expect that the indirect utility function $J(W,x,t)$ solves the HJB equation for all possible values of W and x and all $t \in [0,T)$ and that it satisfies the terminal condition $J(W,x,T)=0$. (We could allow for some utility of terminal consumption or wealth, e.g. representing the utility of leaving money for your heirs. Then the terminal condition should be of the form $J(W,x,T)=\bar{u}(W)$.) This can be supported formally by the so-called verification theorem. The solution procedure is therefore as follows: (1) solve the maximization problem embedded in the HJB-equation giving a candidate for the optimal strategies expressed in terms of the yet unknown indirect utility function and its derivatives. (2) substitute the candidate for the optimal strategies into the HJB-equation, ignore the sup-operator, and solve the resulting partial differential equation for $J(W,x,t)$. Such a solution will then also give the candidate optimal strategies in terms of W, x, and t.[2]

Let us find the first-order conditions of the maximization in (6.63). The first-order condition with respect to c gives us immediately the envelope condition (6.47), which we were really looking for. Nevertheless, let us also look at the first-order condition with respect to $\boldsymbol{\pi}$, i.e.

$$J_W(W,x,t)W\underline{\underline{\sigma}}(x,t)\boldsymbol{\lambda}(x) + J_{WW}(W,x,t)W^2\underline{\underline{\sigma}}(x,t)\underline{\underline{\sigma}}(x,t)^\top\boldsymbol{\pi} + J_{Wx}(W,x,t)W\underline{\underline{\sigma}}(x,t)\boldsymbol{v}(x) = 0$$

so that the optimal portfolio is

$$\boldsymbol{\pi}_t = -\frac{J_W(W_t,x_t,t)}{W_t J_{WW}(W_t,x_t,t)}\left(\underline{\underline{\sigma}}(x_t,t)^\top\right)^{-1}\boldsymbol{\lambda}(x_t) - \frac{J_{Wx}(W_t,x_t,t)}{W_t J_{WW}(W_t,x_t,t)}\left(\underline{\underline{\sigma}}(x_t,t)^\top\right)^{-1}\boldsymbol{v}(x_t). \quad (6.64)$$

As the horizon shrinks, the indirect utility function $J(W,x,t)$ approaches the terminal utility which is independent of the state x. Consequently, the derivative $J_{Wx}(W,x,t)$ and hence the last term of the portfolio will approach zero as $t \to T$. Short-sighted investors pick a portfolio given by the first term on the right-hand side. We can interpret the second term as an intertemporal hedge term since it shows how a long-term investor will deviate from the short-term investor. The last term will also disappear for "non-instantaneous" investors in three special cases:

(1) there is no x: investment opportunities are constant; there is nothing to hedge.

(2) $J_{Wx}(W,x,t) \equiv 0$: The state variable does not affect the marginal utility of the investor. This turns out to be the case for investors with logarithmic utility. Such an investor is *not interested in hedging* changes in the state variable.

(3) $v(x) \equiv 0$: The state variable is uncorrelated with instantaneous returns on the traded assets. In this case the investor is *not able to hedge* changes in the state variable.

In all other cases the state variable induces an additional term to the optimal portfolio relative to the case of constant investment opportunities.

[2]There is really also a third step, namely to check that the assumptions made along the way and the technical conditions needed for the verification theorem to apply are all satisfied. The standard version of the verification theorem is precisely stated and proofed in, e.g. Theorem 11.2.2 in Øksendal (1998) or Theorem III.8.1 in Fleming and Soner (1993). The technical conditions of the standard version are not always satisfied in concrete consumption-portfolio problems, however, but at least for some concrete problems a version with an appropriate set of conditions can be found; see, e.g., Korn and Kraft (2001) and Kraft (2004).

In general, we have three-fund separation in the sense that all investors will combine the risk-free asset, the "tangency portfolio" of risky assets given by the portfolio weights

$$\boldsymbol{\pi}_t^{\tan} = \frac{1}{\mathbf{1}^\top \left(\underline{\underline{\sigma}}(x_t,t)^\top\right)^{-1} \boldsymbol{\lambda}(x_t)} \left(\underline{\underline{\sigma}}(x_t,t)^\top\right)^{-1} \boldsymbol{\lambda}(x_t),$$

and the "hedge portfolio" given by the weights

$$\boldsymbol{\pi}_t^{\mathrm{hdg}} = \frac{1}{\mathbf{1}^\top \left(\underline{\underline{\sigma}}(x_t,t)^\top\right)^{-1} \boldsymbol{v}(x_t)} \left(\underline{\underline{\sigma}}(x_t,t)^\top\right)^{-1} \boldsymbol{v}(x_t).$$

Inserting the definition of $\boldsymbol{\lambda}$ we can rewrite the expression for the tangency portfolio as

$$\boldsymbol{\pi}_t^{\tan} = \frac{1}{\mathbf{1}^\top \underline{\underline{\Sigma}}(x_t,t)^{-1} \left(\boldsymbol{\mu}(x_t,t) - r^f(x_t)\mathbf{1}\right)} \underline{\underline{\Sigma}}(x_t,t)^{-1} \left(\boldsymbol{\mu}(x_t,t) - r^f(x_t)\mathbf{1}\right),$$

which is analogous to the tangency portfolio in the one-period mean-variance analysis, cf. (6.33).

Substituting the candidate optimal values of c and π back into the HJB equation and gathering terms, we get the second order PDE

$$\begin{aligned}\delta J(W,x,t) = {} & u\left(I_u(J_W(W,x,t))\right) - J_W(W,x,t) I_u(J_W(W,x,t)) + \frac{\partial J}{\partial t}(W,x,t) + r(x) W J_W(W,x,t) \\ & - \frac{1}{2}\frac{J_W(W,x,t)^2}{J_{WW}(W,x,t)} \|\boldsymbol{\lambda}(x)\|^2 + J_x(W,x,t) m(x) + \frac{1}{2} J_{xx}(W,x,t) \left(\|\boldsymbol{v}(x)\|^2 + \hat{v}(x)^2\right) \\ & - \frac{1}{2}\frac{J_{Wx}(W,x,t)^2}{J_{WW}(W,x,t)} \|\boldsymbol{v}(x)\|^2 - \frac{J_W(W,x,t) J_{Wx}(W,x,t)}{J_{WW}(W,x,t)} \boldsymbol{\lambda}(x)^\top \boldsymbol{v}(x).\end{aligned} \tag{6.65}$$

Although the PDE (6.65) looks very complicated, closed-form solutions can be found for a number of interesting model specifications. See, e.g., Munk (2005a).

If more than one, say k, variables are necessary to capture the information about investment opportunities, the optimal portfolio will involve k hedge portfolios beside the tangency portfolio and the risk-free asset so that $(k+2)$-fund separation holds.

Nielsen and Vassalou (2006) have shown that the only characteristics of investment opportunities that will induce intertemporal hedging is the short-term risk-free interest rate r_t^f and $\|\boldsymbol{\lambda}_t\|$, which is the maximum Sharpe ratio obtainable at the financial market. Since r_t^f is the intercept and $\|\boldsymbol{\lambda}_t\|$ the slope of the instantaneous mean-variance frontier this result makes good sense. Long-term investors are concerned about the variations in the investments that are good in the short run.

6.6 Concluding remarks

This chapter has characterized the optimal consumption and portfolio choice of an individual by the first-order condition of her utility maximization problem. The characterization provides a link between asset prices and the optimal consumption plan of any individual. In the next chapter we will look at the market equilibrium.

6.7 Exercises

EXERCISE 6.1 Consider a one-period economy and an individual with a time-additive but state-dependent expected utility so that the objective is

$$\max_{\boldsymbol{\theta}} u(c_0, X_0) + e^{-\delta} \mathrm{E}[u(c, X)].$$

The decisions of the individual does not affect X_0 or X. For example, X_0 and X could be the aggregate consumption in the economy at time 0 and time 1, respectively, which are not significantly affected by the consumption of a small individual. What is the link between prices and marginal utility in this case? What if $u(c, X) = \frac{1}{1-\gamma}(c - X)^{1-\gamma}$? What if $u(c, X) = \frac{1}{1-\gamma}(c/X)^{1-\gamma}$?

EXERCISE 6.2 Consider a one-period economy where four basic financial assets are traded without portfolio constraints or transaction costs. There are four possible end-of-period states of the economy. The objective state probabilities and the prices and state-contingent dividends of the assets are given in the following table:

	state 1	state 2	state 3	state 4	
probability	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{3}$	
	state-contingent dividend				price
Asset 1	1	2	2	2	$\frac{3}{2}$
Asset 2	1	3	0	0	$\frac{7}{6}$
Asset 3	0	0	1	1	$\frac{1}{3}$
Asset 4	3	2	1	1	$\frac{3}{2}$

The economy is known to be arbitrage-free.

(a) Show that asset 4 is redundant and verify that the price of asset 4 is identical to the price of the portfolio of the other assets that replicates asset 4.

(b) Is the market complete?

(c) Show that the vector $\boldsymbol{\psi}^* = \left(\frac{1}{6}, \frac{1}{3}, \frac{1}{6}, \frac{1}{6}\right)$ is a valid state-price vector and that it is in the set of dividends spanned by the basic assets. Characterize the set of all valid state-price vectors.

(d) Show that the vector $\boldsymbol{\zeta}^* = \left(1, \frac{4}{3}, \frac{4}{7}, \frac{4}{7}\right)$ is a valid state-price deflator and that it is in the set of dividends spanned by the basic assets. Show that any state-price deflator must be a vector of the form $\left(1, \frac{4}{3}, y, 1 - \frac{3}{4}y\right)$, where $y \in \left(0, \frac{4}{3}\right)$.

(e) Show that it is possible to construct a risk-free asset from the four basic assets. What is the risk-free interest rate?

In the following consider an individual maximizing $u(c_0) + \beta\,\mathrm{E}[u(c)]$, where c_0 denotes consumption at the beginning of the period and c denotes the state-dependent consumption at the end of the period. Assume $u(c) = c^{1-\gamma}/(1-\gamma)$. For $\omega \in \{1, 2, 3, 4\}$, let c_ω denote the end-of-period consumption if state ω is realized.

(f) Show that the optimal consumption plan must satisfy

$$c_2 = c_1 \left(\frac{4}{3}\right)^{-1/\gamma}, \qquad c_4 = c_0 \left(\frac{1}{\beta} - \frac{3}{4}\left(\frac{c_3}{c_0}\right)^{-\gamma}\right)^{-1/\gamma}.$$

For the remainder of the problem it is assumed that the individual has identical income/endowment in states 3 and 4.

(g) Explain why c_3 and c_4 must be identical, and hence that the optimal consumption plan must have

$$c_3 = c_4 = c_0 \left(\frac{4}{7\beta}\right)^{-1/\gamma}.$$

(h) Assuming $\gamma = 2$, $\beta = \frac{6}{7}$, and $c_0 = 1$, find the optimal state-dependent end-of-period consumption, i.e. c_1, c_2, c_3, c_4.

(i) What is the present value of the optimal consumption plan?

(j) Assuming that the individual receives no end-of-period income in any state, find an optimal portfolio for this individual.

EXERCISE 6.3 Consider a one-period economy with four possible, equally likely, states at the end of the period. The agents in the economy consume at the beginning of the period (time 0) and at the end of the period (time 1). The agents can choose between three different consumption plans as shown in the following table:

	consumption	state-contingent time 1 consumption			
	at time 0	state 1	state 2	state 3	state 4
Consumption plan 1	8	9	16	9	4
Consumption plan 2	8	9	9	9	9
Consumption plan 3	8	4	16	25	4

Denote the time 0 consumption by c_0, the uncertain consumption at time 1 by c, and the consumption at time 1 in case state ω is realized by c_ω.

(a) Consider an agent with logarithmic utility,

$$U(c_0, c_1, c_2, c_3, c_4) = \ln c_0 + \mathrm{E}[\ln c] = \ln c_0 + \sum_{\omega=1}^{4} p_\omega \ln c_\omega,$$

where p_ω is the probability that state ω is realized. Compute the utility for each of the three possible consumption plans and determine the optimal consumption plan. Find the associated state-price vector. Using this state-price vector, what is the price at the beginning of the period of an asset that gives a payoff of 2 in states 1 and 4 and a payoff of 1 in states 2 and 3?

(b) Answer the same questions with the alternative time-additive square-root utility,

$$U(c_0, c_1, c_2, c_3, c_4) = \sqrt{c_0} + \mathrm{E}[\sqrt{c}] = \sqrt{c_0} + \sum_{\omega=1}^{4} p_\omega \sqrt{c_\omega}.$$

(c) Answer the same questions with the alternative habit-style square-root utility,

$$U(c_0, c_1, c_2, c_3, c_4) = \sqrt{c_0} + \mathrm{E}[\sqrt{c - 0.5c_0}] = \sqrt{c_0} + \sum_{\omega=1}^{4} p_\omega \sqrt{c_\omega - 0.5c_0}.$$

EXERCISE 6.4 Show Equation (6.24).

EXERCISE 6.5 Using the Lagrangian characterization of the mean-variance frontier, show that for any mean-variance efficient return $R^{\boldsymbol{\pi}}$ different from the minimum-variance portfolio there is a unique mean-variance efficient return $R^{z(\boldsymbol{\pi})}$ with $\operatorname{Cov}[R^{\boldsymbol{\pi}}, R^{z(\boldsymbol{\pi})}] = 0$. Show that $\operatorname{E}[R^{z(\boldsymbol{\pi})}] = (A - B\operatorname{E}[R^{\boldsymbol{\pi}}])/(B - C\operatorname{E}[R^{\boldsymbol{\pi}}])$.

EXERCISE 6.6 Let $R_{\min}$ denote the return on the minimum-variance portfolio. Let R be any other return, efficient or not. Show that $\operatorname{Cov}[R, R_{\min}] = \operatorname{Var}[R_{\min}]$.

EXERCISE 6.7 Let R_1 denote the return on a mean-variance efficient portfolio and let R_2 denote the return on another not necessarily efficient portfolio with $\operatorname{E}[R_2] = \operatorname{E}[R_1]$. Show that $\operatorname{Cov}[R_1, R_2] = \operatorname{Var}[R_1]$ and conclude that R_1 and R_2 are positively correlated.

EXERCISE 6.8 Think of the mean-variance framework in a one-period economy. Show that if there is a risk-free asset, then any two mean-variance efficient returns (different from the risk-free return) are either perfectly positively correlated or perfectly negatively correlated. Is that also true if there is no risk-free asset?

EXERCISE 6.9 In a one-period model where the returns of all the risky assets are normally distributed, any greedy and risk-averse investor will place herself on the upward-sloping part of the mean-variance frontier. But where? Consider an agent that maximizes expected utility of end-of-period wealth with a negative exponential utility function $u(W) = -e^{-aW}$ for some constant a. Suppose that M risky assets (with normally distributed returns) and one risk-free asset are traded. What is the optimal portfolio of the agent? Where is the optimal portfolio located on the mean-variance frontier?

EXERCISE 6.10 Look at an individual with habit formation living in a continuous-time complete market economy. The individual wants to maximize his expected utility

$$\operatorname{E}\left[\int_0^T e^{-\delta t} u(c_t, h_t)\, dt\right],$$

where the habit level h_t is given by

$$h_t = h_0 e^{-\alpha t} + \beta \int_0^t e^{-\alpha(t-u)} c_u\, du.$$

We can write the budget constraint as

$$\operatorname{E}\left[\int_0^T \zeta_t c_t\, dt\right] \leq W_0,$$

where $\zeta = (\zeta_t)$ is the state-price deflator and W_0 is the initial wealth of the agent (including the present value of any future non-financial income).

(a) Show that $dh_t = (\beta c_t - \alpha h_t)\, dt$. What condition on α and β will ensure that the habit level declines, when current consumption equals the habit level?

(b) Show that the state-price deflator is linked to optimal consumption by the relation

$$\zeta_t = ke^{-\delta t}\left\{u_c(c_t,h_t) + \beta\,\mathrm{E}_t\left[\int_t^T e^{-(\delta+\alpha)(s-t)}u_h(c_s,h_s)\,ds\right]\right\} \qquad (*)$$

for some appropriate constant k. *Hint: First consider what effect consumption at time t has on future habit levels.*

(c) How does (*) look when $u(c,h) = \frac{1}{1-\gamma}(c-h)^{1-\gamma}$?

Chapter 7

Market equilibrium

7.1 Introduction

The previous chapter characterized the optimal decisions of utility-maximizing individuals who take asset prices as given. This chapter will focus on the characterization of market equilibrium asset prices. We will work throughout in a one-period model and assume that the state space is finite, $\Omega = \{1, 2, \ldots, S\}$. The results can be generalized to an infinite state space and a multi-period setting.

First, let us define an equilibrium. Consider a one-period economy with I assets and L greedy and risk-averse individuals. Each asset i is characterized by its time 0 price P_i and a random variable D_i representing the time 1 dividend. Each individual is characterized by a (strictly increasing and concave) utility index $\mathcal{U}_l$ and an endowment (e_0^l, e^l). An **equilibrium** for the economy consists of a price vector $\boldsymbol{P}$ and portfolios $\boldsymbol{\theta}^l$, $l = 1, \ldots, L$, satisfying the two conditions

(i) for each $l = 1, \ldots, L$, $\boldsymbol{\theta}^l$ is optimal for individual l given $\boldsymbol{P}$,

(ii) markets clear, i.e. $\sum_{l=1}^{L} \theta_i^l = 0$ for all $i = 1, \ldots, I$.

Associated with an equilibrium $(\boldsymbol{P}; \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^L)$ is an equilibrium consumption allocation (c_0^l, c^l), $l = 1, \ldots, L$, defined by

$$c_0^l = e_0^l - \boldsymbol{\theta}^l \cdot \boldsymbol{P}; \qquad c_\omega^l = e_\omega^l + D_\omega^{\boldsymbol{\theta}^l}, \quad \omega \in \Omega.$$

In the market clearing condition we have assumed that the traded assets are in a net supply of zero and, since the time 0 endowment is a certain number of units of the consumption good, no one owns any assets at time 0. This might seem restrictive but it does cover the case with initial asset holdings and positive net supply of assets. Just interpret $\boldsymbol{\theta}^l$ as individual l's net trade in the assets, i.e. the change in the portfolio relative to the initial portfolio, and interpret the time 1 endowment as the sum of some income from non-financial sources and the dividend from the initial portfolio.

We will assume throughout that individuals have homogeneous beliefs, i.e. that they agree that probabilities of future events are measured by the probability measure $\mathbb{P}$.

Outline of the chapter...

7.2 Pareto-optimality and representative individuals

Define the aggregate initial and future endowment in the economy as

$$\bar{e}_0 = \sum_{l=1}^{L} e_0^l, \quad \bar{e}_\omega = \sum_{l=1}^{L} e_\omega^l.$$

(If we allow for assets in a positive net supply, the dividends of those assets are to be included in the time 1 aggregate endowment.) A consumption allocation $\{(c_0^1, c^1), \ldots, (c_0^L, c^L)\}$ is said to be feasible if

$$\sum_{l=1}^{L} c_0^l \leq \bar{e}_0; \quad \sum_{l=1}^{L} c_\omega^l \leq \bar{e}_\omega, \quad \omega \in \Omega.$$

Here, the left-hand sums define the aggregate consumption,

$$C_0 = \sum_{l=1}^{L} c_0^l, \quad C_\omega = \sum_{l=1}^{L} c_\omega^l.$$

A consumption allocation $\{(c_0^1, \boldsymbol{c}^1), \ldots, (c_0^L, \boldsymbol{c}^L)\}$ is **Pareto-optimal** if it is feasible and there is no other feasible consumption plan $\{(\hat{c}_0^1, \hat{\boldsymbol{c}}^1), \ldots, (\hat{c}_0^L, \hat{\boldsymbol{c}}^L)\}$ such that $\mathcal{U}_l(\hat{c}_0^l, \hat{\boldsymbol{c}}^l) \geq \mathcal{U}_l(c_0^l, \boldsymbol{c}^l)$ for all $l = 1, \ldots, L$ with strict inequality for some l.

Pareto-optimality of consumption allocations is closely linked to the solution to the allocation problem of a hypothetical central planner. Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)^\top$ be a vector of strictly positive numbers, one for each individual. Define the function $\mathcal{U}_{\boldsymbol{\eta}} : \mathbb{R}_+ \times \mathbb{R}_+^S \to \mathbb{R}$ by

$$\mathcal{U}_{\boldsymbol{\eta}}(\bar{e}_0, \bar{\boldsymbol{e}}) = \sup \left\{ \sum_{l=1}^{L} \eta_l \mathcal{U}_l(c_0^l, \boldsymbol{c}^l) \,|\, \sum_{l=1}^{L} c_0^l \leq \bar{e}_0, \sum_{l=1}^{L} c_\omega^l \leq \bar{e}_\omega, c_0^l, c_\omega^l \geq 0, \text{ for all } \omega \text{ and } l \right\}.$$

Here, $\sum_{l=1}^{L} \eta_l \mathcal{U}_l(c_0^l, \boldsymbol{c}^l)$ is a linear combination of the utilities of the individuals when individual l follows the consumption plan $(c_0^l, \boldsymbol{c}^l)$. This linear combination is maximized over all feasible allocations of the total endowment $(\bar{e}_0, \bar{\boldsymbol{e}})$. Given the total endowment and the weights $\boldsymbol{\eta}$ on individuals, $\mathcal{U}_{\boldsymbol{\eta}}$ gives the best linear combination of utilities that can be obtained. As long as the individuals' utility indices are increasing and concave, $\mathcal{U}_{\boldsymbol{\eta}}$ will also be increasing and concave. We can thus think of $\mathcal{U}_{\boldsymbol{\eta}}$ as the utility index of a greedy and risk-averse individual, a central planner giving weights to the individual utility functions. The following theorem, sometimes called the Second Welfare Theorem, gives the link to Pareto-optimality:

Theorem 7.1 *Given any Pareto-optimal consumption allocation with aggregate consumption $(C_0, \boldsymbol{C})$, a strictly positive weighting vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)^\top$ exists so that the same allocation maximizes $\mathcal{U}_{\boldsymbol{\eta}}(C_0, \boldsymbol{C})$.*

Consequently, we can find Pareto-optimal allocations by solving the central planner's problem for a given weighting vector $\boldsymbol{\eta}$. The Lagrangian of the central planner's constrained maximization problem is

$$\mathcal{L} = \sum_{l=1}^{L} \eta_l \mathcal{U}_l(c_0^l, \boldsymbol{c}^l) + \alpha_0 \left(C_0 - \sum_{l=1}^{L} c_0^l \right) + \sum_{\omega=1}^{S} \alpha_\omega \left(C_\omega - \sum_{l=1}^{L} c_\omega^l \right),$$

where $\alpha_0, \alpha_1, \ldots, \alpha_S$ are Lagrange multipliers. The first-order conditions are

$$\frac{\partial \mathcal{L}}{\partial c_0^l} = 0 \quad \Leftrightarrow \quad \eta_l \frac{\partial \mathcal{U}_l}{\partial c_0^l} = \alpha_0, \quad l = 1, \ldots, L, \tag{7.1}$$

$$\frac{\partial \mathcal{L}}{\partial c_\omega^l} = 0 \quad \Leftrightarrow \quad \eta_l \frac{\partial \mathcal{U}_l}{\partial c_\omega^l} = \alpha_\omega, \quad \omega = 1, \ldots, S, \quad l = 1, \ldots, L, \tag{7.2}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_0} = 0 \quad \Leftrightarrow \quad \sum_{l=1}^{L} c_0^l = C_0, \tag{7.3}$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_\omega} = 0 \quad \Leftrightarrow \quad \sum_{l=1}^{L} c_\omega^l = C_\omega, \quad \omega = 1, \ldots, S. \tag{7.4}$$

Note that since we can scale all η_l's by a positive constant without affecting the maximizing consumption allocation, we might as well assume $\alpha_0 = 1$. Given strictly increasing and concave utility functions with infinite marginal utility at zero, the first-order conditions are both necessary and sufficient for optimality. We can thus conclude that a feasible consumption allocation is Pareto-optimal if and only if we can find a weighting vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)^\top$ so that (7.1) and (7.2) are satisfied. In particular, by dividing (7.2) by (7.1), it is clear that we need to have

$$\frac{\frac{\partial \mathcal{U}_l}{\partial c_\omega^l}}{\frac{\partial \mathcal{U}_l}{\partial c_0^l}} = \frac{\alpha_\omega}{\alpha_0}, \quad \omega = 1, \ldots, S, \quad l = 1, \ldots, L, \tag{7.5}$$

with the consequence that

$$\frac{\frac{\partial \mathcal{U}_k}{\partial c_\omega^k}}{\frac{\partial \mathcal{U}_k}{\partial c_0^k}} = \frac{\frac{\partial \mathcal{U}_l}{\partial c_\omega^l}}{\frac{\partial \mathcal{U}_l}{\partial c_0^l}}, \quad \omega = 1, \ldots, S, \tag{7.6}$$

for any individuals k and l, i.e. the individuals align their marginal rates of substitution. This property is often referred to as **efficient risk-sharing**. The central planner will distribute aggregate consumption risk so that all individuals have the same marginal willingness to shift consumption across time and states.

Note that if individuals have time-additive expected utility, i.e.

$$\mathcal{U}_l(c_0, c) = u_l(c_0) + e^{-\delta_l} \operatorname{E}[u_l(c)] = u_l(c_0) + e^{-\delta_l} \sum_{\omega=1}^{S} p_\omega u_l(c_\omega),$$

we can replace (7.1) and (7.2) by

$$\eta_l u_l'(c_0^l) = \alpha_0, \quad l = 1, \ldots, L, \tag{7.7}$$

$$\eta_l e^{-\delta_l} p_\omega u_l'(c_\omega^l) = \alpha_\omega, \quad \omega = 1, \ldots, S, \quad l = 1, \ldots, L. \tag{7.8}$$

Dividing the second equation by the first, we see that a Pareto-optimal consumption allocation has the property that

$$e^{-\delta_k} \frac{u_k'(c_\omega^k)}{u_k'(c_0^k)} = e^{-\delta_l} \frac{u_l'(c_\omega^l)}{u_l'(c_0^l)}, \quad \omega = 1, \ldots, S, \tag{7.9}$$

for any two individuals k and l.

The following theorem shows that all Pareto-optimal consumption allocations have the property that the consumption of individuals move together.

Theorem 7.2 *For any Pareto-optimal consumption allocation, the consumption of any individual will be a strictly increasing function of aggregate consumption, i.e. for any l, $c^l = f_l(C)$ for some strictly increasing function f_l.*

Proof: Given some Pareto-optimal consumption allocation $(c_0^l, \boldsymbol{c}^l)$, $l = 1, \ldots, L$. Assume for simplicity that preferences can be represented by time-additive expected utility. From Theorem 7.1 we know that we can find a weighting vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)^\top$ so that (7.7) and (7.8) hold. In particular, we have

$$\begin{aligned} \eta_k u_k'(c_0^k) &= \eta_l u_l'(c_0^l), \\ \eta_k e^{-\delta_k} u_k'(c_\omega^k) &= \eta_l e^{-\delta_l} u_l'(c_\omega^l), \quad \omega = 1, \ldots, S, \end{aligned}$$

for any two individuals k and l. Moreover, for any two states $\omega, \omega' \in \Omega$, we must have

$$\frac{u_k'(c_\omega^k)}{u_k'(c_{\omega'}^k)} = \frac{u_l'(c_\omega^l)}{u_l'(c_{\omega'}^l)}. \tag{7.10}$$

Aggregate consumption in state ω is $C_\omega = \sum_{l=1}^L c_\omega^l$, where the sum is over all individuals. Suppose that aggregate consumption is higher in state ω than in state ω', i.e. $C_\omega > C_{\omega'}$. Then there must at least one individual, say individual l, who consumes more in state ω than in state ω', $c_\omega^l > c_{\omega'}^l$. Consequently, $u_l'(c_\omega^l) < u_l'(c_{\omega'}^l)$. But then (7.10) implies that $u_k'(c_\omega^k) < u_k'(c_{\omega'}^k)$ and thus $c_\omega^k > c_{\omega'}^k$ for all individuals k. □

A consequence of the previous theorem is that with Pareto-optimal allocations individuals do not have to distinguish between states in which aggregate consumption is the same. Aggregate consumption C at time 1 is a random variable that induces a partition of the state space. Suppose that the possible values of aggregate consumption are $x_1, \ldots, x_K$ and let $\Omega_k = \{\omega \in \Omega | C_\omega = x_k\}$ be the set of states in which aggregate consumption equals x_k. Then $\Omega = \Omega_1 \cup \cdots \cup \Omega_K$. For any individual l, we can define a valid state-price vector by

$$\psi_\omega = e^{-\delta_l} \frac{p_\omega u_l'(c_\omega^l)}{u_l'(c_0^l)} = e^{-\delta_l} \frac{p_\omega u_l'(f_l(C_\omega))}{u_l'(c_0^l)}.$$

Define

$$\begin{aligned} \psi(k) &= \sum_{\omega \in \Omega_k} \psi_\omega = \sum_{\omega \in \Omega_k} e^{-\delta_l} \frac{p_\omega u_l'(f_l(C_\omega))}{u_l'(c_0^l)} \\ &= e^{-\delta_l} \frac{u_l'(f_l(x_k))}{u_l'(c_0^l)} \sum_{\omega \in \Omega_k} p_\omega = e^{-\delta_l} \frac{u_l'(f_l(x_k))}{u_l'(c_0^l)} \mathbb{P}(C = x_k), \end{aligned}$$

which can be interpreted as the value of an asset with a dividend of one if aggregate consumption turns out to be x_k and a zero dividend in other cases. The price of any marketed dividend D can then be written as

$$\begin{aligned} P &= \sum_{\omega=1}^S \psi_\omega D_\omega = \sum_{k=1}^K \sum_{\omega \in \Omega_k} \psi_\omega D_\omega \\ &= \sum_{k=1}^K \sum_{\omega \in \Omega_k} e^{-\delta_l} \frac{p_\omega u_l'(f_l(C_\omega))}{u_l'(c_0^l)} D_\omega = \sum_{k=1}^K e^{-\delta_l} \frac{u_l'(f_l(x_k))}{u_l'(c_0^l)} \sum_{\omega \in \Omega_k} p_\omega D_\omega \\ &= \sum_{k=1}^K \frac{\psi(k)}{\mathbb{P}(C = x_k)} \sum_{\omega \in \Omega_k} p_\omega D_\omega = \sum_{k=1}^K \psi(k) \sum_{\omega \in \Omega_k} \frac{p_\omega}{\mathbb{P}(C = x_k)} D_\omega \\ &= \sum_{k=1}^K \psi(k) \, \mathrm{E}[D | C = x_k], \end{aligned}$$

where $\mathrm{E}[D|C = x_k]$ is the expected dividend conditional on aggregate consumption being x_k. The last equality is due to the fact that $p_\omega/\mathbb{P}(C = x_k)$ is the conditional probability of the state ω given that aggregate consumption is x_k. To price any marketed dividend we thus need only prices of Arrow-Debreu style assets on aggregate consumption and the expectation of the dividend conditional on the aggregate consumption.

The economy is said to have a **representative individual** if for each equilibrium in the economy with L individuals there is a vector $\boldsymbol{\eta}$ such that the equilibrium asset prices are the same in the economy with the single individual with utility $\mathcal{U}_{\boldsymbol{\eta}}$ and endowment $(\bar{e}_0, \bar{\boldsymbol{e}})$. Theorem 7.1 has the following immediate consequence:

Theorem 7.3 *If the equilibrium consumption allocation is Pareto-optimal, the economy has a representative individual.*

Since it is much easier to analyze models with only one individual, many asset pricing models do assume the existence of a representative individual. Of course, it is interesting to know under what conditions this assumption is satisfied, i.e. under what conditions the equilibrium consumption allocation in the underlying multi-individual economy is going to be Pareto-optimal. We will study that question below. Note that in the representative individual economy there can be no trade in the financial assets (who should be the other party in the trade?) and the consumption of the representative individual must equal the total endowment. If you want to model how financial assets are traded, a representative individual formulation is obviously not useful, but if you just want to study equilibrium asset prices it is often convenient.

7.3 Pareto-optimality in complete markets

From the analysis in the previous chapter we know that any individual's marginal rate of substitution is a valid state-price deflator. With time-additive expected utility the state-price deflator induced by individual l is the random variable

$$\zeta^l = e^{-\delta_l} \frac{u_l'(c^l)}{u_l'(c_0^l)},$$

where c_0^l is the optimal time 0 consumption and c^l is the optimal state-dependent time 1 consumption. In general, these state prices may not be identical for different investors so that multiple state-price vectors and deflators can be constructed in this way. However, if the market is complete, we know that the state-price deflator is unique so we must have $\zeta^k = \zeta^l$ for any two individuals k and l. This means that $\zeta_\omega^k = \zeta_\omega^l$ for all possible states $\omega \in \Omega = \{1, 2, \ldots, S\}$. Assuming complete markets and time-additive utility we can conclude that

$$e^{-\delta_k} \frac{u_k'\left(c_\omega^k\right)}{u_k'\left(c_0^k\right)} = e^{-\delta_l} \frac{u_l'\left(c_\omega^l\right)}{u_l'\left(c_0^l\right)} \tag{7.11}$$

for any ω, i.e. the marginal rate of substitution is the same for all individuals. From the discussion above, this will imply efficient risk-sharing and that the consumption allocation is Pareto-optimal. This result is often called the *First Welfare Theorem*:

Theorem 7.4 *If the financial market is complete, then every equilibrium consumption allocation is Pareto-optimal.*

Here is a formal proof:

Proof: Recall from (6.19) that in a complete market the utility maximization problem of individual l can be written as

$$\begin{aligned} \max_{c_0,\boldsymbol{c}} \ & \mathcal{U}_l(c_0,\boldsymbol{c}) \\ \text{s.t. } & c_0+\boldsymbol{\psi}\cdot\boldsymbol{c}\le e_0^l+\boldsymbol{\psi}\cdot\boldsymbol{e}^l \\ & c_0,\boldsymbol{c}\ge 0, \end{aligned}$$

where $\boldsymbol{\psi}$ is the unique state-price vector. Let $\{(c_0^l,\boldsymbol{c}^l), l=1,\dots,L\}$ be an equilibrium consumption allocation, but suppose we can find another consumption allocation $\{(\hat{c}_0^l,\hat{\boldsymbol{c}}^l), l=1,\dots,L\}$ which gives all individuals at least the same utility and some individuals strictly higher utility than $\{(c_0^l,\boldsymbol{c}^l), l=1,\dots,L\}$. Since we assume strictly increasing utility and $(c_0^l,\boldsymbol{c}^l)$ maximizes individual l's utility subject to the constraint $c_0^l+\boldsymbol{\psi}\cdot\boldsymbol{c}^l\le e_0^l+\boldsymbol{\psi}\cdot\boldsymbol{e}^l$, the inequality

$$\hat{c}_0^l+\boldsymbol{\psi}\cdot\hat{\boldsymbol{c}}^l\ge e_0^l+\boldsymbol{\psi}\cdot\boldsymbol{e}^l$$

must hold for all individuals with strict inequality for at least one individual. Summing up over all individuals we get

$$\sum_{l=1}^{L}\left(\hat{c}_0^l+\boldsymbol{\psi}\cdot\hat{\boldsymbol{c}}^l\right)>\sum_{l=1}^{L}\left(e_0^l+\boldsymbol{\psi}\cdot\boldsymbol{e}^l\right)=\bar{e}_0+\boldsymbol{\psi}\cdot\bar{\boldsymbol{e}}.$$

Hence, the consumption allocation $\{(\hat{c}_0^l,\hat{\boldsymbol{c}}^l), l=1,\dots,L\}$ is not feasible. □

A complete market equilibrium provides efficient risk-sharing. From (7.11) it follows that, for any two states ω and ω', we have

$$\frac{u_k'\left(c_\omega^k\right)}{u_k'\left(c_{\omega'}^k\right)}=\frac{u_l'\left(c_\omega^l\right)}{u_l'\left(c_{\omega'}^l\right)}. \tag{7.12}$$

Suppose that (7.12) did not hold. Suppose we could find two individuals k and l and two states ω and ω' such that

$$\frac{u_k'(c_\omega^k)}{u_k'(c_{\omega'}^k)}>\frac{u_l'(c_\omega^l)}{u_l'(c_{\omega'}^l)}. \tag{7.13}$$

Then the two agents could engage in a trade that makes both better off. What trade? Since the market is complete, Arrow-Debreu assets for all states are traded, in particular for states ω and ω'. Consider the following trade: Individual k buys ε_ω Arrow-Debreu assets for state ω from individual l at a unit price of φ_ω. And individual l buys $\varepsilon_{\omega'}$ Arrow-Debreu assets for state ω' from individual k at a unit price of $\varphi_{\omega'}$. The deal is arranged so that the net price is zero, i.e.

$$\varepsilon_\omega\varphi_\omega-\varepsilon_{\omega'}\varphi_{\omega'}=0\quad\Leftrightarrow\varepsilon_{\omega'}=\varepsilon_\omega\frac{\varphi_{\omega'}}{\varphi_\omega}.$$

The deal will change the consumption of the two individuals in states ω and ω' but not in other states, nor at time 0. The total change in the expected utility of individual k will be

$$p_\omega\left(u_k(c_\omega^k+\varepsilon_\omega)-u_k(c_\omega^k)\right)+p_{\omega'}\left(u_k(c_{\omega'}^k-\varepsilon_\omega\frac{\varphi_{\omega'}}{\varphi_\omega})-u_k(c_{\omega'}^k)\right).$$

Dividing by ε_ω and letting $\varepsilon_\omega\to 0$, the additional expected utility approaches

$$p_\omega u_k'(c_\omega^k)-\frac{\varphi_{\omega'}}{\varphi_\omega}u_k'(c_{\omega'}^k),$$

which is strictly positive whenever

$$\frac{u_k'(c_\omega^k)}{u_k'(c_{\omega'}^k)} > \frac{p_{\omega'}\varphi_{\omega'}}{p_\omega\varphi_\omega}. \tag{7.14}$$

On the other hand, the total change in the expected utility of individual l will be

$$p_\omega\left(u_l(c_\omega^l - \varepsilon_\omega) - u_l(c_\omega^l)\right) + p_{\omega'}\left(u_l(c_{\omega'}^l + \varepsilon_\omega\frac{\varphi_{\omega'}}{\varphi_\omega}) - u_l(c_{\omega'}^l)\right).$$

Dividing by ε_ω and letting $\varepsilon_\omega \to 0$, we get

$$-p_\omega u_l'(c_\omega^l) + p_{\omega'}\frac{\varphi_{\omega'}}{\varphi_\omega}u_l'(c_{\omega'}^l),$$

which is strictly positive whenever

$$\frac{u_l'(c_\omega^l)}{u_l'(c_{\omega'}^l)} < \frac{p_{\omega'}\varphi_{\omega'}}{p_\omega\varphi_\omega}. \tag{7.15}$$

If the inequality (7.13) holds, the two individuals can surely find prices φ_ω and $\varphi_{\omega'}$ so that both (7.14) and (7.15) are satisfied, i.e. both increase their expected utility.

If the market is incomplete, the individuals might not be able to implement this trade. In other words, we cannot be sure that (7.12) holds for states for which Arrow-Debreu assets are not traded, i.e. "uninsurable" states.

Combining theorems stated above, we have the following conclusion:

Theorem 7.5 *If the financial market is complete, the economy has a representative individual.*

If we want to study asset pricing in a complete market, we might as well assume that the economy has a single individual. But what is the appropriate weighting vector $\boldsymbol{\eta}$ for the complete market? We must ensure that the first-order conditions of the central planner are satisfied when we plug in the optimal consumption plans of the individuals. Recall that in a complete market the utility maximization problem of individual l can be formulated as in (6.19). The Lagrangian for this problem is

$$\mathcal{L}_l = \mathcal{U}_l(c_0^l, \boldsymbol{c}^l) + \kappa_l\left(\sum_{\omega=1}^S \psi_\omega\left(e_\omega^l - c_\omega^l\right) + e_0^l - c_0^l\right).$$

The first-order conditions, again necessary and sufficient, are

$$\frac{\partial\mathcal{L}_l}{\partial c_0^l} = 0 \quad\Leftrightarrow\quad \frac{\partial\mathcal{U}_l}{\partial c_0^l} = \kappa_l, \tag{7.16}$$

$$\frac{\partial\mathcal{L}_l}{\partial c_\omega^l} = 0 \quad\Leftrightarrow\quad \frac{\partial\mathcal{U}_l}{\partial c_\omega^l} = \kappa_l\psi_\omega, \quad \omega = 1,\ldots,S, \tag{7.17}$$

$$\frac{\partial\mathcal{L}_l}{\partial \kappa_l} = 0 \quad\Leftrightarrow\quad c_0^l + \sum_{\omega=1}^S \psi_\omega c_\omega^l = e_0^l + \sum_{\omega=1}^S \psi_\omega e_\omega^l. \tag{7.18}$$

If we set $\varphi_0 = 1$ and $\eta_l = 1/\kappa_l$ for each $l = 1,\ldots,L$, the Equations (7.1)–(7.4) will indeed be satisfied.

Note that the weight η_l associated with individual l is the inverse of his "shadow price" of his budget constraint and η_l will therefore depend on the initial endowment of individual l. Redistributing the aggregate initial endowment across individuals will thus change the relative values of the weights η_l and, hence, the utility function of the representative individual and equilibrium asset prices.

Let us briefly consider the special case where all individuals have time-additive expected utilities. Then the representative individual will also have time-additive expected utility, i.e.

$$\mathcal{U}_{\boldsymbol{\eta}}(c_0, \boldsymbol{c}) = u_{\boldsymbol{\eta},0}(c_0) + \mathrm{E}\left[u_{\boldsymbol{\eta},1}(c)\right] = u_{\boldsymbol{\eta},0}(c_0) + \sum_{\omega=1}^{S} p_\omega u_{\boldsymbol{\eta},1}(c_\omega),$$

where

$$u_{\boldsymbol{\eta},0}(c_0) = \sup\left\{\sum_{l=1}^{L} \eta_l u_l(c_0^l) \,|\, c_0^1, \ldots, c_0^L > 0 \text{ with } c_0^1 + \cdots + c_0^L \leq c_0\right\}, \tag{7.19}$$

$$u_{\boldsymbol{\eta},1}(c) = \sup\left\{\sum_{l=1}^{L} e^{-\delta_l} \eta_l u_l(c^l) \,|\, c^1, \ldots, c^L > 0 \text{ with } c^1 + \cdots + c^L \leq c\right\}. \tag{7.20}$$

Often a functional form for $u_{\boldsymbol{\eta},0}$ and $u_{\boldsymbol{\eta},1}$ is assumed directly with the property that $u_{\boldsymbol{\eta},1}(c) = e^{-\delta} u_{\boldsymbol{\eta},0}(c) = e^{-\delta} u(c)$, where δ can be interpreted as the average time preference rate in the economy. Then the unique state-price deflator follows by evaluating the derivative of u_η at the aggregate endowment,

$$\zeta_\omega = \frac{u'_{\boldsymbol{\eta},1}(\bar{e}_\omega)}{u'_{\boldsymbol{\eta},0}(\bar{e}_0)} = e^{-\delta} \frac{u'(\bar{e}_\omega)}{u'(\bar{e}_0)}.$$

This will be very useful in order to link asset prices and interest rates to aggregate consumption.

7.4 Pareto-optimality in some incomplete markets

In the previous section we saw that complete market equilibria are Pareto-optimal. However, as discussed earlier, real-life financial markets are probably not complete. Equilibrium consumption allocations in incomplete markets will generally not be Pareto-optimal since the individuals cannot necessarily implement the trades needed to align their marginal rates of substitution. On the other hand, individuals do not have to be able to implement any possible consumption plan so we do not need markets to be complete in the strict sense. If every Pareto-optimal consumption allocation can be obtained by trading in the available assets, the market is said to be **effectively complete**. If the market is effectively complete, we can use the representative individual approach to asset pricing. In this section we will discuss some examples of effectively complete markets.

For any Pareto-optimal consumption allocation we know from Theorem 7.2 that the consumption of any individual is an increasing function of aggregate consumption. Individual consumption is measurable with respect to aggregate consumption. As in the discussion below Theorem 7.2, suppose that the possible values of aggregate consumption are $x_1, \ldots, x_K$ and let $\Omega_k = \{\omega \in \Omega | C_\omega = x_k\}$ be the set of states in which aggregate consumption equals x_k. If it is possible for each k to form a portfolio that provides a payment of 1 if the state is in Ω_k and a zero payment otherwise—an Arrow-Debreu style asset for aggregate consumption—then the market is effectively complete. (See also Exercise 7.3.) The individuals are indifferent between states within a given subset Ω_k. Risk beyond aggregate consumption risk does not carry any premium. Assuming strictly increasing utility functions, aggregate time 1 consumption will equal aggregate time 1 endowment. If we think of the aggregate endowment as the total value of the market, aggregate consumption will equal the value of the market portfolio and we can partition the state space according to the value or return of the market portfolio. Risk beyond market risk is diversified away and does not carry any premium.

If there is a full set of Arrow-Debreu style assets for aggregate consumption, markets will be effectively complete for all strictly increasing and concave utility functions. The next theorem, which is due to Rubinstein (1974), shows that markets are effectively complete under weaker assumptions on the available assets if individuals have utility functions of the HARA class with identical risk cautiousness. Before stating the precise result, let us review a few facts about the HARA utility functions which were defined in Section 5.6. A utility function u is of the HARA class, if the absolute risk tolerance is affine, i.e.

$$\mathrm{ART}(c) \equiv -\frac{u'(c)}{u''(c)} = \alpha c + \beta.$$

The risk cautiousness is $\mathrm{ART}'(c) = \alpha$. Ignoring insignificant constants, the marginal utility must be either

$$u'(c) = e^{-c/\beta},$$

for the case $\alpha = 0$ (corresponding to negative exponential utility), or

$$u'(c) = (\alpha c + \beta)^{-1/\alpha},$$

for the case $\alpha \neq 0$ (which encompasses extended log-utility, satiation HARA utility, and subsistence HARA utility).

Theorem 7.6 *Suppose that*

(i) all individuals have time-additive HARA utility functions with identical risk cautiousness;

(ii) a risk-free asset is traded;

(iii) the time 1 endowment of all individuals are spanned by traded assets, i.e. $e^l \in \mathcal{M}$, $l = 1, \ldots, L$.

Then the equilibrium is Pareto-optimal and the economy has a representative individual. The optimal consumption for any individual is a strictly increasing affine function of aggregate consumption.

Proof: Suppose first that the market is complete. Then we know that the equilibrium consumption allocation will be Pareto-optimal and we can find a weighting vector $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_L)^\top$ so that

$$\eta_k e^{-\delta_k} u_k'(c_\omega^k) = \eta_l e^{-\delta_l} u_l'(c_\omega^l) \tag{7.21}$$

for any two individuals k and l and any ω. Assume that the common risk cautiousness α is different from zero so that

$$u_l'(c) = (\alpha c + \beta_l)^{-1/\alpha}, \quad l = 1, \ldots, L.$$

(The proof for the case $\alpha = 0$ is similar.) Substituting this into the previous equation, we obtain

$$\eta_k e^{-\delta_k} \left(\alpha c_\omega^k + \beta_k\right)^{-1/\alpha} = \eta_l e^{-\delta_l} \left(\alpha c_\omega^l + \beta_l\right)^{-1/\alpha},$$

which implies that

$$\left(\eta_k e^{-\delta_k}\right)^{-\alpha} \left(\alpha c_\omega^k + \beta_k\right) \left(\eta_l e^{-\delta_l}\right)^{\alpha} = \alpha c_\omega^l + \beta_l.$$

Summing up over $l = 1, \ldots, L$, we get

$$\left(\eta_k e^{-\delta_k}\right)^{-\alpha} \left(\alpha c_\omega^k + \beta_k\right) \sum_{l=1}^{L} \left(\eta_l e^{-\delta_l}\right)^{\alpha} = \alpha \sum_{l=1}^{L} c_\omega^l + \sum_{l=1}^{L} \beta_l = \alpha C_\omega + \sum_{l=1}^{L} \beta_l,$$

where C_ω is aggregate consumption (or endowment) in state ω. Solving for c_ω^k, we find that

$$c_\omega^k = \frac{\alpha C_\omega + \sum_{l=1}^{L} \beta_l}{\alpha \left(\eta_k e^{-\delta_k}\right)^{-\alpha} \sum_{l=1}^{L} \left(\eta_l e^{-\delta_l}\right)^{\alpha}} - \frac{\beta_k}{\alpha} \equiv A_k C_\omega + B_k,$$

which is strictly increasing and affine in C_ω.

The same consumption allocation can be obtained in a market where a risk-free asset exists and the time 1 endowments of all individuals are spanned by traded assets. □

Under the additional assumption that the time preference rates of individuals are identical, $\delta_l = \delta$ for all $l = 1, \ldots, L$, we can show a bit more:

Theorem 7.7 *If the assumptions of Theorem 7.6 are satisfied and individuals have identical time preference rates, then the relative weights which the representative individual associates to individuals—and therefore equilibrium asset prices—will be independent of the initial distribution of aggregate endowment across individuals.*

Proof: In order to verify this, we will compute the utility function of the representative individual. Since we are assuming time-additive utility, we will derive $u_{\boldsymbol{\eta},0}$ and $u_{\boldsymbol{\eta},1}$ defined in (7.19) and (7.20) for any given weighting vector $\boldsymbol{\eta}$. We will focus on the case where the common risk cautiousness α is non-zero and different from 1 so that

$$u_l(c) = \frac{1}{1 - 1/\alpha} \left(\alpha c + \beta_l\right)^{1 - 1/\alpha}, \quad u_l'(c) = \left(\alpha c + \beta_l\right)^{-1/\alpha}, \quad l = 1, \ldots, L.$$

(In Exercise 7.2 you are asked to do the same for the cases of extended log-utility ($\alpha = 1$) and negative exponential utility ($\alpha = 0$).) The first-order condition for the maximization in the definition of $u_{\boldsymbol{\eta},0}$ implies that

$$\eta_l \left(\alpha c + \beta_l\right)^{-1/\alpha} = \nu, \tag{7.22}$$

where ν is the Lagrange multiplier. Rearranging, we get

$$\alpha c_0^l + \beta_l = \nu^{-\alpha} \eta_l^{\alpha}.$$

Summing up over l gives

$$\alpha c_0 + \sum_{l=1}^{L} \beta_l = \nu^{-\alpha} \sum_{l=1}^{L} \eta_l^{\alpha},$$

so that the Lagrange multiplier is

$$\nu = \left(\alpha c_0 + \sum_{l=1}^{L} \beta_l\right)^{-1/\alpha} \left(\sum_{l=1}^{L} \eta_l^{\alpha}\right)^{1/\alpha}.$$

Substituting this back into (7.22), we see that the solution to the maximization problem is such that

$$\eta_l \left(\alpha c + \beta_l\right)^{-1/\alpha} = \left(\alpha c_0 + \sum_{l=1}^{L} \beta_l\right)^{-1/\alpha} \left(\sum_{l=1}^{L} \eta_l^{\alpha}\right)^{1/\alpha},$$

which implies that

$$\begin{aligned} u_{\boldsymbol{\eta},0}(c_0) &= \sum_{l=1}^{L} \eta_l \frac{1}{1-1/\alpha} \left(\alpha c_0^l + \beta_l\right)^{1-1/\alpha} \\ &= \frac{1}{1-1/\alpha} \sum_{l=1}^{L} \left(\alpha c_0^l + \beta_l\right) \eta_l \left(\alpha c_0^l + \beta_l\right)^{-1/\alpha} \\ &= \frac{1}{1-1/\alpha} \left(\alpha c_0 + \sum_{l=1}^{L} \beta_l\right)^{-1/\alpha} \left(\sum_{l=1}^{L} \eta_l^{\alpha}\right)^{1/\alpha} \sum_{l=1}^{L} \left(\alpha c_0^l + \beta_l\right) \\ &= \frac{1}{1-1/\alpha} \left(\sum_{l=1}^{L} \eta_l^{\alpha}\right)^{1/\alpha} \left(\alpha c_0 + \sum_{l=1}^{L} \beta_l\right)^{1-1/\alpha}. \end{aligned} \tag{7.23}$$

Almost identical computations lead to

$$u_{\boldsymbol{\eta},1}(c) = e^{-\delta} \frac{1}{1-1/\alpha} \left(\sum_{l=1}^{L} \eta_l^{\alpha}\right)^{1/\alpha} \left(\alpha c + \sum_{l=1}^{L} \beta_l\right)^{1-1/\alpha} \tag{7.24}$$

The utility function of the representative individual is therefore of the same class as the individual utility functions,

$$\mathcal{U}_{\boldsymbol{\eta}}(c_0, \boldsymbol{c}) = u_{\boldsymbol{\eta},0}(c_0) + \mathrm{E}\left[u_{\boldsymbol{\eta},1}(c)\right] = u_{\boldsymbol{\eta}}(c_0) + e^{-\delta}\,\mathrm{E}[u_{\boldsymbol{\eta}}(c)],$$

where

$$u_{\boldsymbol{\eta}}(c) = \frac{1}{1-1/\alpha} \left(\sum_{l=1}^{L} \eta_l^{\alpha}\right)^{1/\alpha} \left(\alpha c + \sum_{l=1}^{L} \beta_l\right)^{1-1/\alpha}.$$

More importantly, the state-price deflator is

$$\begin{aligned} \zeta = e^{-\delta} \frac{u'_{\boldsymbol{\eta}}(c)}{u'_{\boldsymbol{\eta}}(c_0)} &= e^{-\delta} \frac{\left(\sum_{l=1}^{L} \eta_l^{\alpha}\right)^{1/\alpha} \left(\alpha c + \sum_{l=1}^{L} \beta_l\right)^{-1/\alpha}}{\left(\sum_{l=1}^{L} \eta_l^{\alpha}\right)^{1/\alpha} \left(\alpha c_0 + \sum_{l=1}^{L} \beta_l\right)^{-1/\alpha}} \\ &= e^{-\delta} \frac{\left(\alpha c + \sum_{l=1}^{L} \beta_l\right)^{-1/\alpha}}{\left(\alpha c_0 + \sum_{l=1}^{L} \beta_l\right)^{-1/\alpha}}, \end{aligned} \tag{7.25}$$

which is independent of the weighting vector $\boldsymbol{\eta}$. The state-price deflator and, hence, the asset prices are independent of the distribution of endowment across individuals. □

7.5 Exercises

EXERCISE 7.1 Suppose η_1 and η_2 are strictly positive numbers and that $u_1(c) = u_2(c) = c^{1-\gamma}/(1-\gamma)$ for any non-negative real number c. Define the function $u_\eta : \mathbb{R}_+ \to \mathbb{R}$ by

$$u_\eta(x) = \sup\left\{\eta_1 u_1(y_1) + \eta_2 u_2(y_2) \,|\, y_1 + y_2 \leq x; y_1, y_2 \geq 0\right\}.$$

Show that $u_\eta(x) = kx^{1-\gamma}/(1-\gamma)$ for some constant k. What is the implication for the utility of representative individuals?

EXERCISE 7.2 Show Theorem 7.7 for the case of extended log-utility and the case of negative exponential utility.

EXERCISE 7.3 Suppose that aggregate time 1 consumption can only take on the values $1, 2, \ldots, K$ for some finite integer K. Assume that European call options on aggregate consumption are traded for any exercise price $0, 1, 2, \ldots, K$. Consider a portfolio with one unit of the option with exercise price $k-1$, one unit of the option with exercise price $k+1$, and minus two units of the option with exercise price k. What is the payoff of this portfolio? Discuss the consequences of your findings for the (effective) completeness of the market. Could you do just as well with put options?

EXERCISE 7.4 Assume a discrete-time economy with L agents. Each agent l maximizes time-additive expected utility $\mathrm{E}\left[\sum_{t=0}^{T} \beta_l^t u_l(c_{lt})\right]$ where u_l is strictly increasing and concave. Show that

$$\frac{\zeta_{t+1}}{\zeta_t} = \frac{\sum_{l=1}^{L} \beta_l u_l'(c_{l,t+1})}{\sum_{l=1}^{L} u_l'(c_{lt})}$$

is a valid one-period state-price deflator, i.e. that it is strictly positive and satisfies $\mathrm{E}_t[(\zeta_{t+1}/\zeta_t)R_{t+1}] = 1$ for any gross return R_{t+1} over the period $[t, t+1]$.

EXERCISE 7.5 George and John live in a continuous-time economy in which the relevant uncertainty is generated by a one-dimensional standard Brownian motion $z = (z_t)_{t\in[0,T]}$. Both have time-additive utility of their consumption process: George maximizes

$$U_G(c) = \mathrm{E}\left[\int_0^T e^{-0.02t} \ln c_t \, dt\right],$$

while John maximizes

$$U_J(c) = \mathrm{E}\left[\int_0^T e^{-0.02t} \left(-\frac{1}{c_t}\right) dt\right].$$

George's optimal consumption process $c_G = (c_{Gt})$ has a constant expected growth rate of 4% and a constant volatility of 5%, i.e.

$$dc_{Gt} = c_{Gt}\left[0.04\, dt + 0.05\, dz_t\right].$$

Two assets are traded in the economy. One asset is an instantaneously risk-free bank account with continuously compounded rate of return r_t. The other asset is a risky asset with price process $P = (P_t)$ satisfying

$$dP_t = P_t\left[\mu_{Pt}\, dt + 0.4\, dz_t\right]$$

for some drift μ_{Pt}. The market is complete.

(a) What are the relative risk aversions of George and John, respectively?

(b) Using the fact that a state price deflator can be derived from George's consumption process, determine the risk-free rate r_t and the market price of risk λ_t. What can you conclude about the price processes of the two assets?

(c) Find the drift and volatility of John's optimal consumption process, $c_J = (c_{Jt})$.

(d) Suppose George and John are the only two individuals in the economy. What can you say about the dynamics of aggregate consumption? Can the representative agent have constant relative risk aversion?

EXERCISE 7.6 Consider a discrete-time economy with L individuals with identical preferences so that agent $l = 1, \ldots, L$ at time 0 wants to maximize

$$\mathrm{E}\left[\sum_{t=0}^{T} \beta^t \frac{1}{1-\gamma} c_{l,t}^{1-\gamma}\right]$$

where $c_{l,t}$ denotes the consumption rate of individual l at time t. Let $c_{l,t}^*$ be the optimal consumption rate of individual l at time t.

(a) Argue why

$$\frac{\zeta_{t+1}}{\zeta_t} = \frac{\beta}{L} \sum_{l=1}^{L} \left(\frac{c_{l,t+1}^*}{c_{l,t}^*}\right)^{-\gamma} \qquad (*)$$

is a state-price deflator between time t and time $t+1$.

(b) If the market is complete, explain why the next-period state-price deflator in (*) can be written as

$$\frac{\zeta_{t+1}}{\zeta_t} = \beta \left(\frac{\sum_{l=1}^{L} c_{l,t+1}^*}{\sum_{l=1}^{L} c_{l,t}^*}\right)^{-\gamma}.$$

EXERCISE 7.7 (Use spreadsheet or similar computational tool.) Consider a one-period economy with 5 possible states and 5 assets traded. The state-contingent dividends and prices of the assets and the state probabilities are as follows:

	state 1	state 2	state 3	state 4	state 5	price
Asset 1	1	1	1	1	1	0.9
Asset 2	0	2	4	6	8	1.7
Asset 3	4	0	2	4	2	2.3
Asset 4	10	0	0	2	2	4.3
Asset 5	4	4	0	4	4	2.8
probability	0.25	0.25	0.2	0.2	0.1	

(a) Verify that the market is complete and find the unique state-price deflator.

Consider an individual investor, Alex, with access to the above financial market with the given prices. Suppose Alex has time-additive expected utility with a time preference rate of $\delta = 0.03$ and a constant relative risk aversion of $\gamma = 2$. Suppose his optimal consumption at time 0 is 5 and that he will receive an income of 5 at time 1 no matter which state is realized.

(b) What is Alex' optimal time 1 consumption? What does it cost him to finance that consumption? What is the optimal portfolio for Alex?

Suppose now that there is only one other individual, Bob, in the economy. Bob also has time-additive expected utility with a time preference rate of 0.03 but a relative risk aversion of 5. Bob's optimal time 0 consumption is also 5.

(c) What is Bob's optimal time 1 consumption?

(d) What is the aggregate time 0 consumption and the state-dependent aggregate time 1 consumption?

(e) What is Bob's time 0 endowment and state-dependent time 1 endowment? What is Bob's optimal portfolio?

(f) Verify that the markets clear.

EXERCISE 7.8 Bruce lives in a continuous-time complete market economy. He has time-additive logarithmic utility, $u_B(c) = \ln c$, with a time preference rate of $\delta_B = 0.02$, and his optimal consumption process $c_B = (c_{Bt})$ has dynamics

$$dc_{Bt} = c_{Bt}\left[0.03\,dt + 0.1\,dz_t\right],$$

where $z = (z_t)$ is a standard Brownian motion.

(a) Characterize the state-price deflator induced by Bruce's optimal consumption process? Identify the continuously compounded short-term risk-free interest rate and the instantaneous Sharpe ratio of any risky asset.

Patti lives in the same economy as Bruce. She has time-additive expected utility with a HARA utility function $u_P(c) = \frac{1}{1-\gamma}(c-\bar{c})^{1-\gamma}$ and a time preference rate identical to Bruce's, i.e. $\delta_P = 0.02$.

(b) Explain why Patti's optimal consumption strategy $c_P = (c_{Pt})$ must satisfy

$$c_{Pt} = \bar{c} + \left(\frac{c_{Bt}}{c_{B0}}\right)^{1/\gamma}(c_{P0} - \bar{c})\,.$$

Find the dynamics of Patti's optimal consumption process.

Chapter 8

Consumption-based asset pricing

8.1 Introduction

Previous chapters have shown how state-price deflators determine asset prices and how the optimal consumption choices of individuals determine state-price deflators. In this chapter we combine these observations and link asset prices to consumption. Models linking expected returns to the covariance between return and (aggregate) consumption are typically called Consumption-based Capital Asset Pricing Models (CCAPM). Such models date back to Rubinstein (1976) and Breeden (1979).

The outline of the chapter is as follows. Section 8.2 develops the CCAPM in the one-period framework. Section 8.3 derives a general link between asset prices and consumption in a multi-period setting, while Section 8.4 focuses on a simple and tractable specification. Section 8.5 shows that this simple specification is unable to match important empirical features of aggregate consumption and return, leaving a number of apparent asset pricing puzzles. Section 8.6 discusses some problems with such empirical studies. Some extensions of the simple model are presented in Sections 8.7 and 8.8. These extensions help resolve some of the apparent puzzles.

8.2 The one-period CCAPM

For simplicity let us first investigate the link between asset prices and consumption in the one-period framework.

As discussed several times the marginal rate of substitution of any individual defines a state-price deflator. If we assume time-additive expected utility with time preference rate δ and utility function u, this state-price deflator is

$$\zeta = e^{-\delta} \frac{u'(c)}{u'(c_0)},$$

where c_0 is optimal time 0 consumption and c is the state-dependent optimal time 1 consumption. If the economy can be modeled by a representative individual, the equation holds for aggregate consumption.

Assuming that a risk-free asset is traded, we know from Section 4.2.1 that the gross risk-free

return is

$$R^f = \frac{1}{\mathrm{E}[\zeta]} = e^{\delta} \frac{1}{\mathrm{E}[u'(c)/u'(c_0)]} \tag{8.1}$$

and the expected gross return on any asset i is

$$\begin{aligned} \mathrm{E}[R_i] &= \frac{1}{\mathrm{E}[\zeta]} - \frac{\mathrm{Cov}[R_i, \zeta]}{\mathrm{E}[\zeta]} \\ &= R^f - \frac{\mathrm{Cov}[u'(c)/u'(c_0), R_i]}{\mathrm{E}[u'(c)/u'(c_0)]} \\ &= R^f - \frac{\sigma[u'(c)/u'(c_0)]}{\mathrm{E}[u'(c)/u'(c_0)]} \sigma[R_i] \rho\left[R_i, u'(c)/u'(c_0)\right], \end{aligned} \tag{8.2}$$

where $\sigma[x]$ denotes the standard deviation of the random variable x, while $\rho[x, y]$ is the correlation between the random variables x and y. An asset with a return which is positively correlated with the marginal utility of consumption (and hence negatively correlated with the level of consumption) is attractive, has a high equilibrium price, and thus a low expected return.

In order to obtain a relation between expected returns and consumption itself (rather than marginal utility of consumption) we need to make further assumptions or approximations. Below we develop three versions.

8.2.1 The simple one-period CCAPM: version 1

A first-order Taylor approximation of $u'(c)$ around c_0 gives that

$$\frac{u'(c)}{u'(c_0)} \approx \frac{u'(c_0) + u''(c_0)(c - c_0)}{u'(c_0)} = 1 - \gamma(c_0) g,$$

where $\gamma(c_0) = -c_0 u''(c_0)/u'(c_0)$ is the relative risk aversion of the individual evaluated at the time 0 consumption level, and $g = c/c_0 - 1$ is the (state-dependent) relative growth rate of consumption over the period. If we further assume that $\mathrm{E}[u'(c)/u'(c_0)] \approx 1$, we get that

$$\mathrm{E}[R_i] - R^f \approx \gamma(c_0) \,\mathrm{Cov}[g, R_i] = \gamma(c_0) \rho[g, R_i] \sigma[g] \sigma[R_i], \tag{8.3}$$

where $\rho[g, R_i]$ is the correlation between consumption growth and the return and $\sigma[g]$ and $\sigma[R_i]$ are the standard deviations of consumption growth and return, respectively. The above equation links expected excess returns to covariance with consumption growth. We can rewrite the equation as

$$\mathrm{E}[R_i] \approx R^f + \beta[R_i, g] \eta,$$

where $\beta[R_i, g] = \mathrm{Cov}[g, R_i]/\,\mathrm{Var}[g]$ and $\eta = \gamma(c_0)\,\mathrm{Var}[g]$. If we ignore the approximative character of the equation, it shows that the growth rate of the individual's optimal consumption is a pricing factor.

In particular, if there exists a portfolio with a return R^c which is perfectly correlated with consumption growth, i.e. $R^c = a + bg$ for constants a and b, then

$$\mathrm{E}[R^c] - R^f \approx \gamma(c_0) \,\mathrm{Cov}[g, R^c] = \gamma(c_0) b \,\mathrm{Var}[g] \quad \Rightarrow \quad \gamma(c_0) \approx \frac{1}{b \,\mathrm{Var}[g]} \left(\mathrm{E}[R^c] - R^f\right)$$

and substituting this into (8.3) we get

$$\begin{aligned} \mathrm{E}[R_i] - R^f &\approx \frac{1}{b \,\mathrm{Var}[g]} \left(\mathrm{E}[R^c] - R^f\right) \mathrm{Cov}[g, R_i] \\ &= \frac{1}{b} \beta[R_i, g] \left(\mathrm{E}[R^c] - R^f\right) = \beta[R_i, R^c] \left(\mathrm{E}[R^c] - R^f\right). \end{aligned} \tag{8.4}$$

Here the last equality follows from $\mathrm{Cov}[R^c, R_i] = b\,\mathrm{Cov}[g, R_i]$ and $\mathrm{Var}[R^c] = b^2\,\mathrm{Var}[g]$. Equation (8.4) is exactly as the classic CAPM but with the return on a consumption-mimicking portfolio instead of the return on the market portfolio.

If the economy has a representative individual we can use the above relations with aggregate consumption instead of individual consumption. The factor risk premium η should then reflect the relative risk aversion of the representative individual.

What if the economy does not have a representative individual? Eq. (8.3) will still hold for any individual l, i.e.

$$\mathrm{E}[R_i] - R^f \approx \gamma_l(c_0^l)\,\mathrm{Cov}[c^l/c_0^l, R_i] = \mathrm{ARA}_l(c_0^l)\,\mathrm{Cov}[c^l, R_i],$$

where γ_l and ARA_l are the relative and absolute risk aversion of individual l, so that

$$\big(\mathrm{E}[R_i] - R^f\big)\,\frac{1}{\mathrm{ARA}_l(c_0^l)} \approx \mathrm{Cov}[c^l, R_i], \quad l = 1, 2, \ldots, L.$$

Summing up over all individuals, we get

$$\big(\mathrm{E}[R_i] - R^f\big)\sum_{l=1}^{L}\frac{1}{\mathrm{ARA}_l(c_0^l)} \approx \sum_{l=1}^{L}\mathrm{Cov}[c^l, R_i] = \mathrm{Cov}\left[\sum_{l=1}^{L} c^l, R_i\right] = \mathrm{Cov}[C, R_i],$$

where C is aggregate time 1 consumption. If we let C_0 denote aggregate time 0 consumption, we obtain

$$\mathrm{E}[R_i] - R^f \approx \left(\sum_{l=1}^{L}\frac{1}{\mathrm{ARA}_l(c_0^l)}\right)^{-1} C_0\,\mathrm{Cov}[C/C_0, R_i]. \tag{8.5}$$

Again ignoring the approximation, this equation shows that aggregate consumption growth is a pricing factor even if there is no representative individual. We just need to replace the relative risk aversion of the representative individual by some complex average of individual risk aversions. The absolute risk tolerance of individual l is exactly $1/\,\mathrm{ARA}_l(c_0^l)$ so the first term on the right-hand side of the above equation can be interpreted as the inverse of the aggregate absolute risk tolerance in the economy. The higher the aggregate absolute risk tolerance, the lower the equilibrium risk premium on risky assets.

Of course, we prefer exact asset pricing results to approximate. In a later section, we will show that in continuous-time models the analogues to the above relations will hold as exact equalities under appropriate assumptions.

8.2.2 The simple one-period CCAPM: version 2

Suppose that

(i) the individual has constant relative risk aversion, $u(c) = c^{1-\gamma}/(1-\gamma)$,

(ii) consumption growth is lognormally distributed,

$$\ln(1+g) \equiv \ln\left(\frac{c}{c_0}\right) \sim N(\bar{g}, \sigma_g^2).$$

Then

$$\frac{u'(c)}{u'(c_0)} = \left(\frac{c}{c_0}\right)^{-\gamma} = \exp\left\{-\gamma \ln\left(\frac{c}{c_0}\right)\right\}.$$

For a random variable $x \sim N(m, s^2)$, it can be shown (see Appendix B) that

$$\mathrm{E}[e^{-kx}] = e^{-km+\frac{1}{2}k^2s^2}$$

for any constant k. In particular,

$$\mathrm{E}[e^{-\gamma x}] = e^{-\gamma m+\frac{1}{2}\gamma^2 s^2}$$

and

$$\begin{aligned}\mathrm{Var}[e^{-\gamma x}] &= \mathrm{E}[(e^{-\gamma x})^2] - \left(\mathrm{E}[e^{-\gamma x}]\right)^2 = \mathrm{E}[e^{-2\gamma x}] - \left(e^{-\gamma m+\frac{1}{2}\gamma^2 s^2}\right)^2 \\ &= e^{-2\gamma m+2\gamma^2 s^2} - \left(e^{-\gamma m+\frac{1}{2}\gamma^2 s^2}\right)^2 = \left(e^{-\gamma m+\frac{1}{2}\gamma^2 s^2}\right)^2 \left[e^{\gamma^2 s^2} - 1\right].\end{aligned}$$

In our case, we get

$$\mathrm{E}\left[\frac{u'(c)}{u'(c_0)}\right] = \mathrm{E}\left[\exp\left\{-\gamma \ln\left(\frac{c}{c_0}\right)\right\}\right] = \exp\left\{-\gamma\bar{g} + \frac{1}{2}\gamma^2\sigma_g^2\right\} \tag{8.6}$$

and

$$\frac{\sigma\left[u'(c)/u'(c_0)\right]}{\mathrm{E}\left[u'(c)/u'(c_0)\right]} = \sqrt{e^{\gamma^2\sigma_g^2} - 1} \approx \gamma\sigma_g, \tag{8.7}$$

where the approximation is based on $e^x \approx 1 + x$ for $x \approx 0$. The gross risk-free rate of return is then given by

$$R^f = e^{\delta}\left(\mathrm{E}\left[(c/c_0)^{-\gamma}\right]\right)^{-1} = \exp\left\{\delta + \gamma\bar{g} - \frac{1}{2}\gamma^2\sigma_g^2\right\}$$

so that the continuously compounded risk-free rate of return becomes

$$\ln R^f = \delta + \gamma\bar{g} - \frac{1}{2}\gamma^2\sigma_g^2. \tag{8.8}$$

From (8.2) and (8.7) we conclude that in the simple model the excess expected rate of return on a risky asset is

$$\mathrm{E}[R_i] - R^f \approx -\gamma\sigma_g\rho\left[R_i, (c/c_0)^{-\gamma}\right]\sigma[R_i]. \tag{8.9}$$

A first-order Taylor approximation of the function $f(x) = x^{-\gamma}$ around 1 gives $f(x) \approx f(1) + f'(1)(x-1) = 1 - \gamma(x-1)$ and with $x = c/c_0$ we get

$$\left(\frac{c}{c_0}\right)^{-\gamma} \approx 1 - \gamma\left(\frac{c}{c_0} - 1\right)$$

and, consequently,

$$\begin{aligned}\rho\left[R_i, \left(\frac{c}{c_0}\right)^{-\gamma}\right] &\approx \rho\left[R_i, 1 - \gamma\left(\frac{c}{c_0} - 1\right)\right] = \frac{\mathrm{Cov}\left[R_i, 1 - \gamma\left(\frac{c}{c_0} - 1\right)\right]}{\sigma[R_i]\sigma\left[1 - \gamma\left(\frac{c}{c_0} - 1\right)\right]} \\ &= \frac{-\gamma\,\mathrm{Cov}[R_i, c/c_0]}{\sigma[R_i]\gamma\sigma[c/c_0]} = -\rho[R_i, c/c_0].\end{aligned}$$

Therefore,

$$\mathrm{E}[R_i] - R^f \approx \gamma\sigma_g\rho\left[R_i, c/c_0\right]\sigma[R_i], \tag{8.10}$$

as in (8.3).

8.2.3 The simple one-period CCAPM: version 3

Assume that the individual has constant relative risk aversion and that the gross asset return R_i and the consumption growth c/c_0 are jointly lognormally distributed. This is a stronger condition than in version 2, which will allow us to obtain an exact relation between expected return and consumption growth. (The risk-free return will be as in version 2.) Any state-price deflator ζ satisfies $1 = \mathrm{E}[R_i\zeta]$ so

$$1 = \mathrm{E}\left[R_i e^{-\delta}\left(\frac{c}{c_0}\right)^{-\gamma}\right]. \tag{8.11}$$

Due to the distributional assumption, $R_i e^{-\delta}\left(\frac{c}{c_0}\right)^{-\gamma}$ is lognormally distributed. For any lognormally distributed random variable x, we have that

$$\mathrm{E}[x] = \mathrm{E}[e^{\ln x}] = e^{\mathrm{E}[\ln x] + \frac{1}{2}\operatorname{Var}[\ln x]},$$

and hence

$$\ln(\mathrm{E}[x]) = \mathrm{E}[\ln x] + \frac{1}{2}\operatorname{Var}[\ln x]. \tag{8.12}$$

Taking logs in (8.11), we therefore get

$$\begin{aligned}
0 &= \mathrm{E}\left[\ln R_i - \delta - \gamma \ln\left(\frac{c}{c_0}\right)\right] + \frac{1}{2}\operatorname{Var}\left[\ln R_i - \delta - \gamma \ln\left(\frac{c}{c_0}\right)\right] \\
&= \mathrm{E}[\ln R_i] - \delta - \gamma\,\mathrm{E}\left[\ln\left(\frac{c}{c_0}\right)\right] + \frac{1}{2}\operatorname{Var}[\ln R_i] + \frac{1}{2}\gamma^2 \operatorname{Var}\left[\ln\left(\frac{c}{c_0}\right)\right] - \gamma \operatorname{Cov}\left[\ln R_i, \ln\left(\frac{c}{c_0}\right)\right] \\
&= \mathrm{E}[\ln R_i] - \ln R^f + \frac{1}{2}\operatorname{Var}[\ln R_i] - \gamma \operatorname{Cov}\left[\ln R_i, \ln\left(\frac{c}{c_0}\right)\right],
\end{aligned}$$

where the last equality follows from the fact that risk-free rate still satisfies (8.8). Rearranging, we obtain

$$\mathrm{E}[\ln R_i] - \ln R^f + \frac{1}{2}\operatorname{Var}[\ln R_i] = \gamma \operatorname{Cov}\left[\ln R_i, \ln\left(\frac{c}{c_0}\right)\right],$$

which using (8.12) can be rewritten as

$$\ln(\mathrm{E}[R_i]) - \ln R^f = \gamma \operatorname{Cov}\left[\ln R_i, \ln\left(\frac{c}{c_0}\right)\right] = \gamma\sigma_g\rho[\ln R_i, \ln(c/c_0)]\sigma[\ln R_i]. \tag{8.13}$$

This relation is exact but holds only under the more restrictive assumption of joint lognormality of consumption and returns.

8.3 General multi-period link between consumption and asset returns

Now take the analysis to multi-period settings. Assuming time-additive expected utility we can define a state-price deflator from the optimal consumption plan of any individual as

$$\zeta_t = e^{-\delta t}\frac{u'(c_t)}{u'(c_0)}. \tag{8.14}$$

This is true both in the discrete-time and in the continuous-time setting. If a representative individual exists, the equation holds for aggregate consumption.

Not surprisingly, the multi-period discrete-time setting leads to equations very similar to those derived in the one-period framework in the previous section. Since

$$\frac{\zeta_{t+1}}{\zeta_t} = e^{-\delta}\frac{u'(c_{t+1})}{u'(c_t)},$$

we get from (4.23) and (4.24) that the risk-free gross return is

$$R_t^f = e^{\delta}\left(\mathrm{E}_t\left[\frac{u'(c_{t+1})}{u'(c_t)}\right]\right)^{-1} \tag{8.15}$$

and that the excess expected gross return on a risky asset is

$$\begin{aligned}\mathrm{E}_t[R_{i,t+1}] - R_t^f &= -\frac{\mathrm{Cov}_t[u'(c_{t+1})/u'(c_t), R_{i,t+1}]}{\mathrm{E}_t[u'(c_{t+1})/u'(c_t)]} \\ &= \rho_t\left[R_{i,t+1}, \frac{u'(c_{t+1})}{u'(c_t)}\right]\sigma_t[R_{i,t+1}]\left(-\frac{\sigma_t[u'(c_{t+1})/u'(c_t)]}{\mathrm{E}_t[u'(c_{t+1})/u'(c_t)]}\right).\end{aligned} \tag{8.16}$$

These equations are the multi-period equivalents of the Equations (8.1) and (8.2) for the one-period model. As in the one-period case, we can obtain an approximate relation between expected returns and relative consumption growth, $g_{t+1} = c_{t+1}/c_t - 1$, over the next period,

$$\mathrm{E}_t[R_{i,t+1}] - R_t^f \approx \gamma(c_t)\,\mathrm{Cov}_t[g_{t+1}, R_{i,t+1}], \tag{8.17}$$

and also an approximate consumption-beta equation

$$\mathrm{E}_t[R_{i,t+1}] - R_t^f \approx \beta_t[R_{i,t+1}, R_{t+1}^c]\left(\mathrm{E}_t[R_{t+1}^c] - R_t^f\right), \tag{8.18}$$

where $\beta_t[R_{i,t+1}, R_{t+1}^c] = \mathrm{Cov}_t[R_{i,t+1}, R_{t+1}^c]/\mathrm{Var}_t[R_{t+1}^c]$ and R_{t+1}^c is the gross return on a portfolio perfectly correlated with consumption growth over this period.

Now let us turn to the continuous-time setting. Suppose that the dynamics of consumption can be written as

$$dc_t = c_t\left[\mu_{ct}\,dt + \boldsymbol{\sigma}_{ct}^{\top}\,d\boldsymbol{z}_t\right], \tag{8.19}$$

where μ_{ct} is the expected relative growth rate of consumption and $\boldsymbol{\sigma}_{ct}$ is the vector of sensitivities of consumption growth to the exogenous shocks to the economy. In particular, the variance of relative consumption growth is given by $\|\boldsymbol{\sigma}_{ct}\|^2$. Given the dynamics of consumption and the relation (8.14) we can obtain the dynamics of ζ_t by an application of Itô's Lemma on the function $g(t,c) = e^{-\delta t}u'(c)/u'(c_0)$. The relevant derivatives are

$$\frac{\partial g}{\partial t}(t,c) = -\delta e^{-\delta t}\frac{u'(c)}{u'(c_0)}, \quad \frac{\partial g}{\partial c}(t,c) = e^{-\delta t}\frac{u''(c)}{u'(c_0)}, \quad \frac{\partial^2 g}{\partial c^2}(t,c) = e^{-\delta t}\frac{u'''(c)}{u'(c_0)},$$

implying that

$$\begin{aligned}\frac{\partial g}{\partial t}(t,c_t) &= -\delta e^{-\delta t}\frac{u'(c_t)}{u'(c_0)} = -\delta\zeta_t, \\ \frac{\partial g}{\partial c}(t,c_t) &= e^{-\delta t}\frac{u''(c_t)}{u'(c_0)} = \frac{u''(c_t)}{u'(c_t)}\zeta_t = -\gamma(c_t)c_t^{-1}\zeta_t, \\ \frac{\partial^2 g}{\partial c^2}(t,c_t) &= e^{-\delta t}\frac{u'''(c_t)}{u'(c_0)} = \frac{u'''(c_t)}{u'(c_t)}\zeta_t = \eta(c_t)c_t^{-2}\zeta_t,\end{aligned}$$

where $\gamma(c_t) \equiv -c_t u''(c_t)/u'(c_t)$ is the relative risk aversion of the individual, and where $\eta(c_t) \equiv c_t^2 u'''(c_t)/u'(c_t)$ is positive under the very plausible assumption that the absolute risk aversion

of the individual is decreasing in the level of consumption. Consequently, the dynamics of the state-price deflator can be expressed as

$$d\zeta_t = -\zeta_t \left[\left(\delta + \gamma(c_t)\mu_{ct} - \frac{1}{2}\eta(c_t)\|\boldsymbol{\sigma}_{ct}\|^2 \right) dt + \gamma(c_t)\boldsymbol{\sigma}_{ct}^\top \, d\boldsymbol{z}_t \right], \tag{8.20}$$

Comparing the above equation with (4.37), we can draw the conclusions summarized in the following theorem:

Theorem 8.1 *In a continuous-time economy where the optimal consumption process of an individual with time-additive expected utility satisfies* (8.19)*, the continuously compounded risk-free short-term interest rate satisfies*

$$r_t^f = \delta + \gamma(c_t)\mu_{ct} - \frac{1}{2}\eta(c_t)\|\boldsymbol{\sigma}_{ct}\|^2 \tag{8.21}$$

and

$$\boldsymbol{\lambda}_t = \gamma(c_t)\boldsymbol{\sigma}_{ct} \tag{8.22}$$

defines a market price of risk process. Here $\gamma(c_t) = -c_t u''(c_t)/u'(c_t)$ *and* $\eta(c_t) = c_t^2 u'''(c_t)/u'(c_t)$.

If we substitute (8.22) into (4.35) we see that the excess expected rate of return on asset i over the instant following time t can be written as

$$\mu_{it} + \delta_{it} - r_t^f = \gamma(c_t)\boldsymbol{\sigma}_{it}^\top \boldsymbol{\sigma}_{ct} = \gamma(c_t)\rho_{ict} \, \|\boldsymbol{\sigma}_{it}\| \, \|\boldsymbol{\sigma}_{ct}\|. \tag{8.23}$$

Here $\boldsymbol{\sigma}_{it}^\top \boldsymbol{\sigma}_{ct}$ and ρ_{ict} are the covariance and correlation between the rate of return on asset i and the consumption growth rate, respectively, and $\|\boldsymbol{\sigma}_{it}\|$ and $\|\boldsymbol{\sigma}_{ct}\|$ are standard deviations (volatilities) of the rate of return on asset i and the consumption growth rate, respectively. Equation (8.23) is the continuous-time version of (8.17). Note that the continuous-time relation is exact. Again, if we can find a trading strategy "mimicking" the consumption process (same volatility, perfect correlation) we get the "consumption-beta" relation

$$\mu_{it} + \delta_{it} - r_t^f = \beta_{it}^c \left(\mu_t^* + \delta_t^* - r_t^f \right),$$

where $\beta_{it}^c = \boldsymbol{\sigma}_{it}^\top \boldsymbol{\sigma}_{ct} / \|\boldsymbol{\sigma}_{ct}\|^2$.

If the market is effectively complete, the above equations are valid for aggregate consumption if we apply the utility function and time preference rate of the representative individual. The representative individual version of Equation (8.23) says that risky assets are priced so that the expected excess return on an asset is given by the product of the relative risk aversion of the representative individual and the covariance between the asset return and the growth rate of aggregate consumption. This is the key result in the Consumption-based CAPM (or just CCAPM) developed by Breeden (1979).

As already indicated in the one-period framework, we can obtain a relation between expected returns and aggregate consumption also if the market is incomplete and no representative individual exists. Let us stick to the continuous-time setting. Let $c_l = (c_{lt})$ denote the optimal consumption process of individual number l in the economy and assume that

$$dc_{lt} = c_{lt} \left[\mu_{clt} \, dt + \boldsymbol{\sigma}_{clt}^\top \, d\boldsymbol{z}_t \right].$$

If there are L individuals in the economy, aggregate consumption is $C_t = \sum_{l=1}^{L} c_{lt}$ and we have that

$$\begin{aligned} dC_t &= \sum_{l=1}^{L} dc_{lt} = \left(\sum_{l=1}^{L} c_{lt}\mu_{clt}\right) dt + \left(\sum_{l=1}^{L} c_{lt}\boldsymbol{\sigma}_{clt}\right)^{\top} d\boldsymbol{z}_t \\ &\equiv C_t \left[\mu_{Ct}\, dt + \boldsymbol{\sigma}_{Ct}^{\top}\, d\boldsymbol{z}_t\right], \end{aligned}$$

where $\mu_{Ct} \equiv \left(\sum_{l=1}^{L} c_{lt}\mu_{clt}\right)/C_t$ and $\boldsymbol{\sigma}_{Ct} = \left(\sum_{l=1}^{L} c_{lt}\boldsymbol{\sigma}_{clt}\right)/C_t$. We know from (8.23) that

$$\mu_{it} + \delta_{it} - r_t^f = A_l(c_{lt}) c_{lt} \boldsymbol{\sigma}_{it}^{\top} \boldsymbol{\sigma}_{clt}, \quad l = 1, \ldots, L,$$

where $A_l(c_{lt}) \equiv -u_l''(c_{lt})/u_l'(c_{lt})$ is the *absolute* risk aversion of individual l. Consequently,

$$\left(\mu_{it} + \delta_{it} - r_t^f\right) \frac{1}{A_l(c_{lt})} = c_{lt} \boldsymbol{\sigma}_{it}^{\top} \boldsymbol{\sigma}_{clt}, \quad l = 1, \ldots, L,$$

and summing up over l, we get

$$\left(\mu_{it} + \delta_{it} - r_t^f\right) \sum_{l=1}^{L} \left(\frac{1}{A_l(c_{lt})}\right) = \boldsymbol{\sigma}_{it}^{\top} \left(\sum_{l=1}^{L} c_{lt}\boldsymbol{\sigma}_{clt}\right) = \boldsymbol{\sigma}_{it}^{\top} \left(C_t \boldsymbol{\sigma}_{Ct}\right).$$

Therefore, we have the following relation between excess expected returns and aggregate consumption:

$$\mu_{it} + \delta_{it} - r_t^f = \frac{C_t}{\sum_{l=1}^{L} \left(\frac{1}{A_l(c_{lt})}\right)} \boldsymbol{\sigma}_{it}^{\top} \boldsymbol{\sigma}_{Ct}. \tag{8.24}$$

Relative to the complete markets version (8.23), the only difference is that the relative risk aversion of the representative individual is replaced by some complicated average of the risk aversions of the individuals. Note that if all individuals have CRRA utility with identical relative risk aversions, then $A_l(c_{lt}) = \gamma/c_{lt}$ and the multiplier $C_t/\sum_{l=1}^{L} \left(\frac{1}{A_l(c_{lt})}\right)$ in the above equation reduces to γ.

The Consumption-based CAPM is a very general asset pricing result. Basically any asset pricing model can be seen as a special case of the Consumption-based CAPM. On the other hand, the general Consumption-based CAPM is not very useful for applications without further assumptions. Therefore we turn now to more specific consumption-based models.

8.4 The simple multi-period CCAPM

A large part of the asset pricing literature makes (not always explicitly stated, unfortunately) the following additional assumptions:

1. the economy has a representative individual with CRRA time-additive utility, i.e. $u(C) = \frac{1}{1-\gamma} C^{1-\gamma}$,
2. future aggregate consumption is lognormally distributed.

In the discrete-time version of the model, we can proceed as in version 2 of the one-period model. The first assumption leads to a marginal rate of substitution given by

$$\frac{u'(C_{t+1})}{u'(C_t)} = \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma} = \exp\left\{-\gamma \ln\left(\frac{C_{t+1}}{C_t}\right)\right\}.$$

By the second assumption, $\ln(C_{t+1}/C_t) \sim N(\bar{g}, \sigma^2)$, and hence

$$\mathrm{E}_t\left[\frac{u'(C_{t+1})}{u'(C_t)}\right] = \mathrm{E}_t\left[\exp\left\{-\gamma \ln\left(\frac{C_{t+1}}{C_t}\right)\right\}\right] = \exp\left\{-\gamma\bar{g} + \frac{1}{2}\gamma^2\sigma^2\right\} \tag{8.25}$$

and

$$\frac{\sigma_t\left(u'(C_{t+1})/u'(C_t)\right)}{\mathrm{E}_t\left[u'(C_{t+1})/u'(C_t)\right]} = \sqrt{e^{\gamma^2\sigma^2} - 1} \approx \gamma\sigma. \tag{8.26}$$

According to (8.15), the gross risk-free return over the period from t to $t+1$ is then given by

$$R_t^f = e^{\delta}\left(\mathrm{E}_t\left[(C_{t+1}/C_t)^{-\gamma}\right]\right)^{-1} = \exp\left\{\delta + \gamma\bar{g} - \frac{1}{2}\gamma^2\sigma^2\right\}$$

so that the continuously compounded risk-free rate of return becomes

$$\ln R_t^f = \delta + \gamma\bar{g} - \frac{1}{2}\gamma^2\sigma^2. \tag{8.27}$$

From (8.17) we conclude that in the simple model the excess expected gross return on a risky asset is

$$\begin{aligned}\mathrm{E}_t\left[R_{i,t+1}\right] - R_t^f &\approx -\gamma\sigma\rho_t\left[R_{i,t+1}, \left(\frac{C_{t+1}}{C_t}\right)^{-\gamma}\right]\sigma_t\left[R_{i,t+1}\right] \\ &\approx \gamma\sigma\rho_t\left[R_{i,t+1}, \frac{C_{t+1}}{C_t}\right]\sigma_t\left[R_{i,t+1}\right],\end{aligned} \tag{8.28}$$

where the last expression follows from a first-order Taylor approximation as in Section 8.2.2. If we further assume that the future asset price and the future consumption level are simultaneously lognormally distributed, we are back in version 3 of the one-period model, and the following exact relation can be shown:

$$\ln\left(\mathrm{E}_t[R_{i,t+1}]\right) - \ln R_t^f = \gamma\sigma\rho_t\left[\ln R_{i,t+1}, \ln\left(\frac{C_{t+1}}{C_t}\right)\right]\sigma_t\left[\ln R_{i,t+1}\right].$$

In the formulas above the mean $\bar{g}$ and variance σ^2 of consumption growth can vary over time, but in the stationary case where both are constant we see that the risk-free interest rate must also be constant. Furthermore, a risky asset with a constant correlation with consumption growth will have a constant Sharpe ratio $\left(\mathrm{E}_t\left[R_{i,t+1}\right] - R_t^f\right)/\sigma_t[R_{i,t+1}]$. If the standard deviation of the return is constant, the expected excess rate of return will also be constant.

In the continuous-time version of the stationary simple consumption-based model, the second assumption means that μ_{Ct} and $\boldsymbol{\sigma}_{Ct}$ in the consumption process (8.19) are constant, i.e. consumption follows a geometric Brownian motion. It follows from (8.21) and (8.23) that the model with these assumptions generate a constant continuously compounded short-term risk-free interest rate of

$$r^f = \delta + \gamma\mu_C - \frac{1}{2}\gamma(1+\gamma)\|\boldsymbol{\sigma}_C\|^2 \tag{8.29}$$

and a constant Sharpe ratio for asset i given by

$$\frac{\mu_{it} + \delta_{it} - r^f}{\|\boldsymbol{\sigma}_{it}\|} = \gamma\rho_{iC}\|\boldsymbol{\sigma}_C\| \tag{8.30}$$

if the asset has a constant correlation with consumption.

8.5 Theory meets data — asset pricing puzzles

The simple consumption-based model has been exposed to numerous empirical tests. Almost all of these tests focus on the question whether the relation (8.30)—or the similar discrete-time relation (8.28)—holds for a broad-based stock index for reasonable values of the relative risk aversion coefficient. Let us use the following stylized figures that are roughly representative for the U.S. economy over the second half of the 20th century:

- the average annual real excess rate of return on the stock index (relative to the yield on a short-term government bond) is about 8.0%;
- the empirical standard deviation of the annual real rate of return on the stock index is about 20.0%;
- the empirical standard deviation of annual relative changes in aggregate consumption is about 2.0%;
- the empirical correlation between the real return on the stock index and changes in aggregate consumption is about 0.2.

Inserting these estimates into (8.30) it follows that the relative risk aversion coefficient γ must be 100. This is certainly an unrealistically high risk aversion for a typical individual, cf. the discussion in Section 5.6.2. In fact, this computation—which is standard in the literature—exaggerates the problem somewhat. The estimates of the return and consumption moments are based on discrete observations, not continuous observations, so we should use the discrete-time version of the model. If we drop the approximation in (8.26), the discrete-time simple model says that

$$\mathrm{E}_t\left[R_{i,t+1}\right] - R_t^f \approx \sqrt{e^{\gamma^2\sigma^2} - 1}\,\rho_t\left[R_{i,t+1}, \frac{C_{t+1}}{C_t}\right]\sigma_t\left[R_{i,t+1}\right]$$

and plugging in the estimates, we need

$$\gamma \approx \frac{1}{\sigma}\sqrt{\ln\left[1 + \left(\frac{\mathrm{E}_t[R_{i,t+1}] - R_t^f}{\rho\sigma_t[R_{i,t+1}]}\right)^2\right]} = \frac{1}{0.02}\sqrt{\ln(5)} \approx 63.4.$$

But 63.4 is still an unreasonably high relative risk aversion.

With $\gamma = 100$, Equation (8.29) indicates that an increase of one percentage point in the expected growth rate of aggregate consumption will be accompanied by an increase in the short-term interest rate of 100 percentage points—also very unrealistic. Similar results are obtained for other countries and other data periods.

For a reasonable level of risk aversion (probably in the area 2–5), the expected excess rate of return predicted by the theory is much smaller than the historical average. This so-called *equity premium puzzle* was first pointed out by Mehra and Prescott (1985). Weil (1989) notes that the simple model predicts a higher level of interest rates than observed empirically. This is the so-called *risk-free rate puzzle*.

In the simple model, the real short-term interest rate and the Sharpe ratios of risky assets are constant over time, which is also inconsistent with empirical observations. Interest rates vary over time, and recent studies indicate that Sharpe ratios, expected returns, and volatilities on

stocks vary over the business cycle with high values in recessions and low values in periods of high economic growth rates, cf. e.g. Cochrane (2001). The future values of these variables are therefore to some extent predictable, contrasting the simple model. This could be called a *predictability puzzle*.

In addition, empirical studies show that the simple model can explain only a small part of the differences in the average returns on different stocks, cf. e.g. Breeden, Gibbons, and Litzenberger (1989). This is a *cross-sectional stock return puzzle*.

Based on the apparently large discrepancy between the model and the data, it is tempting to conclude that the consumption-based approach to asset pricing is not applicable, and otherwise intelligent economists have not been able to withstand this temptation. The conclusion is not fair, however. Firstly, as explained in the following section, there are a number of problems with the empirical tests of the model which make their conclusions less clear. Secondly, it is only the very simple special case of the general consumption-based asset pricing model which is tested. The consumption-based approach in itself is based on only very few and relatively undisputed assumptions so the lack of empirical support must be blamed on the additional assumptions of the simple model. In the last 10-15 years, a number of alternative specifications of the general model have been developed. Most of these alternatives assume either a different representation of preferences than in the simple model or that the market is incomplete so that the representative individual approach is invalid. We will discuss these two types of model extensions in Sections 8.7 and 8.8.

8.6 Problems with the empirical studies

A number of issues should be taken into account when the conclusions of the empirical studies of the consumption-based model are evaluated. In the following we discuss some selected issues.

Firstly, the 8% estimate of the average excess stock return is relatively imprecise. With 50 observations (one per year) and a standard deviation of 20%, the standard error equals $20\%/\sqrt{50} \approx 2.8\%$ so that a 95% confidence interval is roughly $[2.5\%, 13.5\%]$. With a mean return of 2.5% instead of 8.0%, the required value of the relative risk aversion drops to 31.25, which is certainly still very high but nevertheless considerably smaller than the original value of 100.

Secondly, the consumption data used in the tests is of a poor quality as pointed out by e.g. Breeden, Gibbons, and Litzenberger (1989). The available data on aggregate consumption measures the expenses for purchases of consumption goods over a given period (usually a quarter or a month). This is problematic for several reasons. Many, especially very expensive, consumption goods are durable goods offering consumption "services" beyond the period of purchase. The model addresses the rate of consumption at a given point in time rather than the sum of consumption rates over some time interval; see Grossman, Melino, and Shiller (1987). The consumption data is reported infrequently relative to financial data. Moreover, the aggregate consumption numbers are undoubtedly subject to various sampling and measurement errors; see Wilcox (1992). These problems motivate the development of asset pricing models that do not depend on consumption at all. This is for example the case of the so-called factor models discussed in Chapter 9.

Thirdly, also data for stock returns has to be selected and applied with caution. Most tests of the asset pricing models are based on data from the U.S. or other economies that have experienced

relatively high growth at least throughout the last 50 years. Probably investors in these countries were not so sure 50 years ago that the economy would avoid major financial and political crises and outperform other countries. Brown, Goetzmann, and Ross (1995) point out that due to this *survivorship bias* the realized stock returns overstate the ex-ante expected rate of returns significantly by maybe as much as 2–4 percentage points. Note however that in many crises in which stocks do badly, also bonds and deposits tend to provide low returns so it is not clear how big the effect on the expected *excess* stock return is.

Fourthly, the pricing relations of the model involve the ex-ante expectations of individuals, while the tests of the model are based on a single realized sequence of market prices and consumption. Estimating and testing a model involving ex-ante expectations and other moments requires stationarity in data in the sense that it must be assumed that each of the annual observations is drawn from the same probability distribution. For example, the average of the observed annual stock market returns is only a reasonable estimate of the ex-ante expected annual stock market return if the stock market return in each of the 50 years is drawn from the same probability distribution. Some important changes in the investment environment over the past years invalidate the stationarity assumption. For example, Mehra and Prescott (2003) note two significant changes in the U.S. tax system in the period between 1960 and 2000, a period included in most tests of the consumption-based model:

- The marginal tax rate for stock dividends has dropped from 43% to 17%.
- Stock returns in most pension savings accounts are now tax-exempt, which was not so in the 1960s. Bond returns in savings accounts have been tax-exempt throughout the period.

Both changes have led to increased demands for stocks with stock price increases as a result. These changes in the tax rules were hardly predicted by investors and, hence, they can partly explain why the model has problems explaining the high realized stock returns. Similarly, it can be argued that the reductions in direct and indirect transaction costs and the liberalizations of international financial markets experienced over the last decades have increased the demands for stocks and driven up stock returns above what could be expected ex-ante. The high transaction costs and restrictions on particularly international investments in the past may have made it impossible or at least very expensive for investors to obtain the optimal diversification of their investments so that even unsystematic risks may have been priced with higher required returns as a consequence.[1]

One can also argue theoretically that the returns of a given stock cannot be stationary. Here is the argument: in each period there is a probability that the issuing firm defaults and the stock stops to exist. Then it no longer makes sense to talk about the return probability distribution of that stock. More generally, the probability distribution of the return in one period may very well depend on the returns in previous periods.

Fifthly, as emphasized by Bossaerts (2002), standard tests assume that ex-ante expectations of individuals are correct in the sense that they are confirmed by realizations. The general asset pricing theory does allow individuals to have systematically over-optimistic or over-pessimistic expectations. The usual tests implicitly assume that market data can be seen as realizations of

[1]It is technically complicated to include transaction costs and trading restrictions in asset pricing models, but a study of He and Modest (1995) indicates that such imperfections at least in part can explain the equity premium puzzle.

the ex-ante expectations of individuals. Bossaerts (2002) describes studies which indicate that this assumption is not necessarily valid.

8.7 CCAPM with alternative preferences

An interesting alternative to the simple consumption-based model is to allow the utility of a given consumption level at a given point in time t to depend on some benchmark X_t, i.e. the preferences are modeled by $\mathrm{E}[\int_0^T e^{-\delta t}u(c_t, X_t)\,dt]$ in continuous time or $\mathrm{E}[\sum_{t=0}^T e^{-\delta t}u(c_t, X_t)]$ in discrete time. This incorporates the case of (internal) habit formation where X_t is determined as a weighted average of the previous consumption rates of the individual, and the case of state-dependent (or "external habit") preferences where X_t is a variable not affected by the consumption decisions of the individual. If we assume that a high value of the benchmark means that the individual will be more eager to increase consumption, as is the case with (internal) habit formation, we should have $u_{cX}(c, X) > 0$.

Typically, models apply one of two tractable specifications of the utility function. The first specification is

$$u(c_t, X_t) = \frac{1}{1-\gamma}\left(c_t - X_t\right)^{1-\gamma}, \quad \gamma > 0, \tag{8.31}$$

which is defined for $c_t > X_t$. Marginal utility is

$$u_c(c_t, X_t) = \left(c_t - X_t\right)^{-\gamma},$$

and the relative risk aversion is

$$-\frac{c_t u_{cc}(c_t, X_t)}{u_c(c_t, X_t)} = \gamma\frac{c_t}{c_t - X_t}, \tag{8.32}$$

which is no longer constant and is greater than γ. This will allow us to match historical consumption and stock market data with a lower value of the parameter γ, but it is more reasonable to study whether we can match the data with a fairly low average value of the relative risk aversion given by (8.32). The second tractable specification is

$$u(c_t, X_t) = \frac{1}{1-\gamma}\left(\frac{c_t}{X_t}\right)^{1-\gamma}, \quad \gamma > 0, \tag{8.33}$$

which is defined for $c_t, X_t > 0$. Since

$$u_c(c, X) = c^{-\gamma}X^{\gamma-1}, \qquad u_{cX}(c, X) = (\gamma - 1)c^{-\gamma}X^{\gamma-2}$$

we need $\gamma > 1$ to ensure that marginal utility is increasing in X. Despite the generalization of the utility function relative to the standard model, the relative risk aversion is still constant:

$$-\frac{cu_{cc}(c, X)}{u_c(c, X; \gamma)} = \gamma.$$

8.7.1 Habit formation

In the case with (internal) habit formation the individual will appreciate a given level of consumption at a given date higher if she is used to low consumption than if she is used to high consumption. Such a representation of preferences is still consistent with the von Neumann and

Morgenstern (1944) axioms of expected utility but violates the time-additivity of utility typically assumed. A rational individual with an internal habit will take into account the effects of her current decisions on the future habit levels. We saw in Chapter 6 that this will considerably complicate the formulas for optimal consumption and for the associated state-price deflator.

Individuals with habit formation in preferences will other things equal invest more in the risk-free asset in order to ensure that future consumption will not come very close to (or even below) the future habit level. This extra demand for the risk-free asset will lower the equilibrium interest rate and, hence, help resolve the risk-free rate puzzle. Moreover, we see from (8.32) that the risk aversion is higher in "bad states" where current consumption is close to the habit level than in "good states" of high current consumption relative to the benchmark. This will be reflected by the risk premia of risky assets and has the potential to explain the observed cyclical behavior of expected returns.

As we have seen in Chapter 6 the state-price deflators that can be derived from individuals with internal habit formation are considerably more complex than with time-additive preferences or external habit formation. A general and rather abstract continuous-time analysis for the case of an internal habit defined by a weighted average of earlier consumption rates is given by Detemple and Zapatero (1991).

Only very few concrete asset pricing models with internal habit have been developed with the continuous-time model of Constantinides (1990) as the most frequently cited. The model of Constantinides assumes a representative individual who can invest in a risk-free asset and a single risky asset. A priori, the risk-free rate and the expected rate of return and the volatility of the risky asset are assumed to be constant. The utility of the individual is given by (8.31), where the habit level is a weighted average of consumption at all previous dates. Constantinides solves for the optimal consumption and investment strategies of the individual and shows that the optimal consumption rate varies much less over time in the model with habit formation than with the usual time-additive specification of utility. A calibration of the model to historical data shows that the model is consistent with a large equity premium and a low risk aversion but, on the other hand, the consumption process of the model has an unrealistically high auto-correlation and the variance of long-term consumption growth is quite high. By construction, the model cannot explain variations in interest rates and expected returns on stocks.

8.7.2 State-dependent utility: general results

With an external habit/benchmark, e.g. given by the aggregate consumption level, the state-price deflator is

$$\zeta_t = e^{-\delta t} \frac{u_c(c_t, X_t)}{u_c(c_0, X_0)},$$

where the only difference to the model without habit is that the marginal utilities depend on the habit level. An external habit formalizes the idea that the marginal utility of consumption of one individual is increasing in the consumption level of other individuals (sometimes referred to as the "keeping up with the Jones'es" effect). In this case, there is no effect of the consumption choice of the individual on the future benchmark levels, but of course the individual will include her knowledge of the dynamics of the benchmark when making consumption decisions.

In a discrete-time setting the one-period deflator is

$$\frac{\zeta_{t+1}}{\zeta_t} = e^{-\delta}\frac{u_c(c_{t+1}, X_{t+1})}{u_c(c_t, X_t)}.$$

Using the Taylor approximation

$$u_c(c_{t+1}, X_{t+1}) \approx u_c(c_t, X_t) + u_{cc}(c_t, X_t)\Delta c_{t+1} + u_{cX}(c_t, X_t)\Delta X_{t+1},$$

the approximate relation

$$\mathrm{E}_t[R_{i,t+1}] - R_t^f \approx \left(-\frac{c_t u_{cc}(c_t, X_t)}{u_c(c_t, X_t)}\right)\mathrm{Cov}_t\left[R_{i,t+1}, \frac{\Delta c_{t+1}}{c_t}\right] - \frac{u_{cX}(c_t, X_t)}{u_c(c_t, X_t)}\mathrm{Cov}_t\left[R_{i,t+1}, \Delta X_{t+1}\right] \tag{8.34}$$

can be derived. The covariance of return with the benchmark variable adds a term to the excess expected return of a risky asset. Moreover, the relative risk aversion in the first term will now generally vary with the benchmark variable.

In a continuous-time setting where the dynamics of consumption is again given by (8.19) and the dynamics of the benchmark process $X = (X_t)$ is of the form

$$dX_t = \mu_{Xt}\,dt + \boldsymbol{\sigma}_{Xt}^\top\,d\boldsymbol{z}_t, \tag{8.35}$$

an application of Itô's Lemma will give the dynamics of the state-price deflator. As before, the risk-free rate and the market price of risk can be identified from the drift and the sensitivity, respectively, of the state-price deflator. The following theorem states the conclusion. Exercise 8.3 asks for the proof.

Theorem 8.2 *In a continuous-time economy where the optimal consumption process of an individual with state-dependent expected utility satisfies* (8.19) *and the dynamics of the benchmark is given by* (8.35)*, the continuously compounded risk-free short-term interest rate satisfies*

$$\begin{aligned} r_t^f = \delta &+ \frac{-c_t u_{cc}(c_t, X_t)}{u_c(c_t, X_t)}\mu_{ct} - \frac{1}{2}\frac{c_t^2 u_{ccc}(c_t, X_t)}{u_c(c_t, X_t)}\|\boldsymbol{\sigma}_{ct}\|^2 \\ &- \frac{u_{cX}(c_t, X_t)}{u_c(c_t, X_t)}\mu_{Xt} - \frac{1}{2}\frac{u_{cXX}(c_t, X_t)}{u_c(c_t, X_t)}\|\boldsymbol{\sigma}_{Xt}\|^2 - \frac{c_t u_{ccX}(c_t, X_t)}{u_c(c_t, X_t)}\boldsymbol{\sigma}_{ct}^\top\boldsymbol{\sigma}_{Xt} \end{aligned} \tag{8.36}$$

and

$$\boldsymbol{\lambda}_t = \frac{-c_t u_{cc}(c_t, X_t)}{u_c(c_t, X_t)}\boldsymbol{\sigma}_{ct} - \frac{u_{cX}(c_t, X_t)}{u_c(c_t, X_t)}\boldsymbol{\sigma}_{Xt} \tag{8.37}$$

defines a market price of risk process. In particular, the excess expected rate of return on asset i is

$$\mu_{it} + \delta_{it} - r_t^f = \frac{-c_t u_{cc}(c_t, X_t)}{u_c(c_t, X_t)}\boldsymbol{\sigma}_{it}^\top\boldsymbol{\sigma}_{ct} - \frac{u_{cX}(c_t, X_t)}{u_c(c_t, X_t)}\boldsymbol{\sigma}_{it}^\top\boldsymbol{\sigma}_{Xt}. \tag{8.38}$$

Assuming $u_{CX}(c, X) > 0$, we see from (8.37) that, if $\boldsymbol{\sigma}_{Xt}$ goes in the same direction as $\boldsymbol{\sigma}_{ct}$, state-dependent utility will tend to lower the market price of risk, working against a resolution of the equity premium puzzle. On the other hand there is much more room for time variation in both the risk-free rate and the market price of risk, which can help explain the predictability puzzle.

Again, the above results hold for aggregate consumption if a representative individual with preferences of the given form exists. If $\boldsymbol{\sigma}_{Xt} = \mathbf{0}$, we can link asset prices to aggregate consumption

without assuming the existence of a representative individual, as in the case of standard preferences. We obtain

$$\mu_{it} + \delta_{it} - r_t^f = \frac{C_t}{\sum_{l=1}^{L} \left(\frac{1}{A_l(c_{lt}, X_t)} \right)} \boldsymbol{\sigma}_{it}^{\top} \boldsymbol{\sigma}_{Ct}, \tag{8.39}$$

where C_t is aggregate consumption and $A_l(c_{lt}, X_t) = -u_{cc}^l(c_{lt}, X_t)/u_c^l(c_{lt}, X_t)$ is the (now state-dependent) absolute risk aversion of individual l.

8.7.3 The Campbell and Cochrane model

Campbell and Cochrane (1999) suggest a discrete-time model of an economy with identical individuals with utility functions like (8.31), where X_t is the benchmark or external habit level. The "next-period deflator" is then

$$m_{t+1} \equiv \frac{\zeta_{t+1}}{\zeta_t} = e^{-\delta} \frac{(c_{t+1} - X_{t+1})^{-\gamma}}{(c_t - X_t)^{-\gamma}}.$$

The lognormal distributional assumption for aggregate consumption made in the simple model seems to be empirically reasonable so the Campbell and Cochrane model keeps that assumption. Since it is not obvious what distributional assumption on X_t that will make the model computationally tractable, Campbell and Cochrane define the "surplus consumption ratio" $S_t = (c_t - X_t)/c_t$ in terms of which the "next-period deflator" can be rewritten as

$$m_{t+1} = e^{-\delta} \left(\frac{c_{t+1}}{c_t} \right)^{-\gamma} \left(\frac{S_{t+1}}{S_t} \right)^{-\gamma} = \exp \left\{ -\delta - \gamma \ln \left(\frac{c_{t+1}}{c_t} \right) - \gamma \ln \left(\frac{S_{t+1}}{S_t} \right) \right\}.$$

It is assumed that changes in both consumption growth and the surplus consumption ratio are conditionally lognormally distributed with

$$\begin{aligned} \ln \left(\frac{c_{t+1}}{c_t} \right) &= \bar{g} + \nu_{t+1}, \\ \ln \left(\frac{S_{t+1}}{S_t} \right) &= (1 - \varphi) \left(\ln \bar{S} - \ln S_t \right) + \Lambda(S_t) \nu_{t+1}, \end{aligned}$$

where $\nu_{t+1} \sim N(0, \sigma^2)$ and the function Λ is specified below with the purpose of obtaining some desired properties. With lognormality of c_{t+1} and S_{t+1}, $c_{t+1} - X_{t+1} = c_{t+1} S_{t+1}$ and therefore the next-period deflator will also be lognormal. Note that $\ln S_t$ will fluctuate around $\ln \bar{S}$, which may think of as representing business cycles with low values of $\ln S_t$ corresponding to bad times with relatively low consumption. Also note that the consumption and the surplus consumption ratios are perfectly correlated.

The distributional assumption on the surplus consumption ratio ensures that it stays positive so that $c_t > X_t$, as required by the utility specification. On the other hand, if you want the benchmark X_t to stay positive, the surplus consumption ratio must be smaller than 1, which is not ensured by the lognormal distribution. According to Campbell, Lo, and MacKinlay (1997, p. 331), it can be shown that

$$\ln X_{t+1} \approx \ln(1 - \bar{S}) + \frac{g}{1 - \varphi} + (1 - \varphi) \sum_{j=0}^{\infty} \varphi^j \ln c_{t+j}$$

so that the benchmark is related to a weighted average of past consumption levels.

With the above assumptions, the next-period deflator m_{t+1} is

$$m_{t+1} = \exp\left\{-\delta - \gamma\left[\bar{g} - (1-\varphi)\ln\frac{S_t}{\bar{S}}\right] - \gamma\left(1+\Lambda(S_t)\right)\nu_{t+1}\right\},$$

which is conditionally lognormally distributed with

$$\frac{\sigma_t[m_{t+1}]}{\mathrm{E}_t[m_{t+1}]} = \sqrt{\exp\left\{\gamma^2\sigma^2(1+\Lambda(S_t))^2\right\} - 1} \approx \gamma\sigma\left(1+\Lambda(S_t)\right).$$

(Again the approximation is not that accurate for the parameter values necessary to match consumption and return data.) It follows from (4.24) that the expected excess gross return on a risky asset is

$$\mathrm{E}_t[R_{i,t+1}] - R_t^f \approx \gamma\sigma\left(1+\Lambda(S_t)\right)\rho_t\left[m_{t+1}, R_{i,t+1}\right]\sigma_t[R_{i,t+1}]. \tag{8.40}$$

The variation in risk aversion through S_t increases the risk premium relative to the standard model, cf. (8.28). In order to obtain the counter-cyclical variation in risk premia observed in data, Λ has to be a decreasing function of S.

The continuously compounded short-term risk-free rate is

$$\ln R_t^f = \ln\left(\frac{1}{\mathrm{E}_t[m_{t+1}]}\right) = \delta + \gamma\bar{g} - \frac{1}{2}\gamma^2\sigma^2 - \gamma(1-\varphi)\ln\frac{S_t}{\bar{S}} - \frac{1}{2}\gamma^2\sigma^2\Lambda(S_t)\left(\Lambda(S_t)+2\right).$$

In comparison with the expression (8.27) for the risk-free rate in the simple model, the last two terms on the right-hand side are new. Note that S_t has opposite effects on the two new terms. A low value of S_t means that the marginal utility of consumption is high so that individuals will try to borrow money for current consumption. This added demand for short-term borrowing will drive up the equilibrium interest rate. On the other hand, a low S_t will also increase the risk aversion and, hence, precautionary savings with a lower equilibrium rate as a result. Campbell and Cochrane fix $\Lambda(\cdot)$ and $\bar{S}$ at

$$\Lambda(S_t) = \frac{1}{\bar{S}}\sqrt{1 - 2\ln(S_t/\bar{S})} - 1, \qquad \bar{S} = \sigma\sqrt{\gamma/(1-\varphi)}$$

so that the equilibrium interest rate is given by the constant

$$\ln R^f = \delta + \gamma\bar{g} - \frac{1}{2}\sigma^2\left(\frac{\gamma}{\bar{S}}\right)^2.$$

Note that $\Lambda(\cdot)$ *is* decreasing.

The authors calibrate the model to historical data for consumption growth, interest rates, and the average Sharpe ratio, which for example requires that $\gamma = 2$ and $\bar{S} = 0.057$. The calibrated model is consistent with the observed counter-cyclical variation in expected returns, standard deviations of returns, and the Sharpe ratio. With the given parameter values the model can therefore explain the predictability in those variables. The calibrated model yields empirically reasonable levels of the expected return and standard deviation of stock returns but note that, although the calibrated value of the utility parameter γ is small, the relative risk aversion γ/S_t is still high. With $S_t = \bar{S} = 0.057$, the risk aversion is approximately 35, and the risk aversion is much higher in bad states where S_t is low. The model can therefore not explain the equity premium puzzle. Also observe that the value of $\bar{S}$ implies an average habit level only 5.7% lower than current consumption, which seems to be an extreme degree of habit formation. Finally, note that the dynamics of the business cycle variable S_t is exogenously chosen to obtain the desired properties of the model. It would be interesting to understand how such a process could arise endogenously.

8.7.4 The Chan and Kogan model

In many models for individual consumption and portfolio choice the individual is assumed to have constant relative risk aversion both because this seems quite reasonable and because this simplifies the analysis. If all individuals in an economy have constant relative risk aversion, you might think that a representative individual would also have constant relative risk aversion, but as pointed out by Chan and Kogan (2002) this is only true if they all have the same level of risk aversion. Individuals with a low (constant) relative risk aversion will other things equal invest a larger fraction of their wealth in risky assets, e.g. stocks, than individuals with high (constant) relative risk aversion. Increasing stock prices will imply that the wealth of the relatively risk tolerant individuals will grow more than the wealth of comparably more risk averse individuals. Consequently, the aggregate risk aversion in the economy (corresponding to the risk aversion of a representative individual) will fall. Conversely, the aggregate risk aversion will increase when stock markets drop. This simple observation supports the assumption of Campbell and Cochrane (1999) discussed above that the risk aversion of the representative individual varies counter-cyclically, which helps in resolving asset pricing puzzles.

Let us take a closer look at the model of Chan and Kogan (2002). It is a continuous-time exchange economy in which the aggregate endowment/consumption $Y = (Y_t)$ follows the a geometric Brownian motion

$$dY_t = Y_t[\mu\, dt + \sigma\, dz_t],$$

where $\mu > \sigma^2/2$, $\sigma > 0$, and $z = (z_t)$ is a one-dimensional standard Brownian motion. Two assets are traded: a risky asset which is a unit net supply and pays a continuous dividend equal to the aggregate endowment, and an instantaneously risk-free asset (a bank account) generating a continuously compounded short-term interest rate of r_t^f. The economy is populated with infinitely-lived individuals maximizing time-additive state-dependent utility given by (8.33), i.e.

$$\mathrm{E}\left[\int_0^\infty e^{-\delta t} u(c_t, X_t; \gamma)\, dt\right], \quad u(c, X; \gamma) = \frac{1}{1-\gamma}\left(\frac{c}{X}\right)^{1-\gamma}.$$

Here X is an external benchmark, e.g. an index of the standard of living in the economy. Let $x_t = \ln X_t$ and $y_t = \ln Y_t$. The dynamics of the benchmark is modeled through

$$x_t = e^{-\kappa t} x_0 + \kappa \int_0^t e^{-\kappa(t-s)} y_s\, ds$$

so that

$$dx_t = \kappa\,(y_t - x_t)\; dt.$$

The log-benchmark is a weighted average of past log-consumption. The relative log-consumption variable $\omega_t = y_t - x_t$ will be representative of the state of the economy. A high [low] value of ω_t represents a good [bad] state in terms of aggregate consumption relative to the benchmark. Note that

$$d\omega_t = dy_t - dx_t = \kappa\,(\bar{\omega} - \omega_t)\; dt + \sigma\, dz_t,$$

where $\bar{\omega} = (\mu - \sigma^2/2)/\kappa$.

Individuals are assumed to differ with respect to their relative risk aversion γ but to have identical subjective time preference parameters δ. Since the market is complete, an equilibrium in the economy will be Pareto-optimal and identical to the solution of the problem of a central

planner or representative individual. Let $f(\gamma)$ denote the weight of the individuals with relative risk aversion γ in this problem, where we normalize so that $\int_1^\infty f(\gamma) = 1$. The problem of the central planner is to solve

$$\sup_{\{c_t(Y_t,X_t;\gamma);\gamma>1,t\geq 0\}} \mathrm{E}\left[\int_0^\infty e^{-\delta t}\left(\int_1^\infty f(\gamma)\frac{1}{1-\gamma}\left(\frac{c_t(Y_t,X_t;\gamma)}{X_t}\right)^{1-\gamma} d\gamma\right) dt\right]$$
$$\text{s.t.} \int_1^\infty c_t(Y_t,X_t;\gamma)\, d\gamma \leq Y_t, \quad t \geq 0.$$

In an exchange economy no intertemporal transfer of resources are possible at the aggregate level so the optimal value of the central planner's objective function will be $\mathrm{E}[\int_0^\infty e^{-\delta t}U(Y_t,X_t)\,dt]$, where

$$U(Y_t,X_t) = \sup_{\{c_t(Y_t,X_t;\gamma);\gamma>1\}} \left\{\int_1^\infty f(\gamma)\frac{1}{1-\gamma}\left(\frac{c_t(Y_t,X_t;\gamma)}{X_t}\right)^{1-\gamma} d\gamma \,\Big|\, \int_1^\infty c_t(Y_t,X_t;\gamma)\, d\gamma \leq Y_t\right\}. \tag{8.41}$$

The solution to this problem is characterized in the following theorem, which is the key to the asset pricing results in this model.

Theorem 8.3 *The optimal consumption allocation in the problem* (8.41) *is*

$$c_t(Y_t,X_t;\gamma) = \alpha_t(\omega_t;\gamma)Y_t, \quad \alpha_t(\omega_t;\gamma) = f(\gamma)^{1/\gamma}e^{-\frac{1}{\gamma}h(\omega_t)-\omega_t}, \tag{8.42}$$

where the function h is implicitly defined by the identity

$$\int_1^\infty f(\gamma)^{1/\gamma}e^{-\frac{1}{\gamma}h(\omega_t)-\omega_t}\, d\gamma = 1. \tag{8.43}$$

The utility function of the representative individual is

$$U(Y_t,X_t) = \int_1^\infty \frac{1}{1-\gamma}f(\gamma)^{1/\gamma}e^{-\frac{1-\gamma}{\gamma}h(\omega_t)}\, d\gamma. \tag{8.44}$$

Proof: If we divide the constraint in (8.41) through by X_t and introduce the aggregate consumption share $\alpha_t(Y_t,X_t;\gamma) = c_t(Y_t,X_t;\gamma)/Y_t$, the Lagrangian for the optimization problem is

$$\begin{aligned}\mathcal{L}_t &= \int_1^\infty f(\gamma)\frac{1}{1-\gamma}\left(\alpha_t(Y_t,X_t;\gamma)\frac{Y_t}{X_t}\right)^{1-\gamma} d\gamma + H_t\left(\frac{Y_t}{X_t} - \int_1^\infty \alpha_t(Y_t,X_t;\gamma)\frac{Y_t}{X_t}\, d\gamma\right)\\ &= \frac{Y_t}{X_t}\int_1^\infty \left[f(\gamma)\frac{1}{1-\gamma}\alpha_t(Y_t,X_t;\gamma)^{1-\gamma}\left(\frac{Y_t}{X_t}\right)^{-\gamma} - H_t\alpha_t(Y_t,X_t;\gamma)\right] d\gamma + H_t\frac{Y_t}{X_t},\end{aligned}$$

where H_t is the Lagrange multiplier. The first-order condition for α_t implies that

$$\alpha_t(Y_t,X_t;\gamma) = H_t^{-1/\gamma}f(\gamma)^{1/\gamma}\left(\frac{Y_t}{X_t}\right)^{-1} = f(\gamma)^{1/\gamma}e^{-\frac{1}{\gamma}h_t-\omega_t},$$

where $h_t = \ln H_t$. The consumption allocated to all individuals must add up to the aggregate consumption, which implies the condition (8.43). From the condition, it is clear that h_t and therefore α_t depends on ω_t but not separately on Y_t and X_t.

The maximum of the objective function is

$$\begin{aligned}U(Y_t,X_t) &= \int_1^\infty f(\gamma)\frac{1}{1-\gamma}\left(\alpha_t(Y_t,X_t;\gamma)\frac{Y_t}{X_t}\right)^{1-\gamma} d\gamma\\ &= \int_1^\infty f(\gamma)\frac{1}{1-\gamma}f(\gamma)^{(1-\gamma)/\gamma}e^{-\frac{1-\gamma}{\gamma}h(\omega_t)-(1-\gamma)\omega_t}e^{(1-\gamma)\omega_t}\, d\gamma\\ &= \int_1^\infty \frac{1}{1-\gamma}f(\gamma)^{1/\gamma}e^{-\frac{1-\gamma}{\gamma}h(\omega_t)}\, d\gamma,\end{aligned}$$

which ends the proof. □

The optimal allocation of consumption to a given individual is a fraction of aggregate endowment, a fraction depending on the state of the economy and the relative risk aversion of the individual.

The next lemma summarizes some properties of the function h which will be useful in the following discussion.

Lemma 8.1 *The function h defined in* (8.43) *is decreasing and convex with*

$$h'(\omega) = -\left(\int_1^\infty \frac{1}{\gamma} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\, d\gamma\right)^{-1} < -1. \tag{8.45}$$

Proof: Differentiating with respect to ω_t in (8.43), we get

$$\int_1^\infty f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega_t)-\omega_t}\left(-\frac{1}{\gamma}h'(\omega_t) - 1\right) d\gamma = 0,$$

which implies that

$$h'(\omega_t)\int_1^\infty \frac{1}{\gamma} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega_t)-\omega_t}\, d\gamma = -\int_1^\infty f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega_t)-\omega_t}\, d\gamma = -1,$$

from which the expression for $h'(\omega_t)$ follows. Since we are integrating over $\gamma \geq 1$,

$$\int_1^\infty \frac{1}{\gamma} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega_t)-\omega_t}\, d\gamma < \int_1^\infty f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega_t)-\omega_t}\, d\gamma = 1,$$

which gives the upper bound on $h'(\omega_t)$.

Convexity means $h''(\omega_t) \geq 0$ and by differentiating (8.45) we see that this is true if and only if

$$-h'(\omega)\int_1^\infty \frac{1}{\gamma^2} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\, d\gamma \geq \int_1^\infty \frac{1}{\gamma} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\, d\gamma,$$

i.e. if and only if

$$\int_1^\infty \frac{1}{\gamma^2} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\, d\gamma \geq \left(\int_1^\infty \frac{1}{\gamma} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\, d\gamma\right)^2.$$

In order to show this we apply the Cauchy-Schwartz inequality for integrals,

$$\left(\int_a^b F(x)G(x)\, dx\right)^2 \leq \left(\int_a^b F(x)^2\, dx\right)\left(\int_a^b G(x)^2\, dx\right),$$

with $x = \gamma$, $a = 1$, $b = \infty$, and

$$F(x) = \frac{1}{\gamma}\left(f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\right)^{1/2}, \quad G(x) = \left(f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\right)^{1/2}.$$

By the Cauchy-Schwartz inequality and (8.43), we get exactly

$$\begin{aligned}\left(\int_1^\infty \frac{1}{\gamma} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\, d\gamma\right)^2 &\leq \left(\int_1^\infty \frac{1}{\gamma^2} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\, d\gamma\right)\left(\int_1^\infty f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\, d\gamma\right) \\ &= \int_1^\infty \frac{1}{\gamma^2} f(\gamma)^{1/\gamma} e^{-\frac{1}{\gamma}h(\omega)-\omega}\, d\gamma\end{aligned}$$

as was to be shown. □

Using (8.43) and (8.45), straightforward differentiation of the utility function (8.44) gives that

$$U_Y(Y_t, X_t) = X_t^{-1} e^{h(\omega_t)}, \tag{8.46}$$

$$U_{YY}(Y_t, X_t) = h'(\omega_t) Y_t^{-1} U_Y(Y_t, X_t). \tag{8.47}$$

The relative risk aversion of the representative individual is therefore

$$-\frac{Y_t U_{YY}(Y_t, X_t)}{U_Y(Y_t, X_t)} = -h'(\omega_t). \tag{8.48}$$

It follows from Lemma 8.1 that this relative risk aversion is greater than 1 and decreasing in the state variable ω_t. The intuition is that the individuals with low relative risk aversion will other things equal invest more in the risky asset—and their wealth will therefore fluctuate more—than individuals with high relative risk aversion. In good states a larger fraction of the aggregate resources will be held by individuals with low risk aversion than in bad states. The relative risk aversion of the representative individual, which is some sort of wealth-weighted average of the individual risk aversions, will therefore be lower in good states than in bad states. Aggregate relative risk aversion is counter-cyclical.

We can obtain the equilibrium risk-free rate and the market price of risk from Theorem 8.2. In the present model the dynamics of $X_t = e^{x_t}$ will be

$$dX_t = \kappa X_t \left(y_t - x_t\right) \, dt = \kappa X_t \omega_t \, dt, \tag{8.49}$$

which is locally insensitive to shocks corresponding to $\boldsymbol{\sigma}_{Xt} = \mathbf{0}$. This will simplify the formulas for the risk-free rate and the Sharpe ratio. Moreover,

$$U_{YX}(Y_t, X_t) = -X_t^{-1}(1 + h'(\omega_t)) U_Y(Y_t, X_t), \tag{8.50}$$

$$Y_t^2 U_{YYY}(Y_t, X_t) = \left(h''(\omega_t) + h'(\omega_t)^2 - h'(\omega_t)\right) U_Y(Y_t, X_t). \tag{8.51}$$

We arrive at the following conclusion:

Theorem 8.4 *In the Chan and Kogan model the risk-free short-term interest rate is*

$$r_t^f = \delta - h'(\omega_t)\kappa(\bar{\omega} - \omega_t) + \kappa\omega_t - \frac{1}{2}\sigma^2 \left(h''(\omega_t) + h'(\omega_t)^2\right). \tag{8.52}$$

and the Sharpe ratio of a risky asset is

$$\frac{\mu_t + \delta_t - r_t^f}{\sigma_t} = -h'(\omega_t)\sigma, \tag{8.53}$$

which is decreasing in the state variable ω_t.

In Exercise 8.4 you are asked to provide the details. In Exercise 8.5 you are asked to show how (8.53) follows from (8.39).

The Sharpe ratio in the model varies counter-cyclically. In good states of the economy the average relative risk aversion is relatively low and the risk premium necessary for markets clear is therefore also low. Conversely in bad states where the average relative risk aversion is high.

8.7.5 Durable goods

See Lustig and van Nieuwerburgh (2005), Piazzesi, Schneider, and Tuzel (2006), Yogo (2006). See Exercise 8.10.

8.8 Consumption-based asset pricing with incomplete markets

8.8.1 Evidence of incomplete markets

Some empirical asset pricing studies indicate that markets are incomplete so that a representative individual may be non-existing. The marginal rate of substitution of each individual defines a state-price deflator. It follows from Theorem 4.3 that a weighted average of state-price deflators is also a state-price deflator. Brav, Constantinides, and Geczy (2002) assume that all individuals have time-additive CRRA utility with the same time preference rate δ and the same relative risk aversion γ so that the state-price deflator for individual l is $e^{-\delta t}\left(c_{lt}/c_{l0}\right)^{-\gamma}$. An equally-weighted average over L individuals gives the state-price deflator

$$\zeta_t = e^{-\delta t}\frac{1}{L}\sum_{l=1}^{L}\left(\frac{c_{lt}}{c_{l0}}\right)^{-\gamma}.$$

Using data on the consumption of individual households, the authors find that this state-price deflator is consistent with the historical excess returns on the U.S. stock market for a risk aversion parameter as low as 3. If the market were complete, consumption growth c_{lt}/c_{l0} would be the same for all individuals under these assumptions, and

$$\zeta_t = e^{-\delta t}\left(\frac{\sum_{l=1}^{L} c_{lt}}{\sum_{l=1}^{L} c_{l0}}\right)^{-\gamma} \tag{8.54}$$

would be a valid state-price deflator. Summing up over all individuals in this formula, we get the state-price deflator for a representative individual, who will also have relative risk aversion equal to γ, but this is inconsistent with data except for unreasonably high values of γ. This study therefore indicates that financial markets are incomplete and do not allow individuals to align their marginal rates of substitution.

For various reasons, a large, but declining, fraction of individuals do not invest in stock markets at all or only to a very limited extent. Brav, Constantinides, and Geczy (2002) show that if in Equation (8.54) you only sum up over individuals holding financial assets with a value higher than some threshold, this state-price deflator is consistent with historical data for a relative risk aversion which is relatively high, but much lower than the required risk aversion using aggregate consumption. The higher the threshold, the lower the required risk aversion. This result reflects that only the individuals active in the financial markets contribute to the setting of prices. Other empirical studies report similar findings, and Başak and Cuoco (1998) set up a formal asset pricing model that explicitly distinguishes between individuals owning stocks and individuals not owning stocks.

8.8.2 Labor income risk

The labor income of individuals is not fully insurable, neither through investments in financial assets nor through existing insurance contracts. A number of papers investigate how non-hedgeable income shocks may affect the pricing of financial assets. If unexpected changes in labor income are temporary, individuals may self-insure by building up a buffer of savings in order to even out the consumption effects of the income shocks over the entire life. The effects on equilibrium asset prices will be insignificant, cf. Telmer (1993). If income shocks are to help resolve the equity premium puzzle, the shocks have to affect income beyond the current period. In addition, the magnitude of the income shocks has to be negatively related to the level of stock prices. Both the persistency of income shocks and the counter-cyclical variation in income seem reasonable in light of the risk of lay-offs and find empirical support, cf. Storesletten, Telmer, and Yaron (2004). Individuals facing such an income uncertainty will demand higher risk premia on stocks than in a model without labor income because stocks do badly exactly when individuals face the highest risk of an unexpected decline in income.

In a model where all individuals have time-additive CRRA utility, Constantinides and Duffie (1996) show by construction that if the income processes of different individuals are sufficiently different in a certain sense then their model with any given risk aversion can generate basically any pattern in aggregate consumption and financial prices, including the puzzling historical pattern. However, Cochrane (2001, Ch. 21) argues that with a realistic degree of cross-sectional variation in individual labor income, a relatively high value of the risk aversion parameter is still needed to match historical data. Nevertheless, it seems to be important to incorporate the labor income uncertainty of individuals and in particular the difference between the labor income processes of different individuals in the development of better asset pricing models.

Constantinides, Donaldson, and Mehra (2002) emphasize that individuals choose consumption and investment from a life-cycle perspective and face different opportunities and risks at different ages. They divide individuals into young, middle-aged, and old individuals. Old individuals only consume the savings they have build up earlier, do not receive further income and do not invest in financial assets. Young individuals typically have a low financial wealth and a low current labor income but a high human capital (the present value of their future labor income). Many young individuals would prefer borrowing significant amounts both in order to smooth out consumption over life and to be able to invest in the stock market to generate additional consumption opportunities and obtain the optimal risk/return trade-off. The empirically observed low correlation between labor income and stock returns makes stock investments even more attractive for young individuals. Unfortunately it is difficult to borrow significant amounts just because you expect to get high future income and, therefore, the young individuals can only invest very little, if anything, in the stock market. Middle-aged individuals face a different situation. Their future labor income is limited and relatively certain. Their future consumption opportunities are primarily depending on the return on their investments. For middle-aged individuals the correlation between stock returns and future consumption is thus quite high so stocks are not as attractive for them. Nevertheless, due to the borrowing constraints on the young individuals, the stocks have to be owned by the middle-aged individuals and, hence, the equilibrium expected rate of return has to be quite high. The authors set up a relatively simple model formalizing these thoughts, and the model is able

to explain an equity premium significantly higher than in the standard model, but still below the historically observed premium.

8.9 Concluding remarks

This chapter has developed consumption-based models of asset pricing. Under very weak assumptions, expected excess returns on risky asset will be closely related to the covariance of the asset return with aggregate consumption, giving a conditional consumption-based CAPM. The simple version of the model is unable to match a number of features of consumption and return data, which leads to several asset pricing puzzles. However, a number of relatively recent theoretical and empirical studies identify extensions of the simple model that are able to eliminate (or at least reduce the magnitude of) several of these puzzles. These extensions include state-dependent preferences, heterogenous risk aversion, and labor income risk. The consumption-based asset pricing framework is alive and kicking. A potential problem of the tests and practical implementation of such models, however, is the need for good data on individual or aggregate consumption. The next chapter discusses asset pricing models not relying on consumption data.

8.10 Exercises

EXERCISE 8.1 Carl Smart is currently (at time $t = 0$) considering a couple of investment projects that will provide him with a dividend in one year from now (time $t = 1$) and a dividend in two years from now (time $t = 2$). He figures that the size of the dividends will depend on the growth rate of aggregate consumption over these two years. The first project Carl considers provides a dividend of

$$D_t = 60t + 5(C_t - \mathrm{E}[C_t])$$

at $t = 1$ and at $t = 2$. The second project provides a dividend of

$$D_t = 60t - 5(C_t - \mathrm{E}[C_t])$$

at $t = 1$ and at $t = 2$. Here E[ ] is the expectation computed at time 0 and C_t denotes aggregate consumption at time t. The current level of aggregate consumption is $C_0 = 1000$.
As a valuable input to his investment decision Carl wants to compute the present value of the future dividends of each of the two projects.
First, Carl computes the present values of the two projects using a "risk-ignoring approach", i.e. by discounting the expected dividends using the riskless returns observed in the bond markets. Carl observes that a one-year zero-coupon bond with a face value of 1000 currently trades at a price of 960 and a two-year zero-coupon bond with a face value of 1000 trades at a price of 929.02.

(a) What are the expected dividends of project 1 and project 2?

(b) What are the present values of project 1 and project 2 using the risk-ignoring approach?

Suddenly, Carl remembers that he once took a great course on advanced asset pricing and that the present value of an uncertain dividend of D_1 at time 1 and an uncertain dividend of D_2 at time 2

should be computed as

$$P = \mathrm{E}\left[\frac{\zeta_1}{\zeta_0}D_1 + \frac{\zeta_2}{\zeta_0}D_2\right],$$

where the ζ_t's define a state-price deflator. After some reflection and analysis, Carl decides to value the projects using a conditional Consumption-based CAPM so that the state-price deflator between time t and $t+1$ is of the form

$$\frac{\zeta_{t+1}}{\zeta_t} = a_t + b_t\frac{C_{t+1}}{C_t}, \quad t = 0, 1, \ldots.$$

Carl thinks it's fair to assume that aggregate consumption will grow by either 1% ("low") or 3% ("high") in each of the next two years. Over the first year he believes the two growth rates are equally likely. If the growth rate of aggregate consumption is high in the first year, he believes that there is a 30% chance that it will also be high in the second year and, thus, a 70% chance of low growth in the second year. On the other hand, if the growth rate of aggregate consumption is low in the first year, he estimates that there will be a 70% chance of high growth and a 30% chance of low growth in the second year.

In order to apply the Consumption-based CAPM for valuing his investment projects, Carl has to identify the coefficients a_t and b_t for $t = 0, 1$. The values of a_0 and b_0 can be identified from the prices of two traded assets that only provide dividends at time 1. In addition to the one-year zero-coupon bond mentioned above, a one-year European call option on aggregate consumption is traded. The option has a strike price of $K = 1020$ so that the payoff of the option in one year is $\max(C_1 - 1020, 0)$, where C_1 is the aggregate consumption level at $t = 1$. The option trades at a price of 4.7.

(c) Using the information on the two traded assets, write up two equations that can be used for determining a_0 and b_0. Verify that the equations are solved for $a_0 = 3$ and $b_0 = -2$.

The values of a_1 and b_1 may depend on the consumption growth rate of the first year, i.e. Carl has to find a_1^h, b_1^h that defines the second-year state-price deflator if the first-year growth rate was high and a_1^l, b_1^l that defines the second-year state-price deflator if the first-year growth rate was low. Using the observed market prices of other assets, Carl concludes that

$$a_1^h = 2.5, \quad b_1^h = -1.5, \quad a_1^l = 3.5198, \quad b_1^l = -2.5.$$

(d) Verify that the economy is path-independent in the sense that the current price of an asset that will pay you 1 at $t = 2$ if the growth rate is high in the first year and low in the second year will be identical (at least to five decimal places) to the current price of an asset that will pay you 1 at $t = 2$ if the growth rate is low in the first year and high in the second year.

(e) Illustrate the possible dividends of the two projects in a two-period binomial tree.

(f) What are the correctly computed present values of the two projects?

(g) Carl notes that the expected dividends of the two projects are exactly the same but the present value of project 2 is higher than the present value of project 1. Although Carl is pretty smart, he cannot really figure out why this is so. Can you explain it to him?

EXERCISE 8.2 In the simple consumption-based asset pricing model, the growth rate of aggregate consumption is assumed to have a constant expectation and standard deviation (volatility). For example, in the continuous-time version aggregate consumption is assumed to follow a geometric Brownian motion. Consider the following alternative process for aggregate consumption:

$$dc_t = c_t[\mu\, dt + \sigma c_t^{\alpha-1}\, dz_t],$$

where μ, σ, and α are positive constants, and $z = (z_t)$ is a standard Brownian motion. As in the simple model, assume that a representative individual exists and that this individual has time-additive expected utility exhibiting constant relative risk aversion given by the parameter $\gamma > 0$ and a constant time preference rate $\delta > 0$.

(a) State an equation linking the expected excess return on an arbitrary risky asset to the level of aggregate consumption and the parameters of the aggregate consumption process. How does the expected excess return vary with the consumption level?

(b) State an equation linking the short-term continuously compounded risk-free interest rate r_t^f to the level of aggregate consumption and the parameters of the aggregate consumption process. How does the interest rate vary with the consumption level?

(c) Use Itô's Lemma to find the dynamics of the interest rate, dr_t^f? Can you write the drift and the volatility of the interest rate as functions of the interest rate level only?

EXERCISE 8.3 Give a proof of Theorem 8.2.

EXERCISE 8.4 Consider the Chan and Kogan model. Show the expressions in (8.46), (8.47), (8.49), (8.50), and (8.51). Show Theorem 8.4.

EXERCISE 8.5 In the Chan and Kogan model, show how (8.53) follows from (8.39).

EXERCISE 8.6 Consider a continuous-time economy with complete markets and a representative individual having an "external habit" or "keeping up with the Jones'es" utility function so that, at any time t, the individual maximizes $\mathrm{E}_t\left[\int_t^T e^{-\delta(s-t)}u(c_s, X_s)\, ds\right]$, where $u(c, X) = \frac{1}{1-\gamma}(c - X)^{1-\gamma}$ for $c > X \geq 0$.
Define $Y_t = -\ln\left(1 - \frac{X_t}{c_t}\right)$.

(a) Argue that Y_t is positive. Would you call a situation where Y_t is high for a "good state" or a "bad state"? Explain!

(b) Argue that the unique state-price deflator is given by

$$\zeta_t = e^{-\delta t}\, \frac{c_t^{-\gamma} e^{\gamma Y_t}}{c_0^{-\gamma} e^{\gamma Y_0}}.$$

First, write the dynamics of consumption and the variable Y_t in the general way:

$$dc_t = c_t[\mu_{ct}\,dt + \boldsymbol{\sigma}_{ct}^\top\,d\boldsymbol{z}_t],$$
$$dY_t = \mu_{Yt}\,dt + \boldsymbol{\sigma}_{Yt}^\top\,d\boldsymbol{z}_t,$$

where $\boldsymbol{z} = (\boldsymbol{z}_t)$ is a multi-dimensional standard Brownian motion.

(c) Use Itô's Lemma to find the dynamics of the benchmark X_t. State the drift and sensitivity in terms of c_t and X_t (no Y_t, please).

(d) Use Itô's Lemma to find the dynamics of the state-price deflator and identify the continuously compounded short-term risk-free interest rate r_t^f and the market price of risk $\boldsymbol{\lambda}_t$.

An asset i pays an uncertain terminal dividend but no intermediate dividends. The price dynamics is of the form

$$dP_{it} = P_{it}\left[\mu_{it}\,dt + \boldsymbol{\sigma}_{it}^\top\,d\boldsymbol{z}_t\right].$$

(e) Explain why

$$\mu_{it} - r_t^f = \beta_{ict}\eta_{ct} + \beta_{iYt}\eta_{Yt},$$

where $\beta_{ict} = (\boldsymbol{\sigma}_{it}^\top\boldsymbol{\sigma}_{ct})/\|\boldsymbol{\sigma}_{ct}\|^2$ and $\beta_{iYt} = (\boldsymbol{\sigma}_{it}^\top\boldsymbol{\sigma}_{Yt})/\|\boldsymbol{\sigma}_{Yt}\|^2$. Express η_{ct} and η_{Yt} in terms of previously introduced parameters and variables.

Next, consider the specific model:

$$dc_t = c_t[\mu_c\,dt + \sigma_c\,dz_{1t}],$$
$$dY_t = \kappa[\bar{Y} - Y_t]\,dt + \sigma_Y\sqrt{Y_t}\left(\rho\,dz_{1t} + \sqrt{1-\rho^2}\,dz_{2t}\right),$$

where $(z_1, z_2)^\top$ is a two-dimensional standard Brownian motion, μ_c, σ_c, κ, $\bar{Y}$, and σ_Y are positive constants, and $\rho \in (-1, 1)$.

(f) What is the short-term risk-free interest rate and the market price of risk in the specific model?

Assume that the price dynamics of asset i is

$$dP_{it} = P_{it}\left[\mu_{it}\,dt + \sigma_{it}\left(\psi\,dz_{1t} + \sqrt{1-\psi^2}\,dz_{2t}\right)\right],$$

where $\sigma_{it} > 0$ and $\psi \in (-1, 1)$.

(g) What is the Sharpe ratio of asset i in the specific model? Can the specific model generate counter-cyclical variation in Sharpe ratios (if necessary, provide parameter conditions ensuring this)?

EXERCISE 8.7 Consider the set-up of Exercise 6.10 with $u(c, h) = \frac{1}{1-\gamma}(c-h)^{1-\gamma}$.

(a) Can optimal consumption follow a geometric Brownian motion under these assumptions?

(b) Assume that the *excess* consumption rate $\hat{c}_t = c_t - h_t$ follows a geometric Brownian motion. Show that the state-price deflator will be of the form $\zeta_t = f(t)e^{-\delta t}\hat{c}_t^{-\gamma}$ for some deterministic function $f(t)$ (and find that function). Find an expression for the equilibrium interest rate and the market price of risk. Compare with the "simple model" with no habit, CRRA utility, and consumption following a geometric Brownian motion.

EXERCISE 8.8 (This problem is based on the working paper Menzly, Santos, and Veronesi (2002).) Consider an economy with a representative agent with life-time utility given by

$$U(C) = \mathrm{E}\left[\int_0^\infty e^{-\varphi t}\ln(C_t - X_t)\,dt\right],$$

where X_t is an external habit level and φ is a subjective discount rate. As in Campbell and Cochrane (1999) define the surplus ratio as $S_t = (C_t - X_t)/C_t$. Define $Y_t = 1/S_t$. Aggregate consumption C_t is assumed to follow a geometric Brownian motion

$$dC_t = C_t\left[\mu_C\,dt + \sigma_C\,dz_t\right],$$

where μ_C and σ_C are constants and z is a one-dimensional standard Brownian motion. The dynamics of the habit level is modeled through

$$dY_t = k[\bar{Y} - Y_t]\,dt - \alpha(Y_t - \kappa)\sigma_C\,dz_t,$$

where $k, \bar{Y}, \alpha$, and κ are constants.

(a) Show that $\mathrm{E}_t[Y_\tau] = \bar{Y} + (Y_t - \bar{Y})e^{-k(\tau-t)}$.

(b) For any given dividend process $D = (D_t)$ in this economy, argue that the price is given by

$$P_t^D = (C_t - X_t)\,\mathrm{E}_t\left[\int_t^\infty e^{-\varphi(\tau-t)}\frac{D_\tau}{C_\tau - X_\tau}\,d\tau\right].$$

Let $s_\tau^D = D_\tau/C_\tau$ denote the dividend's share of aggregate consumption. Show that the price P_t^D satisfies

$$\frac{P_t^D}{C_t} = \frac{1}{Y_t}\,\mathrm{E}_t\left[\int_t^\infty e^{-\varphi(\tau-t)}s_\tau^D Y_\tau\,d\tau\right].$$

(c) Show that the price P_t^C of a claim to the aggregate consumption stream $D_\tau = C_\tau$ is given by

$$\frac{P_t^C}{C_t} = \frac{1}{\varphi + k}\left(1 + \frac{k\bar{Y}}{\varphi}S_t\right).$$

(d) Find the dynamics of the state-price deflator. Find and interpret expressions for the risk-free interest rate and the market price of risk in this economy.

EXERCISE 8.9 Consider a continuous-time model of an economy with a representative agent and a single non-durable good. The objective of the agent at any time t is to maximize the expected time-additive CRRA utility,

$$\mathrm{E}_t\left[\int_t^\infty e^{-\delta(s-t)}\frac{C_s^{1-\gamma}}{1-\gamma}\,ds\right],$$

where $\gamma > 0$ and C_s denotes the consumption rate at time s. The agent can invest in a bank account, i.e. borrow and lend at a short-term interest rate of r_t. The bank account is in zero net supply. A single stock with a net supply of one share is available for trade. The agent is initially endowed with this share. The stock pays a continuous dividend at the rate D_t. The agent receives an exogenously given labor income at the rate I_t.

(a) Explain why the equilibrium consumption rate must equal the sum of the dividend rate and the labor income rate, i.e. $C_t = I_t + D_t$.

Let F_t denote the dividend-consumption ratio, i.e. $F_t = D_t/C_t$. Assume that $F_t = \exp\{-X_t\}$, where $X = (X_t)$ is the diffusion process

$$dX_t = (\mu - \kappa X_t)\,dt - \eta\sqrt{X_t}\,dz_{1t}.$$

Here μ, κ, and η are positive constants and $z_1 = (z_{1t})$ is a standard Brownian motion.

(b) Explain why F_t is always between zero and one.

Assume that the aggregate consumption process is given by the dynamics

$$dC_t = C_t\left[\alpha\,dt + \sigma\sqrt{X_t}\rho\,dz_{1t} + \sigma\sqrt{X_t}\sqrt{1-\rho^2}\,dz_{2t}\right],$$

where α, σ, and ρ are constants and $z_2 = (z_{2t})$ is another standard Brownian motion, independent of z_1.

(c) What is the equilibrium short-term interest rate in this economy?

(d) Use Itô's Lemma to derive the dynamics of F_t and of D_t.

(e) Show that the volatility of the dividend rate is greater than the volatility of the consumption rate if $\eta\rho > 0$.

Let P_t denote the price of the stock, i.e. the present value of all the future dividends.

(f) Show that the stock price can be written as

$$P_t = C_t^\gamma\,\mathrm{E}_t\left[\int_t^\infty e^{-\delta(s-t)}C_s^{1-\gamma}F_s\,ds\right].$$

It can be shown that P_t can be written as a function of t, C_t, and F_t:

$$P_t = C_t\int_t^\infty e^{-\delta(s-t)}A(t,s)F_t^{-B(t,s)}\,ds.$$

Here $A(t,s)$ and $B(t,s)$ are some deterministic functions of time that we will leave unspecified.

(g) Use Itô's Lemma to show that

$$dP_t = P_t \left[\ldots \, dt + (\rho\sigma + \eta H_t) \sqrt{X_t} \, dz_{1t} + \sigma\sqrt{1-\rho^2}\sqrt{X_t} \, dz_{2t} \right],$$

where the drift term is left out (you do not have to compute the drift term!) and where

$$H_t = \frac{-\int_t^\infty e^{-\delta(s-t)} A(t,s) B(t,s) F_t^{-B(t,s)} \, ds}{\int_t^\infty e^{-\delta(s-t)} A(t,s) F_t^{-B(t,s)} \, ds}.$$

(h) Show that the expected excess rate of return on the stock at time t can be written as

$$\psi_t = \gamma X_t \left(\sigma^2 + \sigma\rho\eta H_t \right)$$

and as

$$\psi_t = \gamma\sigma_{Ct}^2 + \gamma\rho H_t \sigma_{Ct}\sigma_{Ft},$$

where σ_{Ct} and σ_{Ft} denote the percentage volatility of C_t and F_t, respectively.

(i) What would the expected excess rate of return on the stock be if the dividend-consumption ratio was deterministic? Explain why the model with stochastic dividend-consumption ratio has the potential to resolve (at least partially) the equity premium puzzle.

EXERCISE 8.10 Consider a discrete-time representative individual economy with preferences $\mathrm{E}[\sum_{t=0}^\infty \beta^t u(C_t, D_t^*)]$, where C_t is the consumption of perishable goods and D_t^* is the consumption of services from durable goods. Assume a Cobb-Douglas type utility function,

$$u(C, D^*) = \frac{1}{1-\gamma} \left(C^\alpha (D^*)^{1-\alpha} \right)^{1-\gamma}$$

where $\gamma > 0$ and $\alpha \in [0, 1]$.

(a) Write up the one-period state-price deflator discount factor $M_{t+1} \equiv \frac{\zeta_{t+1}}{\zeta_t} = \beta \frac{u_C(C_{t+1}, D_{t+1}^*)}{u_C(C_t, D_t^*)}$ in this economy.

Assume that the services from the the durable goods are given by $D_t^* = \theta(K_{t-1} + D_t)$, where K_{t-1} is the stock of the durable entering period t and D_t is the additional purchases of the durable good in period t. We can interpret θ as a depreciation rate or as the intensity of usage of the durable. We must have $K_t = (1-\theta)(K_{t-1} + D_t)$.

(b) Argue that if one additional unit of the durable is purchased in period t, then the additional services from the durable in period $t+j$ will be $\theta(1-\theta)^j$.

(c) What is the marginal (life-time expected) utility, MU_t^D, from purchasing an extra unit of the durable in period t?

(d) If p_t is the unit price of the durable good, argue that $\mathrm{MU}_t^D = p_t \alpha C_t^{\alpha(1-\gamma)-1} (D_t^*)^{(1-\alpha)(1-\gamma)}$ in equilibrium.

Now we allow for a stochastic intensity of usage, i.e. $\theta = (\theta_t)$ is a stochastic process. Define $X_t = (K_{t-1} + D_t)/K_{t-1}$, $x_t = \ln X_t$, $c_t = \ln C_t$, and $m_t = \ln M_t$.

(e) Show that $m_{t+1} = \ln\beta + (\alpha(1-\gamma)-1)\Delta c_{t+1} + (1-\alpha)(1-\gamma)\left[\Delta(\ln\theta_{t+1}) + x_{t+1} + \ln(1-\theta_t)\right]$.

Assume now that

$$\begin{aligned}
\Delta c_{t+1} &= g + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N(0, \sigma_\varepsilon^2), \\
\ln\theta_{t+1} &= h + \varphi \ln\theta_t + \nu_{t+1}, \quad \nu_{t+1} \sim N(0, \sigma_\nu^2), \\
x_{t+1} &= (1 - w - \varphi)\ln\theta_t - \ln(1-\theta_t) - h + a\nu_{t+1},
\end{aligned}$$

and that the two exogenous shocks ε_{t+1} and ν_{t+1} have correlation ρ.

(f) Derive the continuously compounded equilibrium short interest rate $r_t^f = \ln R_t^f$. You should find that the interest rate is constant if $w = 0$.

EXERCISE 8.11 Consider a continuous-time economy with a representative agent with time-additive subsistence HARA utility, i.e. the objective of the agent is to maximize

$$\mathrm{E}\left[\int_0^T e^{-\delta t}\frac{1}{1-\gamma}\left(c_t - \bar{c}\right)^{1-\gamma}\,dt\right],$$

where $\bar{c} \geq 0$ is the subsistence consumption level. Assume that aggregate consumption $c = (c_t)$ evolves as

$$dc_t = \mu c_t\,dt + \sigma\sqrt{c_t(c_t - \bar{c})}\,dz_t,$$

where $z = (z_t)$ is a (one-dimensional) standard Brownian motion. Find the equilibrium short-term interest rate r_t^f and the market price of risk λ_t, expressed in terms of c_t and the parameters introduced above. How do r_t^f and λ_t depend on the consumption level? Are r_t^f and λ_t higher or lower or unchanged relative to the standard case in which $\bar{c} = 0$?

Chapter 9

Factor models

9.1 Introduction

The lack of reliable consumption data discussed in Section 8.6 complicates tests and applications of the consumption-based models. As mentioned above, most tests that have been carried out find it problematic to match the (simple) consumption-based model and historical return and consumption data (of poor quality). This motivates a search for models linking returns to other factors than consumption.

The classical CAPM is the Mother of all factor models. It links expected excess returns (on stocks) to the return on the market portfolio (of stocks). It was originally developed in a one-period framework but can be generalized to multi-period settings. The model is based on rather unrealistic assumptions and the empirical success of the CAPM is modest.

Many, many papers have tried to identify factor models that perform better, mostly by adding extra factors. However, this should only be done with extreme care. In a given data set of historical returns it is always possible to find a return that works as a pricing factor, as already indicated in Chapter 4. In fact, any ex-post mean-variance efficient return will work. On the other hand, there is generally no reason to believe that the same return will work as a pricing factor in the future. Factors should be justified by economic arguments or even a theoretical asset pricing model.

It is worth emphasizing that the general theoretical results of the consumption-based asset pricing framework are not challenged by factor models. The problem with the consumption-based models is the implementation. Factor models do not invalidate the consumption-based asset pricing framework but are special cases that may be easier to apply and test. Therefore factors should generally help explain typical individuals' marginal utilities of consumption.

Section 9.2 reviews the classical one-period CAPM and how it fits into the modern consumption-based asset pricing framework. Section 9.3 defines and studies pricing factors in the one-period setting. In particular, pricing factors are linked to state-price deflators. The relation between mean-variance efficient returns and pricing factors is the topic of Section 9.4. Multi-period pricing factors are introduced in Section 9.5 with a discussion of the distinction between conditional and unconditional pricing factors. Section 9.6 offers a brief introduction to empirical studies of factor models. Finally, Section 9.7 discusses how pricing factors can be derived theoretically. It also derives an intertemporal version of the CAPM.

9.2 The classical one-period CAPM

The classical CAPM developed by Sharpe (1964), Lintner (1965), and Mossin (1966) says that return on the market portfolio is a pricing factor so that

$$\mathrm{E}[R_i] = \alpha + \beta[R_i, R_M]\left(\mathrm{E}[R_M] - \alpha\right), \quad i = 1, 2, \ldots, I, \tag{9.1}$$

for some zero-beta return α, which is identical to the risk-free rate if such exists. Here the market-beta is defined as $\beta[R_i, R_M] = \mathrm{Cov}[R_i, R_M]/\mathrm{Var}[R_M]$.

The classical CAPM is usually derived from mean-variance analysis. If all individuals have quadratic utility or returns are normally distributed, any individual will optimally pick a mean-variance efficient portfolio. If R_l denotes the return on the portfolio chosen by individual l and w_l denotes individual l's share of total wealth, the return on the market portfolio will be $R_M = \sum_{l=1}^{L} w_l R_l$ and the market portfolio will be mean-variance efficient. As was already stated in Theorem 4.6 (demonstrated later in this chapter), the return on any mean-variance efficient portfolio will satisfy an equation like (9.1). In particular, this is true for the market portfolio when it is efficient.

To see how the classical CAPM fits into the consumption-based asset pricing framework, consider a model in which all individuals have time-additive expected utility and are endowed with some time 0 wealth but receive no time 1 income from non-financial sources. Let us consider an arbitrary individual with initial wealth endowment e_0. If the individual consumes c_0 at time 0 she will invest $e_0 - c_0$ in the financial assets. Representing the investment by the portfolio weight vector $\boldsymbol{\pi}$, the gross return on the portfolio will be $R^{\boldsymbol{\pi}} = \boldsymbol{\pi} \cdot \boldsymbol{R} = \sum_{i=1}^{I} \pi_i R_i$. The time 1 consumption will equal the total dividend of the portfolio, which is the gross return multiplied by the initial investment, i.e.

$$c = R^{\boldsymbol{\pi}}(e_0 - c_0).$$

We can substitute this into the marginal rate of substitution of the individual so that the associated state-price deflator becomes

$$\zeta = e^{-\delta}\frac{u'(c)}{u'(c_0)} = e^{-\delta}\frac{u'\left(R^{\boldsymbol{\pi}}(e_0 - c_0)\right)}{u'(c_0)}. \tag{9.2}$$

If the economy has a representative individual, she has to own all the assets, i.e. she has to invest in the market portfolio. We can then replace $R^{\boldsymbol{\pi}}$ by R_M, the gross return on the market portfolio. We can obtain the classical CAPM from this relation if we either assume that the utility function is quadratic or that the return on the market portfolio is normally distributed.

Quadratic utility. The quadratic utility function $u(c) = -(\bar{c} - c)^2$ is a special case of the satiation HARA utility functions. Marginal utility $u'(c) = 2(\bar{c} - c)$ is positive for $c < \bar{c}$ so that consumption in excess of $\bar{c}$ will decrease utility. Another problem is that the absolute risk aversion $\mathrm{ARA}(c) = 1/(\bar{c} - c)$ is increasing in the level of consumption. For quadratic utility, Eq. (9.2) becomes

$$\zeta = e^{-\delta}\frac{\bar{c} - R^{\boldsymbol{\pi}}(e_0 - c_0)}{\bar{c} - c_0} = e^{-\delta}\frac{\bar{c}}{\bar{c} - c_0} - e^{-\delta}\frac{e_0 - c_0}{\bar{c} - c_0}R^{\boldsymbol{\pi}}, \tag{9.3}$$

which is affine in the portfolio return. It now follows from the discussion in Section 4.5.2 (also see later section in the present chapter) that the portfolio return is a pricing factor so that

$$\mathrm{E}[R_i] = \alpha + \beta[R_i, R^{\boldsymbol{\pi}}]\left(\mathrm{E}[R^{\boldsymbol{\pi}}] - \alpha\right).$$

Again, if this applies to a representative individual, we can replace $R^{\boldsymbol{\pi}}$ by the market portfolio return R_M and we have the classical CAPM.

For the quadratic utility function the absolute risk tolerance is $\mathrm{ART}(c) = -c+\bar{c}$. If all individuals have quadratic utility functions (possibly with different $\bar{c}$'s) and we assume that a risk-free asset is traded and all time 1 endowments are spanned by traded assets, Theorem 7.6 implies that the optimal consumption of any individual is affine in the aggregate endowment and therefore it can be implemented by investing in a portfolio of the risk-free asset and the market portfolio of all assets. The return on the portfolio will be a weighted average of the risk-free return and the market return, $R^{\boldsymbol{\pi}} = w_f R^f + (1 - w_f)R_M$, and substituting this into (9.3) we see that the state-price deflator associated with any given individual will be affine in R_M. Again, Theorem 9.3 will then give us the classical CAPM.

Normally distributed returns. We will show that for almost any utility function we can derive the classical CAPM relation if returns are jointly normally distributed. We need the following result called Stein's Lemma:

Lemma 9.1 (Stein's Lemma) *If x and y are jointly normally distributed random variables and $g : \mathbb{R} \to \mathbb{R}$ is a differentiable function with $\mathrm{E}[|g'(y)|] < \infty$, then*

$$\mathrm{Cov}[x, g(y)] = \mathrm{E}[g'(y)]\,\mathrm{Cov}[x, y].$$

Proof: Define the random variable ε by $\varepsilon = x - \alpha - \beta y$, where $\beta = \mathrm{Cov}[x, y]/\,\mathrm{Var}[y]$, $\alpha = \mathrm{E}[x] - \beta\,\mathrm{E}[y]$, and $\mathrm{Cov}[\varepsilon, y] = 0$. Since ε and y are jointly normally distributed, the fact that they are uncorrelated implies that they will be independent. It follows that $\mathrm{Cov}[\varepsilon, g(y)] = 0$ for any function g. Therefore,

$$\mathrm{Cov}[x, g(y)] = \beta\,\mathrm{Cov}[y, g(y)] + \mathrm{Cov}[\varepsilon, g(y)] = \beta\,\mathrm{Cov}[y, g(y)].$$

Let us write the mean and variance of y as μ_y and σ_y^2, respectively. Then

$$\mathrm{Cov}[y, g(y)] = \mathrm{E}[yg(y)] - \mathrm{E}[y]\,\mathrm{E}[g(y)] = \mathrm{E}[(y - \mu_y)g(y)] = \int_{-\infty}^{\infty} (y - \mu_y)g(y)f(y)\,dy,$$

where

$$f(y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_y^2}(y - \mu_y)^2\right\}$$

is the probability density function of y. Noting that $f'(y) = -f(y)(y - \mu_y)/\sigma_y^2$, integration by parts gives

$$\begin{aligned}\int_{-\infty}^{\infty} (y - \mu_y)g(y)f(y)\,dy &= -\sigma_y^2 \int_{-\infty}^{\infty} g(y)f'(y)\,dy \\ &= \sigma_y^2 \int_{-\infty}^{\infty} g'(y)f(y)\,dy - \sigma_y^2\Big[g(y)f(y)\Big]_{y=-\infty}^{\infty} \\ &= \sigma_y^2\,\mathrm{E}[g'(y)]\end{aligned}$$

provided that $g(y)$ does not approach plus or minus infinity faster than $f(y)$ approaches zero as $y \to \pm\infty$. Hence,

$$\mathrm{Cov}[x, g(y)] = \beta\,\mathrm{Cov}[y, g(y)] = \beta\sigma_y^2\,\mathrm{E}[g'(y)] = \mathrm{Cov}[x, y]\,\mathrm{E}[g'(y)]$$

as claimed. □

For any state-price deflator of the form $\zeta = g(x)$ where x and R_i are jointly normally distributed, we thus have

$$\begin{aligned}
1 &= \mathrm{E}[g(x)R_i] = \mathrm{E}[g(x)]\,\mathrm{E}[R_i] + \mathrm{Cov}[g(x), R_i] \\
&= \mathrm{E}[g(x)]\,\mathrm{E}[R_i] + \mathrm{E}[g'(x)]\,\mathrm{Cov}[x, R_i] \\
&= \mathrm{E}[g(x)]\,\mathrm{E}[R_i] + \mathrm{E}[g'(x)]\,\mathrm{E}[(x - \mathrm{E}[x])R_i] \\
&= \mathrm{E}\left[\{\mathrm{E}[g(x)] - \mathrm{E}[x]E[g'(x)] + \mathrm{E}[g'(x)]x\}\,R_i\right] \\
&= \mathrm{E}[(a + bx)R_i],
\end{aligned}$$

for some constants a and b. Therefore, we can safely assume that $g(x)$ is affine in x.

In (9.2) we have

$$\zeta = e^{-\delta}\frac{u'\left(R^{\boldsymbol{\pi}}(e_0 - c_0)\right)}{u'(c_0)} = g(R^{\boldsymbol{\pi}}),$$

and if the individual asset returns are jointly normally distributed, the return on any portfolio and the return on any individual asset will also be jointly normally distributed. According to Stein's Lemma we can then safely assume that ζ is affine in $R^{\boldsymbol{\pi}}$. Again, this implies that $R^{\boldsymbol{\pi}}$ is a pricing factor. Note, however, that to apply Stein's Lemma, we have to check that $\mathrm{E}[|g'(R^{\boldsymbol{\pi}})|]$ is finite. In our case,

$$g'(R^{\boldsymbol{\pi}}) = e^{-\delta}(e_0 - c_0)\frac{u''\left(R^{\boldsymbol{\pi}}(e_0 - c_0)\right)}{u'(c_0)}.$$

With log-utility, $u''(c) = -1/c^2$, and since $\mathrm{E}[1/(R^{\boldsymbol{\pi}})^2]$ is infinite (or undefined if you like) when $R^{\boldsymbol{\pi}}$ is normally distributed, we cannot apply Stein's Lemma. In fact, when $R^{\boldsymbol{\pi}}$ is normally distributed, we really need the utility function to be defined on the entire real line, which is not the case for the most reasonable utility functions. For negative exponential utility, there is no such problem.

The assumptions leading to the classical CAPM are clearly problematic. Preferences are poorly represented by quadratic utility functions or other mean-variance utility functions. Returns are not normally distributed. A more fundamental problem is the static nature of the one-period CAPM. Later in this chapter we will discuss how the CAPM can be extended to an dynamic setting. It turns out that we can derive an intertemporal CAPM under much more appropriate assumptions about utility functions and return distributions. We will need CRRA utility and lognormally distributed returns. In addition, we will need the return distribution to be stationary, i.e. the same for all future periods of the same length.

9.3 Pricing factors in a one-period framework

In Section 4.5.2 we defined a one-dimensional pricing factor in a one-period framework and discussed the relation between pricing factors and state-price deflators. Below we generalize this to the case of multi-dimensional factors and give a more rigorous treatment.

9.3.1 Definition and basic properties

We will say that a K-dimensional random variable $\boldsymbol{x} = (x_1, \ldots, x_K)^\top$ is a pricing factor for the market if there exists some $\alpha \in \mathbb{R}$ and some $\boldsymbol{\eta} \in \mathbb{R}^K$ so that

$$\mathrm{E}[R_i] = \alpha + \boldsymbol{\beta}[R_i, \boldsymbol{x}]^\top \boldsymbol{\eta}, \quad i = 1, \ldots, I, \tag{9.4}$$

where the factor-beta of asset i is the K-dimensional vector $\boldsymbol{\beta}[R_i, \boldsymbol{x}]$ given by

$$\boldsymbol{\beta}[R_i, \boldsymbol{x}] = (\mathrm{Var}[\boldsymbol{x}])^{-1} \mathrm{Cov}[\boldsymbol{x}, R_i]. \tag{9.5}$$

Here $\mathrm{Var}[\boldsymbol{x}]$ is the $K \times K$ variance-covariance matrix of $\boldsymbol{x}$ and $\mathrm{Cov}[\boldsymbol{x}, R_i]$ is the K-vector with elements $\mathrm{Cov}[x_k, R_i]$. Saying that $\boldsymbol{x}$ is a pricing factor we implicitly require that $\mathrm{Var}[\boldsymbol{x}]$ is non-singular. The vector $\boldsymbol{\eta}$ is called a factor risk premium and α is called the zero-beta return.

We can write (9.4) more compactly as

$$\mathrm{E}[\boldsymbol{R}] = \alpha \mathbf{1} + \underline{\underline{\beta}}[\boldsymbol{R}, \boldsymbol{x}]\boldsymbol{\eta},$$

where $\boldsymbol{R} = (R_1, \ldots, R_I)^\top$ is the return vector, $\mathbf{1} = (1, \ldots, 1)^\top$, and $\underline{\underline{\beta}}[\boldsymbol{R}, \boldsymbol{x}]$ is the $I \times K$ matrix with $\boldsymbol{\beta}[R_i, \boldsymbol{x}]$ as the i'th row. Due to the linearity of expectations and covariance, (9.4) will also hold for all portfolios of the I assets. Note that if a risk-free asset is traded in the market, it will have a zero factor-beta and, consequently, $\alpha = R^f$.

The equation (9.4) involves the gross return R_i on asset i. What about the rate of return $r_i = R_i - 1$? Clearly, $\mathrm{E}[R_i] = 1 + \mathrm{E}[r_i]$, and the properties of covariance give

$$\mathrm{Cov}[\boldsymbol{x}, R_i] = \mathrm{Cov}[\boldsymbol{x}, 1 + r_i] = \mathrm{Cov}[\boldsymbol{x}, r_i] \quad \Rightarrow \quad \boldsymbol{\beta}[R_i, \boldsymbol{x}] = \boldsymbol{\beta}[r_i, \boldsymbol{x}].$$

Consequently, (9.4) implies that

$$\mathrm{E}[r_i] = (\alpha - 1) + \boldsymbol{\beta}[r_i, \boldsymbol{x}]^\top \eta. \tag{9.6}$$

If a risk-free asset exists, $\alpha - 1 = R^f - 1 = r^f$, the risk-free net rate of return.

The relation (9.4) does not directly involve prices. But since the expected gross return is $\mathrm{E}[R_i] = \mathrm{E}[D_i]/P_i$, we have $P_i = \mathrm{E}[D_i]/\mathrm{E}[R_i]$ and hence the equivalent relation

$$P_i = \frac{\mathrm{E}[D_i]}{\alpha + \boldsymbol{\beta}[R_i, \boldsymbol{x}]^\top \boldsymbol{\eta}}. \tag{9.7}$$

The price is equal to the expected dividend discounted by a risk-adjusted rate. You may find this relation unsatisfactory since the price implicitly enters the right-hand side through the "return-beta" $\boldsymbol{\beta}[R_i, \boldsymbol{x}]$. However, we can define a "dividend-beta" by

$$\boldsymbol{\beta}[D_i, \boldsymbol{x}] = (\mathrm{Var}[\boldsymbol{x}])^{-1} \mathrm{Cov}[\boldsymbol{x}, D_i]$$

and inserting $D_i = R_i P_i$ we see that $\boldsymbol{\beta}[D_i, \boldsymbol{x}] = P_i \boldsymbol{\beta}[R_i, \boldsymbol{x}]$. Equation (9.4) now implies that

$$\frac{\mathrm{E}[D_i]}{P_i} = \alpha + \frac{1}{P_i} \boldsymbol{\beta}[D_i, \boldsymbol{x}]^\top \boldsymbol{\eta}$$

so that

$$P_i = \frac{\mathrm{E}[D_i] - \boldsymbol{\beta}[D_i, \boldsymbol{x}]^\top \boldsymbol{\eta}}{\alpha}. \tag{9.8}$$

Think of the numerator as a certainty equivalent of the risky dividend. The current price is the certainty equivalent discounted by the zero-beta return, which is the risk-free return if this exists.

The following result shows that pricing factors are not unique.

Theorem 9.1 *If the K-dimensional random variable $\boldsymbol{x}$ is a pricing factor, then any $\hat{\boldsymbol{x}}$ of the form $\hat{\boldsymbol{x}} = \boldsymbol{a} + \underline{\underline{A}}\boldsymbol{x}$ where $\boldsymbol{a} \in \mathbb{R}^K$ and $\underline{\underline{A}}$ is a non-singular $K \times K$ matrix is also a pricing factor.*

Proof: According to (A.1), (A.2), and (1.1), we have

$$\begin{gathered}\operatorname{Cov}[\hat{\boldsymbol{x}}, R_i] = \operatorname{Cov}[\boldsymbol{a} + \underline{\underline{A}}\boldsymbol{x}, R_i] = \underline{\underline{A}} \operatorname{Cov}[\boldsymbol{x}, R_i], \\ (\operatorname{Var}[\hat{\boldsymbol{x}}])^{-1} = \big(\operatorname{Var}[\boldsymbol{a} + \underline{\underline{A}}\boldsymbol{x}]\big)^{-1} = \big(\operatorname{Var}[\underline{\underline{A}}\boldsymbol{x}]\big)^{-1} = \big(\underline{\underline{A}} \operatorname{Var}[\boldsymbol{x}] \underline{\underline{A}}^\top\big)^{-1} = \big(\underline{\underline{A}}^\top\big)^{-1} (\operatorname{Var}[\boldsymbol{x}])^{-1} \underline{\underline{A}}^{-1},\end{gathered}$$

and thus

$$\boldsymbol{\beta}[R_i, \hat{\boldsymbol{x}}] = (\operatorname{Var}[\hat{\boldsymbol{x}}])^{-1} \operatorname{Cov}[\hat{\boldsymbol{x}}, R_i] = \big(\underline{\underline{A}}^\top\big)^{-1} (\operatorname{Var}[\boldsymbol{x}])^{-1} \underline{\underline{A}}^{-1} \underline{\underline{A}} \operatorname{Cov}[\boldsymbol{x}, R_i] = \big(\underline{\underline{A}}^\top\big)^{-1} \boldsymbol{\beta}[R_i, \boldsymbol{x}].$$

If we define $\hat{\boldsymbol{\eta}} = \underline{\underline{A}}\boldsymbol{\eta}$, we obtain

$$\boldsymbol{\beta}[R_i, \hat{\boldsymbol{x}}]^\top \hat{\boldsymbol{\eta}} = \Big(\big(\underline{\underline{A}}^\top\big)^{-1} \boldsymbol{\beta}[R_i, \boldsymbol{x}]\Big)^\top \underline{\underline{A}}\boldsymbol{\eta} = \boldsymbol{\beta}[R_i, \boldsymbol{x}]^\top \boldsymbol{\eta}$$

and, hence,

$$\operatorname{E}[R_i] = \alpha + \boldsymbol{\beta}[R_i, \hat{\boldsymbol{x}}]^\top \hat{\boldsymbol{\eta}}, \quad i = 1, \ldots, I, \tag{9.9}$$

which confirms that $\hat{\boldsymbol{x}}$ is a pricing factor. □

Let us look at some important consequences of this theorem.

In general the k'th element of the factor beta $\boldsymbol{\beta}[R_i, \boldsymbol{x}]$ is not equal to $\operatorname{Cov}[x_k, R_i]/\operatorname{Var}[x_k]$. This will be the case, however, if the elements in the pricing factor are mutually uncorrelated, i.e. $\operatorname{Cov}[x_j, x_k] = 0$ for $j \neq k$. In fact, we can orthogonalize the pricing factor so that this will be satisfied. Given any pricing factor $\boldsymbol{x}$, we can find a non-singular $K \times K$ matrix $\underline{\underline{V}}$ so that $\underline{\underline{V}}\,\underline{\underline{V}}^\top = \operatorname{Var}[\boldsymbol{x}]$. Defining $\hat{\boldsymbol{x}} = \underline{\underline{V}}^{-1}\boldsymbol{x}$, we know from above that $\hat{\boldsymbol{x}}$ is also a pricing factor and the variance-covariance matrix is

$$\operatorname{Var}[\hat{\boldsymbol{x}}] = \underline{\underline{V}}^{-1} \operatorname{Var}[\boldsymbol{x}] \big(\underline{\underline{V}}^{-1}\big)^\top = \underline{\underline{V}}^{-1} \underline{\underline{V}}\,\underline{\underline{V}}^\top \big(\underline{\underline{V}}^\top\big)^{-1} = \underline{\underline{I}},$$

i.e. the $K \times K$ identity matrix. It is therefore no restriction to look only for uncorrelated pricing factors.

We can also obtain a pricing factor with mean zero. If $\boldsymbol{x}$ is any pricing factor, just define $\hat{\boldsymbol{x}} = \boldsymbol{x} - \operatorname{E}[\boldsymbol{x}]$. Clearly, $\hat{\boldsymbol{x}}$ has mean zero and, due to the previous theorem, it is also a pricing factor. It is therefore no restriction to look only for zero-mean pricing factors.

Finally, note that we can replace the constant vector $\boldsymbol{a}$ in the above theorem with a K-dimensional random variable $\boldsymbol{\varepsilon}$ with the property that $\operatorname{Cov}[R_i, \boldsymbol{\varepsilon}] = \boldsymbol{0}$ for all i. In particular we have that if $\boldsymbol{x}$ is a pricing factor and $\boldsymbol{\varepsilon}$ is such a random variable, then $\boldsymbol{x} + \boldsymbol{\varepsilon}$ is also a pricing factor.

9.3.2 Returns as pricing factors

Suppose now that the pricing factor is a vector of returns on portfolios of the I assets. Then (9.4) holds with each f_k replacing R_i. We have $\operatorname{Cov}[\boldsymbol{x}, \boldsymbol{x}] = \operatorname{Var}[\boldsymbol{x}]$ and hence $\underline{\underline{\beta}}[\boldsymbol{x}, \boldsymbol{x}] = \underline{\underline{I}}$, the $K \times K$ identity matrix. Consequently,

$$\operatorname{E}[\boldsymbol{x}] = \alpha\boldsymbol{1} + \boldsymbol{\eta} \quad \Rightarrow \quad \boldsymbol{\eta} = \operatorname{E}[\boldsymbol{x}] - \alpha\boldsymbol{1},$$

where $\mathbf{1}$ is a K-dimensional vector of ones. In this case we can therefore rewrite (9.4) as

$$\mathrm{E}[R_i] = \alpha + \boldsymbol{\beta}[R_i, \boldsymbol{x}]^\top \left(\mathrm{E}[\boldsymbol{x}] - \alpha\mathbf{1}\right), \quad i = 1, \ldots, I. \tag{9.10}$$

It is now clear that the classical CAPM has the return on the market portfolio as the single pricing factor. More generally, we will demonstrate in Section 9.4.3 that a return works as a single pricing factor if and only if it is the return on a mean-variance efficient portfolio (different from the minimum-variance portfolio).

For the case of a one-dimensional pricing factor x, Section 4.5.2 explained that the return R^x on the factor-mimicking portfolio also works as a pricing factor. We can generalize this to a multi-dimensional pricing factor in the following way. Given a pricing factor $\boldsymbol{x} = (x_1, \ldots, x_K)^\top$, orthogonalize to obtain $\hat{\boldsymbol{x}} = (\hat{x}_1, \ldots, \hat{x}_K)^\top$. For each $\hat{x}_k$ construct a factor-mimicking portfolio with corresponding return $R^{\hat{x}_k}$. Then the return vector $\boldsymbol{R}^{\hat{\boldsymbol{x}}} = \left(R^{\hat{x}_1}, \ldots, R^{\hat{x}_K}\right)^\top$ will work as a pricing factor and we have an equation like

$$\mathrm{E}[R_i] = \alpha + \boldsymbol{\beta}[R_i, \boldsymbol{R}^{\hat{\boldsymbol{x}}}] \cdot \left(\mathrm{E}[\boldsymbol{R}^{\hat{\boldsymbol{x}}}] - \alpha\mathbf{1}\right), \quad i = 1, \ldots, I. \tag{9.11}$$

It is therefore no restriction to assume that pricing factors are returns.

9.3.3 From a state-price deflator to a pricing factor

From the definition of a covariance we have that $\mathrm{Cov}[R_i, \zeta] = \mathrm{E}[R_i\zeta] - \mathrm{E}[\zeta]\,\mathrm{E}[R_i]$. From (4.3), we now get that

$$\mathrm{E}[R_i] = \frac{1}{\mathrm{E}[\zeta]} - \frac{\mathrm{Cov}[R_i, \zeta]}{\mathrm{E}[\zeta]}. \tag{9.12}$$

With $\beta[R_i, \zeta] = \mathrm{Cov}[R_i, \zeta]/\,\mathrm{Var}[\zeta]$ and $\eta = -\,\mathrm{Var}[\zeta]/\,\mathrm{E}[\zeta]$, we can rewrite the above equation as

$$\mathrm{E}[R_i] = \frac{1}{\mathrm{E}[\zeta]} + \beta[R_i, \zeta]\eta, \tag{9.13}$$

which shows that the state-price deflator is a pricing factor. Although the proof is simple, the results is important enough to deserve its own theorem:

Theorem 9.2 *Any state-price deflator ζ is a pricing factor.*

In the above argument we did not use positivity of the state-price deflator, only the pricing equation (4.1) or, rather, the return version (4.3). Any random variable x that satisfies $P_i = \mathrm{E}[xD_i]$ for all assets works as a pricing factor. In particular, this is true for the random variable ζ^* defined in (4.41) whether it is positive or not. We therefore have that

$$\mathrm{E}[R_i] = \alpha^* + \beta[R_i, \zeta^*]\eta^*, \quad i = 1, \ldots, I, \tag{9.14}$$

where $\alpha^* = 1/\,\mathrm{E}[\zeta^*]$ and $\eta^* = -\,\mathrm{Var}[\zeta^*]/\,\mathrm{E}[\zeta^*]$. Alternatively, we can scale by the price of ζ^* and use the return R^* defined in (4.49) as the factor so that

$$\mathrm{E}[R_i] = \alpha^* + \beta[R_i, R^*]\eta^*, \quad i = 1, \ldots, I, \tag{9.15}$$

where $\alpha^* = 1/\,\mathrm{E}[\zeta^*]$ as before but now $\eta^* = -\,\mathrm{Var}[\zeta^*]/\left(\mathrm{E}[\zeta^*]\,\mathrm{E}[(\zeta^*)^2]\right)$.

More generally, we have the following result:

Theorem 9.3 *If the K-dimensional random variable $\boldsymbol{x}$ satisfies*

(i) $\mathrm{Var}[\boldsymbol{x}]$ *is non-singular;*

(ii) $a \in \mathbb{R}$ *and* $\boldsymbol{b} \in \mathbb{R}^K$ *exist so that* $\zeta = a + \boldsymbol{b}^\top \boldsymbol{x}$ *has the properties* $\mathrm{E}[\zeta] \neq 0$ *and* $P_i = \mathrm{E}[\zeta D_i]$ *for* $i = 1, \ldots, I$,

then $\boldsymbol{x}$ is a pricing factor.

Proof: Substituting $\zeta = a + \boldsymbol{b}^\top \boldsymbol{x}$ into (9.12), we get

$$\begin{aligned}
\mathrm{E}[R_i] &= \frac{1}{a + \boldsymbol{b}^\top \mathrm{E}[\boldsymbol{x}]} - \frac{\boldsymbol{b}^\top \mathrm{Cov}[R_i, \boldsymbol{x}]}{a + \boldsymbol{b}^\top \mathrm{E}[\boldsymbol{x}]} \\
&= \frac{1}{a + \boldsymbol{b}^\top \mathrm{E}[\boldsymbol{x}]} - \frac{(\mathrm{Var}[\boldsymbol{x}]\boldsymbol{b})^\top (\mathrm{Var}[\boldsymbol{x}])^{-1} \mathrm{Cov}[R_i, \boldsymbol{x}]}{a + \boldsymbol{b}^\top \mathrm{E}[\boldsymbol{x}]} \\
&= \alpha + \beta[R_i, \boldsymbol{x}]^\top \boldsymbol{\eta},
\end{aligned}$$

where $\alpha = 1/\left(a + \boldsymbol{b}^\top \mathrm{E}[\boldsymbol{x}]\right)$ and $\boldsymbol{\eta} = -\alpha \mathrm{Var}[\boldsymbol{x}]\boldsymbol{b}$. □

Whenever we have a state-price deflator of the form $\zeta = a + \boldsymbol{b}^\top \boldsymbol{x}$, we can use $\boldsymbol{x}$ as a pricing factor.

9.3.4 From a pricing factor to a (candidate) state-price deflator

Conversely:

Theorem 9.4 *Assume the K-dimensional random variable $\boldsymbol{x}$ is a pricing factor with an associated zero-beta return α different from zero. Then we can find $a \in \mathbb{R}$ and $\boldsymbol{b} \in \mathbb{R}^K$ so that $\zeta = a + \boldsymbol{b}^\top \boldsymbol{x}$ satisfies $P_i = \mathrm{E}[\zeta D_i]$ for $i = 1, \ldots, I$.*

Proof: Let $\boldsymbol{\eta}$ denote the factor risk premium associated with the pricing factor $\boldsymbol{x}$. Define

$$\boldsymbol{b} = -\frac{1}{\alpha} (\mathrm{Var}[\boldsymbol{x}])^{-1} \boldsymbol{\eta}, \quad a = \frac{1}{\alpha} - \boldsymbol{b}^\top \mathrm{E}[\boldsymbol{x}].$$

Then $\zeta = a + \boldsymbol{b}^\top \boldsymbol{x}$ works since

$$\begin{aligned}
\mathrm{E}[\zeta R_i] &= a \,\mathrm{E}[R_i] + \boldsymbol{b}^\top \mathrm{E}[R_i \boldsymbol{x}] \\
&= a \,\mathrm{E}[R_i] + \boldsymbol{b}^\top \left(\mathrm{Cov}[R_i, \boldsymbol{x}] + \mathrm{E}[R_i]\,\mathrm{E}[\boldsymbol{x}]\right) \\
&= \left(a + \boldsymbol{b}^\top \mathrm{E}[\boldsymbol{x}]\right) \mathrm{E}[R_i] + \mathrm{Cov}[R_i, \boldsymbol{x}]^\top \boldsymbol{b} \\
&= \frac{1}{\alpha} \left(\mathrm{E}[R_i] - \mathrm{Cov}[R_i, \boldsymbol{x}]^\top (\mathrm{Var}[\boldsymbol{x}])^{-1} \boldsymbol{\eta}\right) \\
&= \frac{1}{\alpha} \left(\mathrm{E}[R_i] - \boldsymbol{\beta}[R_i, \boldsymbol{x}]^\top \boldsymbol{\eta}\right) \\
&= 1
\end{aligned}$$

for any $i = 1, \ldots, I$. □

Inserting a and $\boldsymbol{b}$ from the proof, we get

$$\zeta = a + \boldsymbol{b}^\top \boldsymbol{x} = \frac{1}{\alpha} \left(1 - \boldsymbol{\eta}^\top (\mathrm{Var}[\boldsymbol{x}])^{-1} (\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}])\right).$$

Any pricing factor $\boldsymbol{x}$ gives us a candidate $a + \boldsymbol{b}^\top \boldsymbol{x}$ for a state-price deflator but it will only be a true state-price deflator if it is strictly positive. The fact that we can find a pricing factor for a given market does not imply that the market is arbitrage-free.

9.3.5 The Arbitrage Pricing Theory

Ross (1976) introduced the Arbitrage Pricing Theory as an alternative to the classical CAPM. The basic assumption is that a K-dimensional random variable $\boldsymbol{x} = (x_1, \ldots, x_K)^\top$ exists so that the return on any asset $i = 1, \ldots, I$ can be decomposed as

$$R_i = \mathrm{E}[R_i] + \boldsymbol{\beta}[R_i, \boldsymbol{x}]\boldsymbol{x} + \varepsilon_i = \mathrm{E}[R_i] + \sum_{k=1}^{K} \beta_{ik} x_k + \varepsilon_i,$$

where $\mathrm{E}[x_k] = 0$, $\mathrm{E}[\varepsilon_i] = 0$, $\mathrm{Cov}[\varepsilon_i, x_k] = 0$, and $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0$ for all i, $j \neq i$, and k. Due to the constraints on means and covariances, we have $\boldsymbol{\beta}[R_i, \boldsymbol{x}] = (\mathrm{Var}[\boldsymbol{x}])^{-1} \mathrm{Cov}[R_i, \boldsymbol{x}]$ as before. Note that one can always make a decomposition as in the equation above. Just think of regressing returns on the vector $\boldsymbol{x}$. The real content of the assumption lies in the restriction that the residuals are uncorrelated, i.e. $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0$ whenever $i \neq j$. The vector $\boldsymbol{x}$ is the source of all the common variations in returns across assets.

Suppose you have invested a given wealth in a portfolio. We can represent a zero net investment deviation from this portfolio by a vector $\boldsymbol{w} = (w_1, \ldots, w_I)^\top$ satisfying $\boldsymbol{w} \cdot \mathbf{1} = 0$, where w_i is the fraction of wealth additionally invested in asset i. In other words, we increase the investment in some assets and decrease the investment in other assets. The additional portfolio return is

$$R^{\boldsymbol{w}} = \boldsymbol{w}^\top \boldsymbol{R} = \sum_{i=1}^{I} w_i R_i = \sum_{i=1}^{I} w_i \,\mathrm{E}[R_i] + \sum_{i=1}^{I} w_i \beta_{i1} x_1 + \cdots + \sum_{i=1}^{I} w_i \beta_{iK} x_K + \sum_{i=1}^{I} w_i \varepsilon_i.$$

Suppose we can find $w_1, \ldots, w_I$ so that

(i) $\sum_{i=1}^{I} w_i \beta_{ik} = 0$ for $k = 1, \ldots, K$,

(ii) $\sum_{i=1}^{I} w_i \varepsilon_i = 0$,

then

$$R^{\boldsymbol{w}} = \sum_{i=1}^{I} w_i \,\mathrm{E}[R_i],$$

i.e. the added return is risk-free. To rule out arbitrage, a risk-free zero net investment should give a zero return so we can conclude that

$$R^{\boldsymbol{w}} = \sum_{i=1}^{I} w_i \,\mathrm{E}[R_i] = 0.$$

In linear algebra terms, we have thus seen that if a vector $\boldsymbol{w}$ is orthogonal to $\mathbf{1}$ and to each of the vectors $(\beta_{1k}, \ldots, \beta_{Ik})^\top$, $k = 1, \ldots, K$, then it must also be orthogonal to the vector of expected returns $\mathrm{E}[\boldsymbol{R}]$. It follows that $\mathrm{E}[\boldsymbol{R}]$ must be spanned by the vectors $\mathbf{1}$, $(\beta_{1k}, \ldots, \beta_{Ik})^\top$, $k = 1, \ldots, K$, i.e. that constants $\alpha, \eta_1, \ldots, \eta_K$ exist so that

$$\mathrm{E}[R_i] = \alpha + \beta_{i1}\eta_1 + \cdots + \beta_{iK}\eta_K = \alpha + \boldsymbol{\beta}[R_i, \boldsymbol{x}]\boldsymbol{\eta}, \quad i = 1, 2, \ldots, I,$$

i.e. $\boldsymbol{x}$ is a pricing factor.

With at least K sufficiently different assets, we can satisfy condition (i) above. We just need the $I \times K$ matrix $\underline{\underline{\beta}}[\boldsymbol{R}, \boldsymbol{x}]$ to have rank K. What about condition (ii)? The usual argument given is that if we pick w_i to be of the order $1/I$ and I is a very large number, then $\sum_{i=1} w_i \varepsilon_i$ will

be close to zero and we can ignore it. But close to zero does not mean equal to zero and if the residual portfolio return is non-zero, the portfolio is not risk-free and the argument breaks down. Even a very small dividend or return in a particular state can have a large influence on the current price and, hence, the expected return. With finitely many assets, we can only safely ignore the residual portfolio return if the residual returns of all the individual assets are zero, i.e. $\varepsilon_i = 0$ for all $i = 1, \ldots, I$.

Theorem 9.5 *If individual asset returns are of the form*

$$R_i = \mathrm{E}[R_i] + \boldsymbol{\beta}[R_i, \boldsymbol{x}]\boldsymbol{x}, \quad i = 1, 2, \ldots, I,$$

and the $I \times K$ matrix $\underline{\underline{\beta}}[\boldsymbol{R}, \boldsymbol{x}]$ has rank K, then $\boldsymbol{x}$ is a pricing factor, i.e. $\alpha \in \mathbb{R}$ and $\boldsymbol{\eta} \in \mathbb{R}^K$ exist so that

$$\mathrm{E}[R_i] = \alpha + \boldsymbol{\beta}[R_i, \boldsymbol{x}]^\top \boldsymbol{\eta}, \quad i = 1, 2, \ldots, I.$$

It is, however, fairly restrictive to assume that *all* the variation in the returns on a large number of assets can be captured by a low number of factors.

9.4 Mean-variance efficient returns and pricing factors

We have introduced the mean-variance frontier earlier in Sections 4.5.3 and 6.2.5. Here we provide an alternative characterization of the mean-variance efficient returns and study the link between mean-variance efficiency and asset pricing theory.

As before let $\boldsymbol{R} = (R_1, \ldots, R_I)^\top$ denote the vector of gross returns on the I traded assets and define $\boldsymbol{\mu} = \mathrm{E}[\boldsymbol{R}]$ and $\underline{\underline{\Sigma}} = \mathrm{Var}[\boldsymbol{R}]$. A portfolio $\boldsymbol{\pi}$ is mean-variance efficient if there is an $m \in \mathbb{R}$ so that $\boldsymbol{\pi}$ solves

$$\min_{\boldsymbol{\pi}} \boldsymbol{\pi}^\top \underline{\underline{\Sigma}} \boldsymbol{\pi} \quad \text{s.t.} \quad \boldsymbol{\pi}^\top \boldsymbol{\mu} = m, \quad \boldsymbol{\pi}^\top \mathbf{1} = 1, \tag{6.21}$$

i.e. $\boldsymbol{\pi}$ has the lowest return variance among all portfolios with expected return m.

9.4.1 Orthogonal characterization

Following Hansen and Richard (1987) we will show that a return R is mean-variance efficient if and only if it can be written on the form $R = R^* + wR^{e*}$ for some number w. Here R^* is the return on the portfolio corresponding to the dividend ζ^* defined in (4.41) and R^{e*} is a particular excess return to be defined shortly. This characterization of the mean-variance portfolios turn out to be preferable for discussing the link between mean-variance analysis and asset pricing models.

R^* and mean-variance analysis

How does R^* fit into the mean-variance framework? The following lemma shows that it is a mean-variance efficient return on the downward-sloping part of the mean-variance frontier. Furthermore, it is the return with minimum second moment. (Recall that the second moment of a random variable x is $\mathrm{E}[x^2]$.)

Lemma 9.2 *R^* has the following properties:*

(a) R^* is the return that has minimum second moment,

(b) R^* is a mean-variance efficient return located on the downward-sloping part of the efficient frontier.

Proof: The return on a portfolio $\boldsymbol{\pi}$ is the random variable $R^{\boldsymbol{\pi}} = \boldsymbol{\pi}^\top \boldsymbol{R}$. The second moment of this return is $\mathrm{E}[(R^{\boldsymbol{\pi}})^2] = \boldsymbol{\pi}^\top \mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\boldsymbol{\pi}$. Consider the minimization problem

$$\min_{\boldsymbol{\pi}} \boldsymbol{\pi}^\top \mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\boldsymbol{\pi} \qquad \text{s.t.} \qquad \boldsymbol{\pi}^\top \mathbf{1} = 1.$$

The Lagrangian is $\mathcal{L} = \boldsymbol{\pi}^\top \mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\boldsymbol{\pi} + \lambda\,(1 - \boldsymbol{\pi}^\top \mathbf{1})$, where λ is the Lagrange multiplier. The first-order condition for $\boldsymbol{\pi}$ is

$$2\,\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\boldsymbol{\pi} - \lambda \mathbf{1} = 0 \quad \Rightarrow \quad \boldsymbol{\pi} = \frac{\lambda}{2}\left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1} \mathbf{1}.$$

Imposing the constraint $\boldsymbol{\pi}^\top \mathbf{1} = 1$, we get $\lambda/2 = 1/(\mathbf{1}^\top \mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\mathbf{1})$, and substituting this into the above expression for $\boldsymbol{\pi}$, we see that $\boldsymbol{\pi}$ is indeed identical to $\boldsymbol{\pi}^*$ in (4.48).

Since $\boldsymbol{\pi}^*$ is the portfolio minimizing the second moment of gross returns among all portfolios, it is also the portfolio minimizing the second moment of gross returns among the portfolios with the same mean return as $\boldsymbol{\pi}^*$, i.e. portfolios with $\mathrm{E}[R^{\boldsymbol{\pi}}] = \mathrm{E}[R^*]$. For these portfolios the variance of return is

$$\mathrm{Var}\,[R^{\boldsymbol{\pi}}] = \mathrm{E}[(R^{\boldsymbol{\pi}})^2] - (\mathrm{E}[R^{\boldsymbol{\pi}}])^2 = \mathrm{E}[(R^{\boldsymbol{\pi}})^2] - (\mathrm{E}[R^*])^2\,.$$

It then follows that $\boldsymbol{\pi}^*$ is the portfolio minimizing the variance of return among all the portfolios having the same mean return as $\boldsymbol{\pi}^*$. Hence, $\boldsymbol{\pi}^*$ is indeed a mean-variance efficient portfolio. In a (standard deviation, mean)-diagram returns with same second moment $K = \mathrm{E}[(R^{\boldsymbol{\pi}})^2]$ yield points on a circle with radius $\sqrt{K}$ centered in (0,0), since $(\sigma(R^{\boldsymbol{\pi}}))^2 + (\mathrm{E}[R^{\boldsymbol{\pi}}])^2 = \mathrm{E}[(R^{\boldsymbol{\pi}})^2]$. The return R^* therefore corresponds to a point on the downward-sloping part of the efficient frontier. □

The constant-mimicking return

In a market without a risk-free asset you may wonder how close you can get to a risk-free dividend. Of course this will depend on what you mean by "close." If you apply a mean-square measure, the distance between the dividend $D^{\boldsymbol{\theta}} = \boldsymbol{\theta}^\top \boldsymbol{D}$ of a portfolio $\boldsymbol{\theta}$ and a risk-free dividend of 1 is

$$\begin{aligned}
\mathrm{E}\left[(D^{\boldsymbol{\theta}} - 1)^2\right] &= \mathrm{E}\left[(\boldsymbol{\theta}^\top \boldsymbol{D} - 1)^2\right] \\
&= \mathrm{E}\left[\boldsymbol{\theta}^\top \boldsymbol{D}\boldsymbol{D}^\top \boldsymbol{\theta} + 1 - 2\boldsymbol{\theta}^\top \boldsymbol{D}\right] \\
&= \boldsymbol{\theta}^\top \mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right] \boldsymbol{\theta} + 1 - 2\boldsymbol{\theta}^\top \mathrm{E}\left[\boldsymbol{D}\right].
\end{aligned}$$

Minimizing with respect to $\boldsymbol{\theta}$, we get $\boldsymbol{\theta}_{\mathrm{cm}} = \left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} \mathrm{E}\left[\boldsymbol{D}\right]$ where the subscript "cm" is short for "constant-mimicking." Let us transform this to a vector $\boldsymbol{\pi}_{\mathrm{cm}}$ of portfolio weights by using (3.6). Applying (3.2) and (4.47), we get

$$\begin{aligned}
\mathrm{diag}(\boldsymbol{P})\boldsymbol{\theta}_{\mathrm{cm}} &= \mathrm{diag}(\boldsymbol{P})\left(\mathrm{E}\left[\boldsymbol{D}\boldsymbol{D}^\top\right]\right)^{-1} \mathrm{E}\left[\boldsymbol{D}\right] \\
&= \mathrm{diag}(\boldsymbol{P})[\mathrm{diag}(\boldsymbol{P})]^{-1}\left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1}[\mathrm{diag}(\boldsymbol{P})]^{-1} \mathrm{E}\left[\boldsymbol{D}\right] \\
&= \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathrm{E}\left[\boldsymbol{R}\right]
\end{aligned}$$

so that

$$\boldsymbol{\pi}_{\text{cm}} = \frac{1}{\mathbf{1}^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathrm{E}\left[\boldsymbol{R}\right]} \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathrm{E}\left[\boldsymbol{R}\right] = \frac{\mathrm{E}[(R^*)^2]}{\mathrm{E}[R^*]} \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \mathrm{E}\left[\boldsymbol{R}\right], \tag{9.16}$$

where the last equality comes from (4.53). The constant-mimicking return is thus

$$R_{\text{cm}} = \left(\boldsymbol{\pi}_{\text{cm}}\right)^\top \boldsymbol{R} = \frac{\mathrm{E}[(R^*)^2]}{\mathrm{E}[R^*]} \mathrm{E}\left[\boldsymbol{R}\right]^\top \left(\mathrm{E}\left[\boldsymbol{R}\boldsymbol{R}^\top\right]\right)^{-1} \boldsymbol{R}. \tag{9.17}$$

Excess returns and R^{e*}

An excess return is simply the difference between two returns. Since any return corresponds to a dividend for a unit initial payment, an excess return can be seen as a dividend for a zero initial payment. Of course, in absence of arbitrage, a non-zero excess return will turn out positive in some states and negative in other states.

Typically, excess returns on different portfolios relative to the same "reference" return are considered. For a reference return $\check{R} = \check{\boldsymbol{\pi}}^\top \boldsymbol{R}$, the set of all possible excess returns are given by

$$\underline{R}^e[\check{R}] = \left\{\boldsymbol{\pi}^\top \boldsymbol{R} - \check{R} \mid \boldsymbol{\pi}^\top \mathbf{1} = 1\right\},$$

where as before $\boldsymbol{R}$ is the I-dimensional vector of returns on the basis assets, and $\boldsymbol{\pi}$ denotes a portfolio weight vector of these I assets.

It is useful to observe that the set of excess returns is the same for all reference returns. An excess return relative to a reference return $\check{R}$ is given by $\boldsymbol{\pi}^\top \boldsymbol{R} - \check{R} = (\boldsymbol{\pi} - \check{\boldsymbol{\pi}})^\top \boldsymbol{R}$ for some portfolio weight vector $\boldsymbol{\pi}$. We can obtain the same excess return relative to another reference return $\dot{R} = \dot{\boldsymbol{\pi}}^\top \boldsymbol{R}$ using the portfolio $\boldsymbol{\pi} + \dot{\boldsymbol{\pi}} - \check{\boldsymbol{\pi}}$. (Check for yourself!) Hence, we can simply talk about *the* set of excess returns, $\underline{R}^e$, without specifying any reference return, and we can write it as

$$\underline{R}^e = \{(\boldsymbol{\pi}^e)^\top \boldsymbol{R} \mid (\boldsymbol{\pi}^e)^\top \mathbf{1} = 0\}.$$

The set of excess returns is a linear subspace of the set of all random variables (in our case with S possible outcomes this is equivalent to $\mathbb{R}^S$) in the sense that

1. if w is a number and R^e is an excess return, then wR^e is an excess return;
   Check: $R^e = (\boldsymbol{\pi}^e)^\top \boldsymbol{R}$ with $(\boldsymbol{\pi}^e)^\top \mathbf{1} = 0$ implies that $wR^e = (w\boldsymbol{\pi}^e)^\top \boldsymbol{R}$ with $(w\boldsymbol{\pi}^e)^\top \mathbf{1} = w(\boldsymbol{\pi}^e)^\top \mathbf{1} = 0$

2. if R^{e1} and R^{e2} are two excess returns, then $R^{e1} + R^{e2}$ is also an excess return.
   Check: $R^{ei} = (\boldsymbol{\pi}^{ei})^\top \boldsymbol{R}$ with $(\boldsymbol{\pi}^{ei})^\top \mathbf{1} = 0$ implies that $R^{e1} + R^{e2} = (\boldsymbol{\pi}^{e1} + \boldsymbol{\pi}^{e2})^\top \boldsymbol{R}$, where $(\boldsymbol{\pi}^{e1} + \boldsymbol{\pi}^{e2})^\top \mathbf{1} = 0$

Define

$$R^{e*} = \frac{\mathrm{E}[R^*]}{\mathrm{E}[(R^*)^2]} \left(R_{\text{cm}} - R^*\right) = \mathrm{E}[\boldsymbol{R}]^\top \left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1} \boldsymbol{R} - \frac{\mathrm{E}[R^*]}{\mathrm{E}[(R^*)^2]} R^*. \tag{9.18}$$

Of course, $R_{\text{cm}} - R^*$ is an excess return, and since R^{e*} is a multiplum of that, we can conclude that R^{e*} is an excess return. Here are some important properties of R^{e*}:

Lemma 9.3 *(a) For any excess return R^e, we have*

$$\mathrm{E}[R^e R^*] = 0. \tag{9.19}$$

In particular,

$$\mathrm{E}[R^{e*}R^*] = 0. \tag{9.20}$$

(b) For any excess return R^e we have

$$\mathrm{E}[R^e] = \mathrm{E}[R^{e*}R^e]. \tag{9.21}$$

(c) $\mathrm{E}[R^{e*}] = \mathrm{E}[(R^{e*})^2]$ *and* $\mathrm{Var}[R^{e*}] = \mathrm{E}[R^{e*}](1 - \mathrm{E}[R^{e*}])$.

Proof: (a) For any return R^i, we have $\mathrm{E}[R^*R^i] = \mathrm{E}[(R^*)^2]$, and hence for any excess return $R^e = R^i - R^j$, we have

$$\mathrm{E}[R^*R^e] = \mathrm{E}[R^*(R^i - R^j)] = \mathrm{E}[R^*R^i] - \mathrm{E}[R^*R^j] = \mathrm{E}[(R^*)^2] - \mathrm{E}[(R^*)^2] = 0.$$

In particular, this is true for the excess return R^{e*}.

(b) Write the excess return R^e as the difference between the return on some portfolio $\boldsymbol{\pi}$ and R^*, i.e. $R^e = \boldsymbol{\pi}^\top \boldsymbol{R} - R^*$. Then

$$\mathrm{E}[R^{e*}R^e] = \mathrm{E}[R^{e*}\,(\boldsymbol{\pi}^\top \boldsymbol{R} - R^*)] = \boldsymbol{\pi}^\top \mathrm{E}[R^{e*}\boldsymbol{R}] - \mathrm{E}[R^{e*}R^*] = \boldsymbol{\pi}^\top \mathrm{E}[R^{e*}\boldsymbol{R}]. \tag{9.22}$$

Using (9.18), we get

$$\mathrm{E}[R^{e*}\boldsymbol{R}] = \mathrm{E}\left[\left(\mathrm{E}[\boldsymbol{R}]^\top \left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1} \boldsymbol{R}\right)\boldsymbol{R}\right] - \frac{\mathrm{E}[R^*]}{\mathrm{E}[(R^*)^2]}\,\mathrm{E}[R^*\boldsymbol{R}]. \tag{9.23}$$

Let us first consider the last part. From Lemma 4.1, we have $\mathrm{E}[R^*R_i] = \mathrm{E}[(R^*)^2]$ for each i so that $\mathrm{E}[R^*\boldsymbol{R}] = \mathrm{E}[(R^*)^2]\mathbf{1}$. Now look at the first part of (9.23). It can be shown in general that, for any vector $\boldsymbol{x}$, $\mathrm{E}[\boldsymbol{x}^\top \boldsymbol{R}\boldsymbol{R}] = \mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\boldsymbol{x}$. Applying this with $\boldsymbol{x} = \left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1}\mathrm{E}[\boldsymbol{R}]$ we obtain

$$\mathrm{E}\left[\left(\mathrm{E}[\boldsymbol{R}]^\top \left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1} \boldsymbol{R}\right)\boldsymbol{R}\right] = \mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1}\mathrm{E}[\boldsymbol{R}] = \mathrm{E}[\boldsymbol{R}].$$

We can now rewrite (9.23):

$$\mathrm{E}[R^{e*}\boldsymbol{R}] = \mathrm{E}[\boldsymbol{R}] - \mathrm{E}[R^*]\mathbf{1}.$$

Going back to (9.22), we have

$$\mathrm{E}[R^{e*}R^e] = \boldsymbol{\pi}^\top \mathrm{E}[R^{e*}\boldsymbol{R}] = \boldsymbol{\pi}^\top (\mathrm{E}[\boldsymbol{R}] - \mathrm{E}[R^*]\mathbf{1}) = \boldsymbol{\pi}^\top \mathrm{E}[\boldsymbol{R}] - \mathrm{E}[R^*] = \mathrm{E}[R^e],$$

as had to be shown.

(c) The first part follows immediately from (b). The second part comes from

$$\mathrm{Var}[R^{e*}] = \mathrm{E}[(R^{e*})^2] - (\mathrm{E}[R^{e*}])^2 = \mathrm{E}[R^{e*}] - (\mathrm{E}[R^{e*}])^2 = \mathrm{E}[R^{e*}]\,(1 - \mathrm{E}[R^{e*}])$$

where we have used the first part. □

A characterization of the mean-variance frontier in terms of R^* and R^{e*}

First we show that any return can be decomposed using R^*, R^{e*}, and some "residual" excess return.

Theorem 9.6 *For any return R_i, we can find a number w_i and an excess return $\eta_i \in \underline{R}^e$ so that*

$$R_i = R^* + w_i R^{e*} + \eta_i \tag{9.24}$$

and

$$\mathrm{E}[\eta_i] = \mathrm{E}[R^* \eta_i] = \mathrm{E}[R^{e*} \eta_i] = 0. \tag{9.25}$$

Proof: Define $w_i = (\mathrm{E}[R_i] - \mathrm{E}[R^*]) \,/\, \mathrm{E}[R^{e*}]$ and $\eta_i = R_i - R^* - w_i R^{e*}$. Then (9.24) and $\mathrm{E}[\eta_i] = 0$ hold by construction. η_i is the difference between the two excess returns $R_i - R^*$ and $w_i R^{e*}$ and therefore itself an excess return. Now $\mathrm{E}[R^* \eta_i] = 0$ follows from (9.19) and from (9.21) we have $\mathrm{E}[\eta_i R^{e*}] = \mathrm{E}[\eta_i]$, which we know is zero. □

Due to the relations $\mathrm{E}[R^* R^{e*}] = \mathrm{E}[R^* \eta_i] = \mathrm{E}[R^{e*} \eta_i] = 0$, the decomposition is said to be orthogonal.

Note that the same w_i applies for all returns with the same expected value. The return variance is

$$\begin{aligned}
\mathrm{Var}[R_i] &= \mathrm{Var}\left[R^* + w_i R^{e*}\right] + \mathrm{Var}[\eta_i] + 2\,\mathrm{Cov}\left[R^* + w_i R^{e*}, \eta_i\right] \\
&= \mathrm{Var}\left[R^* + w_i R^{e*}\right] + \mathrm{Var}[\eta_i] + 2\left\{\mathrm{E}\left[\left(R^* + w_i R^{e*}\right)\eta_i\right] - \mathrm{E}[R^* + w_i R^{e*}]\,\mathrm{E}[\eta_i]\right\} \\
&= \mathrm{Var}\left[R^* + w_i R^{e*}\right] + \mathrm{Var}[\eta_i].
\end{aligned}$$

Clearly the minimum variance for a given mean, i.e. a given w_i is obtained for $\eta_i = 0$. We therefore have the following result:

Theorem 9.7 *A return R_i is mean-variance efficient if and only if it can be written as*

$$R_i = R^* + w_i R^{e*}$$

for some number w_i.

Varying w_i from $-\infty$ to $+\infty$, $R_i = R^* + w_i R^{e*}$ runs through the entire mean-variance frontier in the direction of higher and higher expected returns (since $\mathrm{E}[R^{e*}] = \mathrm{E}[(R^{e*})^2] > 0$).

Note that from (9.18) that the constant-mimicking return can be written as

$$R_{\mathrm{cm}} = R^* + \frac{\mathrm{E}\left[(R^*)^2\right]}{\mathrm{E}[R^*]} R^{e*} \tag{9.26}$$

so the constant-mimicking return is mean-variance efficient.

Allowing for a risk-free asset

Now let us assume that a risk-free asset with return R^f exists (or can be constructed as a portfolio of the basic assets). Then from Lemma 4.1 we have that $R^f\,\mathrm{E}[R^*] = \mathrm{E}[R^* R^f] = \mathrm{E}[(R^*)^2]$, and hence

$$R^f = \frac{\mathrm{E}[(R^*)^2]}{\mathrm{E}[R^*]} = \frac{1}{\mathbf{1}^\top \left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1} \mathrm{E}[\boldsymbol{R}]}. \tag{9.27}$$

In addition we have

$$\mathrm{E}[\boldsymbol{R}]^\top \left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1} \boldsymbol{R} = 1. \tag{9.28}$$

Let us just show this for the case with two assets, a risk-free and a risky so that $\boldsymbol{R} = (\tilde{R}, R^f)^\top$, where $\tilde{R}$ is the return on the risky asset. Then

$$\begin{aligned}
\mathrm{E}[\boldsymbol{R}]^\top \left(\mathrm{E}[\boldsymbol{R}\boldsymbol{R}^\top]\right)^{-1} \boldsymbol{R} &= \begin{pmatrix} \mathrm{E}[\tilde{R}], R^f \end{pmatrix}^\top \begin{pmatrix} \mathrm{E}[\tilde{R}^2] & R^f \,\mathrm{E}[\tilde{R}] \\ R^f \,\mathrm{E}[\tilde{R}] & (R^f)^2 \end{pmatrix}^{-1} \begin{pmatrix} \tilde{R} \\ R^f \end{pmatrix} \\
&= \frac{1}{(R^f)^2 \left(\mathrm{E}[\tilde{R}^2] - (\mathrm{E}[\tilde{R}])^2\right)} \begin{pmatrix} \mathrm{E}[\tilde{R}], R^f \end{pmatrix}^\top \begin{pmatrix} (R^f)^2 & -R^f \,\mathrm{E}[\tilde{R}] \\ -R^f \,\mathrm{E}[\tilde{R}] & \mathrm{E}[\tilde{R}^2] \end{pmatrix} \begin{pmatrix} \tilde{R} \\ R^f \end{pmatrix} \\
&= \frac{1}{(R^f)^2 \left(\mathrm{E}[\tilde{R}^2] - (\mathrm{E}[\tilde{R}])^2\right)} \begin{pmatrix} \mathrm{E}[\tilde{R}], R^f \end{pmatrix}^\top \begin{pmatrix} (R^f)^2 \left(\tilde{R} - \mathrm{E}[\tilde{R}]\right) \\ R^f \left(\mathrm{E}[\tilde{R}^2] - \mathrm{E}[\tilde{R}]\tilde{R}\right) \end{pmatrix} \\
&= 1.
\end{aligned}$$

Substituting (9.27) and (9.28) into the general definition of R^{e*} in (9.18), we get

$$R^{e*} = 1 - \frac{1}{R^f} R^*. \tag{9.29}$$

Consequently,

$$R^f = R^* + R^f R^{e*} \tag{9.30}$$

so the w_i corresponding to the risk-free return is R^f itself. With $R^f > 1$, we see that $R^* + R^{e*}$ corresponds to a point on the frontier below the risk-free rate. (Again we use the fact that $\mathrm{E}[R^{e*}] > 0$.)

The minimum-variance return

With the decomposition in Theorem 9.7, it is easy to find the minimum-variance return. The variance of any mean-variance efficient return is

$$\begin{aligned}
\mathrm{Var}\left[R^* + wR^{e*}\right] &= \mathrm{E}\left[(R^* + wR^{e*})^2\right] - \left(\mathrm{E}\left[R^* + wR^{e*}\right]\right)^2 \\
&= \mathrm{E}\left[(R^*)^2\right] + w^2 \,\mathrm{E}\left[(R^{e*})^2\right] + 2w \,\mathrm{E}\left[R^* R^{e*}\right] \\
&\quad - \left(\mathrm{E}[R^*]\right)^2 - w^2 \left(\mathrm{E}[R^{e*}]\right)^2 - 2w \,\mathrm{E}[R^*] \,\mathrm{E}[R^{e*}] \\
&= \mathrm{E}\left[(R^*)^2\right] + w^2 \,\mathrm{E}\left[R^{e*}\right] \left(1 - \mathrm{E}\left[R^{e*}\right]\right) - \left(\mathrm{E}[R^*]\right)^2 - 2w \,\mathrm{E}[R^*] \,\mathrm{E}[R^{e*}],
\end{aligned}$$

where the simplifications leading to the last expression are due to Lemma 9.3. The first-order condition with respect to w implies that

$$w = \frac{\mathrm{E}[R^*]}{1 - \mathrm{E}[R^{e*}]}.$$

The minimum-variance return is thus

$$R_{\min} = R^* + \frac{\mathrm{E}[R^*]}{1 - \mathrm{E}[R^{e*}]} R^{e*} = R^* + \frac{\mathrm{E}[R^*] \,\mathrm{E}[R^{e*}]}{\mathrm{Var}[R^{e*}]} R^{e*}. \tag{9.31}$$

When a risk-free asset exists, this simplifies to $R_{\min} = R^f$, because $\mathrm{E}[R^*]/(1 - \mathrm{E}[R^{e*}]) = R^f$ using (9.29).

9.4.2 Link between mean-variance efficient returns and state-price deflators

Theorem 9.8 *Let R denote a gross return. Then there exists $a, b \in \mathbb{R}$ so that $\zeta = a + bR$ satisfies $P_i = \mathrm{E}[\zeta D_i]$ for $i = 1, \ldots, I$ if and only if R is a mean-variance efficient return different from the constant-mimicking return.*

Proof: According to Theorem 9.6 we can decompose the return R as

$$R = R^* + wR^{e*} + \eta$$

for some $w \in \mathbb{R}$ and some excess return η with $\mathrm{E}[\eta_i] = \mathrm{E}[R^*\eta_i] = \mathrm{E}[R^{e*}\eta_i] = 0$. R is mean-variance efficient if and only if $\eta = 0$. We need to show that, for suitable $a, b \in \mathbb{R}$,

$$\zeta = a + bR = a + b\left(R^* + wR^{e*} + \eta\right)$$

will satisfy $P_i = \mathrm{E}[\zeta D_i]$ for all i if and only if $\eta = 0$ and $w \neq \mathrm{E}[(R^*)^2]/\,\mathrm{E}[R^*]$.

Recall that $P_i = \mathrm{E}[\zeta D_i]$ for all i implies that $\mathrm{E}[\zeta R_i] = 1$ for all returns R_i and $\mathrm{E}[\zeta R^e] = 0$ for all excess returns R^e. In particular,

$$\begin{aligned}
1 = \mathrm{E}[\zeta R^*] &= \mathrm{E}\left[\left(a + b\left(R^* + wR^{e*} + \eta\right)\right)R^*\right] = a\,\mathrm{E}[R^*] + b\,\mathrm{E}\left[(R^*)^2\right] \\
0 = \mathrm{E}[\zeta R^{e*}] &= \mathrm{E}\left[\left(a + b\left(R^* + wR^{e*} + \eta\right)\right)R^{e*}\right] = a\,\mathrm{E}[R^{e*}] + bw\,\mathrm{E}\left[(R^{e*})^2\right] = (a + bw)\,\mathrm{E}[R^{e*}].
\end{aligned}$$

Solving for a and b, we get

$$a = \frac{w}{w\,\mathrm{E}[R^*] - \mathrm{E}\left[(R^*)^2\right]}, \quad b = -\frac{1}{w\,\mathrm{E}[R^*] - \mathrm{E}\left[(R^*)^2\right]}$$

so that

$$\zeta = \frac{w - \left(R^* + wR^{e*} + \eta\right)}{w\,\mathrm{E}[R^*] - \mathrm{E}\left[(R^*)^2\right]}.$$

Obviously, we have to assume that $w\,\mathrm{E}[R^*] \neq \mathrm{E}\left[(R^*)^2\right]$, which rules out the constant-mimicking return, cf. (9.26).

Now consider any other return R_i and decompose to $R_i = R^* + w_i R^{e*} + \eta_i$. Then

$$\begin{aligned}
\mathrm{E}[\zeta R_i] &= \frac{1}{w\,\mathrm{E}[R^*] - \mathrm{E}\left[(R^*)^2\right]}\,\mathrm{E}\left[\left(w - \left(R^* + wR^{e*} + \eta\right)\right)\left(R^* + w_i R^{e*} + \eta_i\right)\right] \\
&= \frac{1}{w\,\mathrm{E}[R^*] - \mathrm{E}\left[(R^*)^2\right]}\left(w\,\mathrm{E}[R^*] - \mathrm{E}\left[(R^*)^2\right] - \mathrm{E}[\eta\eta_i]\right),
\end{aligned}$$

where we have applied various results from earlier. We can now see that we will have $\mathrm{E}[\zeta R_i] = 1$ for all returns R_i if and only if $\mathrm{E}[\eta\eta_i] = 0$ for all excess returns η_i. In particular $\mathrm{E}[\eta^2] = 0$, which implies that $\eta = 0$. □

9.4.3 Link between mean-variance efficient returns and pricing factors

Theorem 9.9 *A return R^{mv} is a pricing factor, i.e. an $\alpha \in \mathbb{R}$ exists so that*

$$\mathrm{E}[R_i] = \alpha + \beta[R_i, R^{mv}]\left(\mathrm{E}[R^{mv}] - \alpha\right), \quad i = 1, \ldots, I, \tag{9.32}$$

if and only if R^{mv} is a mean-variance efficient return different from the minimum-variance return.

The constant α must then be equal to the zero-beta return corresponding to R^{mv}, which is equal to the risk-free return if a risk-free asset exists.

Proof: ("If" part.) First, let us show that if R^{mv} is mean-variance efficient and different from the minimum-variance return, it will work as a pricing factor. This result is originally due to Roll (1977). For some w, we have $R^{\text{mv}} = R^* + wR^{e*}$. Consider a general return R_i and decompose as

$$R_i = R^* + w_i R^{e*} + \eta_i$$

as in Theorem 9.6. Then

$$\mathrm{E}[R_i] = \mathrm{E}[R^*] + w_i \,\mathrm{E}[R^{e*}]$$

and

$$\begin{aligned}\mathrm{Cov}[R_i, R^{\text{mv}}] &= \mathrm{Var}[R^*] + ww_i \,\mathrm{Var}[R^{e*}] + (w + w_i)\,\mathrm{Cov}[R^*, R^{e*}] \\ &= \mathrm{Var}[R^*] + ww_i \,\mathrm{Var}[R^{e*}] - (w + w_i)\,\mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]\end{aligned}$$

which implies that

$$\mathrm{Cov}[R_i, R^{\text{mv}}] - \mathrm{Var}[R^*] + w\,\mathrm{E}[R^*]\,\mathrm{E}[R^{e*}] = w_i\left(w\,\mathrm{Var}[R^{e*}] - \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]\right).$$

If $w \neq \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]/\,\mathrm{Var}[R^{e*}]$, which according to (9.31) means that R^{mv} is different from the minimum-variance return, then we can solve the above equation for w_i with the solution

$$w_i = \frac{\mathrm{Cov}[R_i, R^{\text{mv}}] - \mathrm{Var}[R^*] + w\,\mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]}{w\,\mathrm{Var}[R^{e*}] - \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]}.$$

Hence

$$\begin{aligned}\mathrm{E}[R_i] &= \mathrm{E}[R^*] + w_i\,\mathrm{E}[R^{e*}] \\ &= \mathrm{E}[R^*] + \frac{\mathrm{Cov}[R_i, R^{\text{mv}}] - \mathrm{Var}[R^*] + w\,\mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]}{w\,\mathrm{Var}[R^{e*}] - \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]}\,\mathrm{E}[R^{e*}] \\ &= \alpha + \frac{\mathrm{Cov}[R_i, R^{\text{mv}}]}{w\,\mathrm{Var}[R^{e*}] - \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]}\,\mathrm{E}[R^{e*}]\end{aligned}$$

where we have defined the constant α as

$$\alpha = \mathrm{E}[R^*] + \frac{w\,\mathrm{E}[R^*]\,\mathrm{E}[R^{e*}] - \mathrm{Var}[R^*]}{w\,\mathrm{Var}[R^{e*}] - \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]}\,\mathrm{E}[R^{e*}].$$

This applies to any return R_i and in particular to R^{mv} itself, which implies that

$$\mathrm{E}[R^{\text{mv}}] = \alpha + \frac{\mathrm{Var}[R^{\text{mv}}]}{w\,\mathrm{Var}[R^{e*}] - \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]}\,\mathrm{E}[R^{e*}]$$

and hence

$$\frac{\mathrm{E}[R^{e*}]}{w\,\mathrm{Var}[R^{e*}] - \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]} = \frac{\mathrm{E}[R^{\text{mv}}] - \alpha}{\mathrm{Var}[R^{\text{mv}}]}.$$

Substituting this back in, we get

$$\mathrm{E}[R_i] = \alpha + \frac{\mathrm{Cov}[R_i, R^{\text{mv}}]}{\mathrm{Var}[R^{\text{mv}}]}\left(\mathrm{E}[R^{\text{mv}}] - \alpha\right) = \alpha + \beta[R_i, R^{\text{mv}}]\left(\mathrm{E}[R^{\text{mv}}] - \alpha\right),$$

which verifies that R^{mv} is a valid pricing factor.

("Only if" part.) Next, let us show that if a return works as a pricing factor, it must be mean-variance efficient and different from the minimum-variance return. This proof is due to Hansen and Richard (1987). Assume that R^{mv} is a pricing factor. Decomposing as in Theorem 9.6

$$R^{\text{mv}} = R^* + wR^{e*} + \eta,$$

we need to show that $\eta = 0$ (so R^{mv} is mean-variance efficient) and that $w \neq \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]/\operatorname{Var}[R^{e*}]$ (so R^{mv} is different from the minimum-variance return).

Define a new return R as the "efficient part" of R^{mv}, i.e.

$$R = R^* + wR^{e*}.$$

Since

$$\operatorname{Cov}[\eta, R^{\text{mv}}] = \mathrm{E}[\eta R^{\text{mv}}] - \mathrm{E}[\eta]\,\mathrm{E}[R^{\text{mv}}] = \mathrm{E}[\eta R^{\text{mv}}] = \mathrm{E}[\eta^2],$$

we get

$$\operatorname{Cov}[R, R^{\text{mv}}] = \operatorname{Cov}[R^{\text{mv}} - \eta, R^{\text{mv}}] = \operatorname{Cov}[R^{\text{mv}}, R^{\text{mv}}] - \operatorname{Cov}[\eta, R^{\text{mv}}] = \operatorname{Var}[R^{\text{mv}}] - \mathrm{E}[\eta^2].$$

On the other hand, $\mathrm{E}[R] = \mathrm{E}[R^{\text{mv}}]$ so applying (9.32) for $R_i = R^{\text{mv}}$, we obtain

$$\mathrm{E}[R^{\text{mv}}] - \alpha = \frac{\operatorname{Cov}[R, R^{\text{mv}}]}{\operatorname{Var}[R^{\text{mv}}]}\,(\mathrm{E}[R^{\text{mv}}] - \alpha)$$

and hence $\operatorname{Cov}[R, R^{\text{mv}}] = \operatorname{Var}[R^{\text{mv}}]$. We conclude that $\mathrm{E}[\eta^2] = 0$, which implies $\eta = 0$.

Suppose that $w = \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}]/\operatorname{Var}[R^{e*}]$ and define a new gross return

$$R = R^{\text{mv}} + \frac{1}{\mathrm{E}[R^{e*}]} R^{e*}.$$

Clearly, $\mathrm{E}[R] = \mathrm{E}[R^{\text{mv}}] + 1$, and furthermore

$$\operatorname{Cov}[R, R^{\text{mv}}] = \operatorname{Var}[R^{\text{mv}}] + \frac{1}{\mathrm{E}[R^{e*}]} \operatorname{Cov}[R^{e*}, R^{\text{mv}}] = \operatorname{Var}[R^{\text{mv}}]$$

since

$$\begin{aligned}\operatorname{Cov}[R^{e*}, R^{\text{mv}}] &= \operatorname{Cov}[R^{e*}, R^* + wR^{e*}] = \operatorname{Cov}[R^{e*}, R^*] + w \operatorname{Var}[R^{e*}] \\ &= \operatorname{Cov}[R^{e*}, R^*] + \mathrm{E}[R^*]\,\mathrm{E}[R^{e*}] = \mathrm{E}[R^{e*} R^*] = 0.\end{aligned}$$

Applying (9.32) to the return R, we get

$$\mathrm{E}[R] = \alpha + \frac{\operatorname{Cov}[R, R^{\text{mv}}]}{\operatorname{Var}[R^{\text{mv}}]}\,(\mathrm{E}[R^{\text{mv}}] - \alpha) = \alpha + (\mathrm{E}[R^{\text{mv}}] - \alpha) = \mathrm{E}[R^{\text{mv}}],$$

which contradicts our early conclusion that $\mathrm{E}[R] = \mathrm{E}[R^{\text{mv}}] + 1$. Hence our assumption about w cannot hold. □

One implication of this theorem is that we can always find some returns that work as a pricing factor, namely the mean-variance efficient returns. Another implication is that the conclusion of the classical CAPM can be restated as "the market portfolio is mean-variance efficient."

9.5 Pricing factors in a multi-period framework

In the one-period framework we defined a pricing factor to be a K-dimensional random variable $\boldsymbol{x}$ such that there exists some $\alpha \in \mathbb{R}$ and some $\boldsymbol{\eta} \in \mathbb{R}^K$ so that

$$\mathrm{E}[R_i] = \alpha + \boldsymbol{\beta}[R_i, \boldsymbol{x}]^\top \boldsymbol{\eta}, \quad i = 1, \ldots, I,$$

where $\boldsymbol{\beta}[R_i, \boldsymbol{x}] = (\mathrm{Var}[\boldsymbol{x}])^{-1} \mathrm{Cov}[\boldsymbol{x}, R_i]$. We saw that any state-price deflator works as a pricing factor and, more generally, if $\zeta = a + \boldsymbol{b}^\top \boldsymbol{x}$ is a state-price deflator for constants $a, \boldsymbol{b}$ then $\boldsymbol{x}$ is a pricing factor. On the other hand, given any pricing factor $\boldsymbol{x}$, we can find constants $a, \boldsymbol{b}$ such that $\zeta = a + \boldsymbol{b}^\top \boldsymbol{x}$ is a candidate state-price deflator (not necessarily strictly positive, alas).

In a multi-period discrete-time framework we will say that a K-dimensional adapted stochastic process $\boldsymbol{x} = (\boldsymbol{x}_t)$ is a **conditional pricing factor**, if there exist adapted stochastic processes $\alpha = (\alpha_t)$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_t)$ so that

$$\mathrm{E}_t[R_{i,t+1}] = \alpha_t + \boldsymbol{\beta}_t[R_{i,t+1}, \boldsymbol{x}_{t+1}]^\top \boldsymbol{\eta}_t, \quad i = 1, \ldots, I, \tag{9.33}$$

for any $t = 0, 1, 2, \ldots, T-1$. Here, the conditional factor beta is defined as

$$\boldsymbol{\beta}_t[R_{i,t+1}, \boldsymbol{x}_{t+1}] = (\mathrm{Var}_t[\boldsymbol{x}_{t+1}])^{-1} \mathrm{Cov}_t[\boldsymbol{x}_{t+1}, R_{i,t+1}]. \tag{9.34}$$

If a conditionally risk-free asset exists, then $\alpha_t = R_t^f$ implying that

$$\mathrm{E}_t[R_{i,t+1}] = R_t^f + \boldsymbol{\beta}_t[R_{i,t+1}, \boldsymbol{x}_{t+1}]^\top \boldsymbol{\eta}_t, \quad i = 1, \ldots, I. \tag{9.35}$$

Suppose $\boldsymbol{x}$ is a conditional pricing factor, and let $a = (a_t)$ be an adapted one-dimensional process and $\underline{\underline{A}} = (\underline{\underline{A}}_t)$ be an adapted process whose values $\underline{\underline{A}}_t$ are non-singular $K \times K$ matrices. Then $\hat{\boldsymbol{x}}$ defined by

$$\hat{\boldsymbol{x}}_{t+1} = a_t + \underline{\underline{A}}_t \boldsymbol{x}_{t+1} \tag{9.36}$$

will also be a conditional pricing factor.

If $\zeta = (\zeta_t)$ is a state-price deflator process, the one-period analysis implies that the ratios ζ_{t+1}/ζ_t define a conditional pricing factor. Since $\zeta_{t+1} = 0 + \zeta_t(\zeta_{t+1}/\zeta_t)$ is a transformation of the form (9.36), we see that any state-price deflator is a conditional pricing factor. As in the one-period case, we will have that if $\zeta_{t+1} = a_t + \boldsymbol{b}_t^\top \boldsymbol{x}_{t+1}$ is a state-price deflator for some adapted process $\boldsymbol{x}$, then $\boldsymbol{x}$ is a conditional pricing factor. And for any conditional pricing factor $\boldsymbol{x}$, we can find adapted process $a = (a_t)$ and $\boldsymbol{b} = (\boldsymbol{b}_t)$ so that $\zeta_{t+1} = a_t + \boldsymbol{b}_t^\top \boldsymbol{x}_{t+1}$ defines a candidate state-price deflator (not necessarily positive, however).

We will say that a K-dimensional adapted stochastic process $\boldsymbol{x} = (\boldsymbol{x}_t)$ is an **unconditional pricing factor**, if there exist constants α and $\boldsymbol{\eta}$ so that

$$\mathrm{E}[R_{i,t+1}] = \alpha + \boldsymbol{\beta}[R_{i,t+1}, \boldsymbol{x}_{t+1}]^\top \boldsymbol{\eta}, \quad i = 1, \ldots, I, \tag{9.37}$$

for any $t = 0, 1, 2, \ldots, T-1$. Here, the unconditional factor beta is defined as

$$\boldsymbol{\beta}[R_{i,t+1}, \boldsymbol{x}_{t+1}] = (\mathrm{Var}[\boldsymbol{x}_{t+1}])^{-1} \mathrm{Cov}[\boldsymbol{x}_{t+1}, R_{i,t+1}]. \tag{9.38}$$

This is true if the state-price deflator can be written as $\frac{\zeta_{t+1}}{\zeta_t} = a + \boldsymbol{b}^\top \boldsymbol{x}_{t+1}$ for constants a and $\boldsymbol{b}$. Hence, an unconditional pricing factor is also a conditional pricing factor. The converse is not true.

Testing factor models on actual data really requires an unconditional model since we need to replace expected returns by average returns, etc. To go from a conditional pricing factor model to an unconditional, we need to link the variation in the coefficients over time to some observable variables. Suppose for example that $\frac{\zeta_{t+1}}{\zeta_t} = a_t + b_t x_{t+1}$ is a state-price deflator so that $x = (x_t)$ is a *conditional* pricing factor, assumed to be one-dimensional for notational simplicity. If we can write

$$a_t = A_0 + A_1 y_t, \quad b_t = B_0 + B_1 y_t$$

for some observable adapted process $y = (y_t)$, then

$$\frac{\zeta_{t+1}}{\zeta_t} = A_0 + A_1 y_t + B_0 x_{t+1} + B_1 y_t x_{t+1}$$

which defines a 3-dimensional *unconditional* pricing factor given by the vector $(y_t, x_{t+1}, y_t x_{t+1})^\top$.

Now let us turn to the continuous-time setting. By a K-dimensional conditional pricing factor in a continuous-time model we mean an adapted K-dimensional process $\boldsymbol{x} = (\boldsymbol{x}_t)$ with the property that there exist some one-dimensional adapted process $\alpha = (\alpha_t)$ and some K-dimensional adapted process $\boldsymbol{\eta} = (\boldsymbol{\eta}_t)$ such that for any asset i (or trading strategy), the expected rate of return per time period satisfies

$$\mu_{it} + \delta_{it} = \alpha_t + (\boldsymbol{\beta}_t^{ix})^\top \boldsymbol{\eta}_t, \tag{9.39}$$

where again $\boldsymbol{\beta}_t^{i\boldsymbol{x}}$ is the factor-beta of asset i at time t. To understand the factor-beta write the price dynamics of risky assets in the usual form

$$dP_{it} = P_{it}\left[\mu_{it}\, dt + \boldsymbol{\sigma}_{it}^\top\, d\boldsymbol{z}_t\right].$$

and the dynamics of $\boldsymbol{x}$ as

$$d\boldsymbol{x}_t = \boldsymbol{\mu}_{xt}\, dt + \underline{\underline{\sigma}}_{xt}\, d\boldsymbol{z}_t,$$

where $\boldsymbol{\mu}_x$ is an adapted process valued in $\mathbb{R}^K$ and $\underline{\underline{\sigma}}_x$ is an adapted process with values being $K \times d$ matrices. Then the factor-beta is defined as

$$\boldsymbol{\beta}_t^{ix} = \left(\underline{\underline{\sigma}}_{xt}\underline{\underline{\sigma}}_{xt}^\top\right)^{-1} \underline{\underline{\sigma}}_{xt}\boldsymbol{\sigma}_{it}. \tag{9.40}$$

If a "bank account" is traded, it then follows that $\alpha_t = r_t^f$. In the following we will assume that this is the case.

Factors are closely linked to market prices of risk and hence to risk-neutral measures and state-price deflators. If $\boldsymbol{x} = (\boldsymbol{x}_t)$ is a factor in an expected return-beta relation, then we can define a market price of risk as (note that it is d-dimensional)

$$\boldsymbol{\lambda}_t = \underline{\underline{\sigma}}_{xt}^\top \left(\underline{\underline{\sigma}}_{xt}\underline{\underline{\sigma}}_{xt}^\top\right)^{-1} \boldsymbol{\eta}_t,$$

since we then have

$$\boldsymbol{\sigma}_{it}^\top \boldsymbol{\lambda}_t = \boldsymbol{\sigma}_{it}^\top \underline{\underline{\sigma}}_{xt}^\top \left(\underline{\underline{\sigma}}_{xt}\underline{\underline{\sigma}}_{xt}^\top\right)^{-1} \boldsymbol{\eta}_t = \left(\boldsymbol{\beta}_t^{ix}\right)^\top \boldsymbol{\eta}_t = \mu_{it} + \delta_{it} - r_t^f.$$

Conversely, let $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_t)$ be any market price of risk and let $\zeta = (\zeta_t)$ be the associated state-price deflator so that

$$d\zeta_t = -\zeta_t \left(r_t^f\, dt + \boldsymbol{\lambda}_t^\top\, d\boldsymbol{z}_t\right).$$

Then we can use ζ as a one-dimensional factor in an expected return-beta relation. Since this corresponds to a factor "sensitivity" vector $-\zeta_t \boldsymbol{\lambda}_t$ replacing the matrix $\underline{\underline{\sigma}}_{xt}$, the relevant "beta" is

$$\beta_t^{i\zeta} = \frac{-\zeta_t \boldsymbol{\sigma}_{it}^\top \boldsymbol{\lambda}_t}{\zeta_t^2 \boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t} = -\frac{1}{\zeta_t} \frac{\boldsymbol{\sigma}_{it}^\top \boldsymbol{\lambda}_t}{\boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t}.$$

We can use $\eta_t = -\zeta_t \boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t$, since then

$$\beta_t^{i\zeta} \eta_t = -\frac{1}{\zeta_t} \frac{\boldsymbol{\sigma}_{it}^\top \boldsymbol{\lambda}_t}{\boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t} \left(-\zeta_t \boldsymbol{\lambda}_t^\top \boldsymbol{\lambda}_t\right) = \boldsymbol{\sigma}_{it}^\top \boldsymbol{\lambda}_t = \mu_{it} + \delta_{it} - r_t^f$$

for any asset i. We can even use $a_t + b_t \zeta_t$ as a factor for any sufficiently well-behaved adapted processes $a = (a_t)$ and $b = (b_t)$.

If we use ζ^* as the factor, the relevant η is $\eta_t = -\zeta_t^* \left(\boldsymbol{\lambda}_t^*\right)^\top \boldsymbol{\lambda}_t^* = -\zeta_t^* \|\boldsymbol{\lambda}_t^*\|^2$. From (4.46), we see that $\|\boldsymbol{\lambda}_t^*\|^2$ is exactly the excess expected rate of return of the growth-optimal strategy, we which can also write as $\mu_t^* - r_t^f$. Hence, we can write the excess expected rate of return on any asset (or trading strategy) as

$$\mu_{it} + \delta_{it} - r_t^f = \beta_t^{i\zeta^*} \left(-\zeta_t^* [\mu_t^* - r_t^f]\right) = \frac{\boldsymbol{\sigma}_{it}^\top \boldsymbol{\lambda}_t^*}{\left(\boldsymbol{\lambda}_t^*\right)^\top \boldsymbol{\lambda}_t^*} [\mu_t^* - r_t^f] \equiv \beta_t^{i\boldsymbol{\lambda}^*} [\mu_t^* - r_t^f]. \tag{9.41}$$

Whether we want to use a discrete-time or a continuous-time model, the key question is what factors to include in order to get prices or returns that are consistent with the data. Due to the link between state-price deflators and (marginal utility of) consumption, we should look for factors among variables that may affect (marginal utility of) consumption.

9.6 Empirical factors

A large part of the literature on factor models is based on empirical studies which for a given data set identifies a number of priced factors so that most of the differences between the returns on different financial assets—typically only various portfolios of stocks—can be explained by their different factor betas. The best known studies of this kind were carried out by Fama and French, who find support for a model with three factors:

1. the return on a broad stock market index;
2. the return on a portfolio of stocks in small companies (according to the market value of all stocks issued by the firm) minus the return on a portfolio of stocks in large companies;
3. the return on a portfolio of stocks issued by firms with a high book-to-market value (the ratio between the book value of the assets of the firm to the market value of all the assets) minus the return on a portfolio of stocks in firms with a low book-to-market value.

According to Fama and French (1996) such a model gives a good fit of U.S. stock market data over the period 1963–1993. However, the empirical analysis does not explain *why* this three-factor model does well and what the underlying pricing mechanisms might be. Also note that while a given factor model does well in a given market over a given period it may perform very badly in other markets and/or other periods (and risk premia may be different for other data sets). Some recent studies indicate that the second of the three factors (the "size effect") seems to have disappeared. Another critique is that over the last 30 years empirical researchers have tried so many factors that

it is hardly surprising that they have found some statistically significant factors. In fact, we know that in any given sample of historical returns it is possible to find a portfolio so that a factor model with the return on this portfolio as the only factor will perfectly explain all returns in the sample! Also note that the Fama-French model is a partial pricing model since the factors themselves are derived from prices of financial assets. For these reasons the purely empirically based "models" do not contribute much to the understanding of the pricing mechanisms of financial markets. However, some interesting recent studies explore whether the Fama-French factors can be seen as proxies for macro-economic variables that can logically be linked to asset prices.

Linking Fama-French factors to risk and variations in investment opportunities: Liew and Vassalou (2000), Lettau and Ludvigson (2001), Vassalou (2003), Petkova (2006).

Data snooping/biases explanations of the success of FF: Lo and MacKinlay (1990), Kothari, Shanken, and Sloan (1995).

Problems in measurement of beta: Berk, Green, and Naik (1999), Gomes, Kogan, and Zhang (2003).

9.7 Theoretical factors

Factor models can be obtained through the general consumption-based asset pricing model by relating optimal consumption to various factors. As discussed in Section 6.5 the optimal consumption plan of an individual with time-additive expected utility must satisfy the so-called envelope condition

$$u'(c_t) = J_W(W_t, x_t, t). \tag{6.47}$$

Here J is the indirect utility function of the individual, i.e. the maximum obtainable expected utility of future consumption. W_t is the financial wealth of the investor at time t. x_t is the time t value of a variable that captures the variations in investment opportunities (captured by the risk-free interest rate, expected returns and volatilities on risky assets, and correlations between risky assets) and investor-specific variables (e.g. labor income). For notational simplicity, x is assumed to be one-dimensional, but this could be generalized.

In a continuous-time framework write the dynamics of wealth compactly as

$$dW_t = W_t \left[\mu_{Wt}\, dt + \boldsymbol{\sigma}_{Wt}^\top\, d\boldsymbol{z}_t\right]$$

and assume that the state variable x follows a diffusion process

$$dx_t = \mu_{xt}\, dt + \boldsymbol{\sigma}_{xt}^\top\, d\boldsymbol{z}_t, \qquad \mu_{xt} = \mu_x(x_t, t), \quad \boldsymbol{\sigma}_{xt} = \boldsymbol{\sigma}_x(x_t, t).$$

From (8.14) it follows that the state-price deflator derived from this individual can be written as

$$\zeta_t = e^{-\delta t}\, \frac{J_W(W_t, x_t, t)}{J_W(W_0, x_0, 0)}.$$

An application of Itô's Lemma yields a new expression for the dynamics of ζ, which again can be compared with (4.37). It follows from this comparison that

$$\boldsymbol{\lambda}_t = \left(\frac{-W_t J_{WW}(W_t, x_t, t)}{J_W(W_t, x_t, t)}\right) \boldsymbol{\sigma}_{Wt} + \left(\frac{-J_{Wx}(W_t, x_t, t)}{J_W(W_t, x_t, t)}\right) \boldsymbol{\sigma}_{xt},$$

is a market price of risk. Consequently, the expected excess rate of return on asset i can be written as

$$\mu_{it} + \delta_{it} - r_t^f = \left(\frac{-W_t J_{WW}(W_t, x_t, t)}{J_W(W_t, x_t, t)}\right) \boldsymbol{\sigma}_{it}^\top \boldsymbol{\sigma}_{Wt} + \left(\frac{-J_{Wx}(W_t, x_t, t)}{J_W(W_t, x_t, t)}\right) \boldsymbol{\sigma}_{it}^\top \boldsymbol{\sigma}_{xt}, \tag{9.42}$$

which can be rewritten as

$$\mu_{it} + \delta_{it} - r_t^f = \beta_{iWt}\eta_{Wt} + \beta_{ixt}\eta_{xt}, \tag{9.43}$$

where

$$\beta_{iWt} = \frac{\boldsymbol{\sigma}_{it}^\top \boldsymbol{\sigma}_{Wt}}{\|\boldsymbol{\sigma}_{Wt}\|^2}, \quad \beta_{ixt} = \frac{\boldsymbol{\sigma}_{it}^\top \boldsymbol{\sigma}_{xt}}{\|\boldsymbol{\sigma}_{xt}\|^2},$$
$$\eta_{Wt} = \|\boldsymbol{\sigma}_{Wt}\|^2 \left(\frac{-W_t J_{WW}(W_t, x_t, t)}{J_W(W_t, x_t, t)}\right), \quad \eta_{xt} = \|\boldsymbol{\sigma}_{xt}\|^2 \left(\frac{-J_{Wx}(W_t, x_t, t)}{J_W(W_t, x_t, t)}\right).$$

We now have a continuous-time version of (9.35) with the wealth of the individual and the state variable as the factors. If it takes m state variables to describe the variations in investment opportunities, labor income, etc., we get an $(m+1)$-factor model.

If the individual is taken to be a representative individual, her wealth will be identical to the aggregate value of all assets in the economy, including all traded financial assets and non-traded asset such as human capital. This is like the market portfolio in the traditional static CAPM. The first term on the right-hand side of (9.42) is then the product of the relative risk aversion of the representative individual (derived from her indirect utility) and the covariance between the rate of return on asset i and the rate of return on the market portfolio. In the special case where the indirect utility is a function of wealth and time only, the last term on the right-hand side will be zero, and we get the well-known relation

$$\mu_{it} + \delta_{it} - r_t^f = \beta_{iWt}\left(\mu_{Wt} + \delta_{Wt} - r_t^f\right),$$

where β_{iWt} is the "market-beta" of asset i. This is a continuous-time version of the traditional static CAPM. This is only true under the strong assumption that individuals do not care about variations in investment opportunities, income, etc. In general we have to add factors describing the future investment opportunities, future labor income, etc. This extension of the CAPM is called the Intertemporal CAPM and was first derived by Merton (1973b).

Only few empirical studies of factor models refer to Merton's Intertemporal CAPM when motivating the choice of factors. Brennan, Wang, and Xia (2004) set up a simple model with the short-term real interest rate and the slope of the capital market line as the factors since these variables capture the investment opportunities. In an empirical test, this model performs as well as the Fama-French model, which is encouraging for the development of theoretically well-founded and empirically viable factor models.

9.8 Exercises

EXERCISE 9.1 Consider a discrete-time economy with a one-dimensional conditional pricing factor $x = (x_t)$ so that, for some adapted processes $\alpha = (\alpha_t)$ and some $\eta = (\eta_t)$,

$$\mathrm{E}_t[R_{i,t+1}] = \alpha_t + \beta_t[R_{i,t+1}, x_{t+1}]\,\eta_t,$$

for all assets i and all $t = 0, 1, \ldots, T-1$.

(a) Show that

$$\frac{\zeta_{t+1}}{\zeta_t} = \frac{1}{\alpha_t}\left(1 - \frac{x_{t+1} - \mathrm{E}_t[x_{t+1}]}{\mathrm{Var}_t[x_{t+1}]}\eta_t\right)$$

satisfies the pricing condition for a state-price deflator, i.e.

$$\mathrm{E}_t\left[\frac{\zeta_{t+1}}{\zeta_t}R_{i,t+1}\right] = 1$$

for all assets i.

(b) Show that ζ_{t+1}/ζ_t is only a true one-period state-price deflator for the period between time t and time $t+1$ if you to impose some condition on parameters and/or distributions and provide that condition. If this condition is not satisfied, what can you conclude about the asset prices in this economy?

(c) Now suppose that the factor is the return on some particular portfolio, i.e. $x_{t+1} = \tilde{R}_{t+1}$. Answer question (b) again.

(d) Consider the good (?) old (!) CAPM where $x_{t+1} = R_{M,t+1}$, the return on the market portfolio. Suppose that $\mathrm{E}_t[R_{M,t+1}] = 1.05$, $\sigma_t[R_{M,t+1}] = 0.2$, and $R_t^f = 1.02$. What do you need to assume about the distribution of the market return to ensure that the model is free of arbitrage? Is an assumption like this satisfied in typical derivations of the CAPM?

EXERCISE 9.2 Using the orthogonal characterization of the mean-variance frontier, show that for any mean-variance efficient return $R^{\boldsymbol{\pi}}$ different from the minimum-variance portfolio there is a unique mean-variance efficient return $R^{z(\boldsymbol{\pi})}$ with $\mathrm{Cov}[R^{\boldsymbol{\pi}}, R^{z(\boldsymbol{\pi})}] = 0$. Show that

$$\mathrm{E}[R^{z(\boldsymbol{\pi})}] = \mathrm{E}[R^*] - \mathrm{E}[R^{e*}]\frac{\mathrm{E}[(R^*)^2] - \mathrm{E}[R^{\boldsymbol{\pi}}]\,\mathrm{E}[R^*]}{\mathrm{E}[R^{\boldsymbol{\pi}}] - \mathrm{E}[R^*] - \mathrm{E}[R^{\boldsymbol{\pi}}]\,\mathrm{E}[R^{e*}]}.$$

EXERCISE 9.3 Consider a discrete-time economy in which asset prices are described by an unconditional linear factor model

$$\frac{\zeta_{t+1}}{\zeta_t} = a + \boldsymbol{b}\cdot\boldsymbol{x}_{t+1}, \quad t = 0, 1, \ldots, T-1,$$

where the conditional mean and second moments of the factor are constant, i.e. $\mathrm{E}_t[\boldsymbol{x}_{t+1}] = \boldsymbol{\mu}$ and $\mathrm{E}_t[\boldsymbol{x}_{t+1}\boldsymbol{x}_{t+1}^\top] = \underline{\underline{\Sigma}}$ for all t.

(a) What is the one-period risk-free rate of return r_t^f? What is the time t annualized yield $\hat{y}_t^{t+s}$ on a zero-coupon bond maturing at time $t+s$? (If B_t^{t+s} denotes the price of the bond, the annualized gross yield is defined by the equation $B_t^{t+s} = (1+\hat{y}_t^{t+s})^{-s}$.)

(b) What can you say about the expected excess one-period returns on risky assets?

You want to value an uncertain stream of dividends $D = (D_t)$. You are told that dividends evolve as

$$\frac{D_{t+1}}{D_t} = m + \boldsymbol{\psi}\cdot\boldsymbol{x}_{t+1} + \varepsilon_{t+1}, \quad t = 0, 1, \ldots, T-1,$$

where $\mathrm{E}_t[\varepsilon_{t+1}] = 0$ and $\mathrm{E}_t[\varepsilon_{t+1}\boldsymbol{x}_{t+1}] = \boldsymbol{0}$ for all t.

(c) Show that, for any $t = 0, 1, \ldots, T-1$,

$$\mathrm{E}_t\left[\frac{D_{t+1}}{D_t}\frac{\zeta_{t+1}}{\zeta_t}\right] = ma + (m\boldsymbol{b} + a\boldsymbol{\psi}) \cdot \boldsymbol{\mu} + \boldsymbol{\psi}^\top \underline{\underline{\Sigma}} \boldsymbol{b} \equiv A$$

(d) Show that

$$\mathrm{E}_t\left[\frac{D_{t+s}}{D_t}\frac{\zeta_{t+s}}{\zeta_t}\right] = A^s.$$

(e) Show that the value at time t of the future dividends is

$$P_t = D_t \frac{A}{1-A}\left(1 - A^{T-t}\right).$$

From (4.27) we know that we can also value the dividends by the formula

$$P_t = \sum_{s=1}^{T-t} \frac{\mathrm{E}_t[D_{t+s}] - \beta_t\left[D_{t+s}, \frac{\zeta_{t+s}}{\zeta_t}\right]\eta_{t,t+s}}{(1+\hat{y}_t^{t+s})^s}, \qquad (*)$$

where

$$\beta_t\left[D_{t+s}, \frac{\zeta_{t+s}}{\zeta_t}\right] = \mathrm{Cov}_t\left[D_{t+s}, \frac{\zeta_{t+s}}{\zeta_t}\right] / \mathrm{Var}_t\left[\frac{\zeta_{t+s}}{\zeta_t}\right], \quad \eta_{t,t+s} = -\mathrm{Var}_t\left[\frac{\zeta_{t+s}}{\zeta_t}\right] / \mathrm{E}_t\left[\frac{\zeta_{t+s}}{\zeta_t}\right].$$

In the next questions you have to compute the ingredients to this valuation formula.

(f) Show that $\mathrm{E}_t[D_{t+s}] = D_t\,(m + \boldsymbol{\psi} \cdot \boldsymbol{\mu})^s$.

(g) Compute $\mathrm{Cov}_t\left[D_{t+s}, \frac{\zeta_{t+s}}{\zeta_t}\right]$.

(h) Compute $\beta_t\left[D_{t+s}, \frac{\zeta_{t+s}}{\zeta_t}\right]$.

(i) Compute $\eta_{t,t+s}$

(j) Verify that the time t value of the future dividends satisfies (*).

EXERCISE 9.4 In a continuous-time framework an individual with time-additive expected power utility induces the state-price deflator

$$\zeta_t = e^{-\delta t}\left(\frac{c_t}{c_0}\right)^{-\gamma},$$

where γ is the constant relative risk aversion, δ is the subjective time preference rate, and $c = (c_t)_{t\in[0,T]}$ is the optimal consumption process of the individual. If the dynamics of the optimal consumption process is of the form

$$dc_t = c_t\left[\mu_{ct}\,dt + \boldsymbol{\sigma}_{ct}^\top\,d\boldsymbol{z}_t\right]$$

then the dynamics of the state-price deflator is

$$d\zeta_t = -\zeta_t\left[\left(\delta + \gamma\mu_{ct} - \frac{1}{2}\gamma(1+\gamma)\|\boldsymbol{\sigma}_{ct}\|^2\right)dt + \gamma\boldsymbol{\sigma}_{ct}^\top\,d\boldsymbol{z}_t\right].$$

(a) State the market price of risk $\boldsymbol{\lambda}_t$ in terms of the preference parameters and the expected growth rate and sensitivity of the consumption process.

In many concrete models of the individual consumption and portfolio decisions, the optimal consumption process will be of the form

$$c_t = W_t e^{f(X_t,t)},$$

where W_t is the wealth of the individual at time t, X_t is the time t value of some state variable, and f is some smooth function. Here X can potentially be multi-dimensional.

(b) Give some examples of variables other than wealth that may affect the optimal consumption of an individual and which may therefore play the role of X_t.

Suppose the state variable X is one-dimensional and write the dynamics of the wealth of the individual and the state variable as

$$\begin{aligned} dW_t &= W_t \left[\left(\mu_{Wt} - e^{f(X_t,t)} \right) dt + \boldsymbol{\sigma}_{Wt}^\top \, d\boldsymbol{z}_t \right], \\ dX_t &= \mu_{Xt} \, dt + \boldsymbol{\sigma}_{Xt}^\top \, d\boldsymbol{z}_t \end{aligned}$$

(c) Characterize the market price of risk in terms of the preference parameters and the drift and sensitivity terms of W_t and X_t.

Hint: Apply Itô's Lemma to $c_t = W_t e^{f(X_t,t)}$ to express the required parts of the consumption process in terms of W and X.

(d) Show that the instantaneous excess expected rate of return on risky asset i can be written as

$$\mu_{it} + \delta_{it} - r_t^f = \beta_{iW,t} \eta_{Wt} + \beta_{iX,t} \eta_{Xt},$$

where $\beta_{iW,t}$ and $\beta iX, t$ are the instantaneous beta's of the asset with respect to wealth and the state variable, respectively. Relate η_{Wt} and η_{Xt} to preference parameters and the drift and sensitivity terms of W and X.

Chapter 10

The economics of the term structure of interest rates

10.1 Introduction

The previous two chapters focused on the implications of asset pricing models for the level of stock market excess returns and the cross-section of stock returns. In this chapter focuses on the consequences of asset pricing theory for the pricing of bonds and the term structure of interest rates implied by bond prices.

A bond is nothing but a standardized and transferable loan agreement between two parties. The issuer of the bond is borrowing money from the holder of the bond and promises to pay back the loan according to a predefined payment scheme. The presence of the bond market allows individuals to trade consumption opportunities at different points in time among each other. An individual who has a clear preference for current capital to finance investments or current consumption can borrow by issuing a bond to an individual who has a clear preference for future consumption opportunities. The price of a bond of a given maturity is, of course, set to align the demand and supply of that bond, and will consequently depend on the attractiveness of the real investment opportunities and on the individuals' preferences for consumption over the maturity of the bond. The term structure of interest rates will reflect these dependencies.

After a short introduction to notation and bond market terminology in Section 10.2, we derive in Sections 10.3 and 10.4 relations between equilibrium interest rates and aggregate consumption and production in settings with a representative individual. In Section 10.5 we give some examples of equilibrium term structure models that are derived from the basic relations between interest rates, consumption, and production. The famous Vasicek model and Cox-Ingersoll-Ross model are presented.

Since individuals are concerned with the number of units of goods they consume and not the dollar value of these goods, the relations found in those sections apply to real interest rates. However, most traded bonds are nominal, i.e. they promise the delivery of certain dollar amounts, not the delivery of a certain number of consumption goods. The real value of a nominal bond depends on the evolution of the price of the consumption good. In Section 10.6 we explore the relations between real rates, nominal rates, and inflation. We consider both the case where money

has no real effects on the economy and the case where money does affect the real economy.

The development of arbitrage-free dynamic models of the term structure was initiated in the 1970s. Until then, the discussions among economists about the shape of the term structure were based on some relatively loose hypotheses. The most well-known of these is the expectation hypothesis, which postulates a close relation between current interest rates or bond yields and expected future interest rates or bond returns. Many economists still seem to rely on the validity of this hypothesis, and a lot of man power has been spend on testing the hypothesis empirically. In Section 10.7, we review several versions of the expectation hypothesis and discuss the consistency of these versions. We argue that neither of these versions will hold for any reasonable dynamic term structure model. Some alternative traditional hypotheses are briefly reviewed in Section 10.8.

10.2 Basic interest rate concepts and relations

As in earlier chapters we will denote by B_t^T the price at time t of a zero-coupon bond paying a dividend of one at time T and no other dividends. If many zero-coupon bonds with different maturities are traded, we can form the function $T \mapsto B_t^T$, which we refer to as the **discount function** prevailing at time t. Of course, we must have $B_t^t = 1$, and we expect the discount function to be decreasing since all individuals will presumably prefer getting the dividend sooner than later.

Next, consider a coupon bond with payment dates $t_1, t_2, \ldots, t_n$, where we assume without loss of generality that $t_1 < t_2 < \cdots < t_n$. The payment at date t_i is denoted by Y_i. Such a coupon bond can be seen as a portfolio of zero-coupon bonds, namely a portfolio of Y_1 zero-coupon bonds maturing at t_1, Y_2 zero-coupon bonds maturing at t_2, etc. If all these zero-coupon bonds are traded in the market, the unique no-arbitrage price of the coupon bond at any time t is

$$B_t = \sum_{t_i > t} Y_i B_t^{t_i}. \tag{10.1}$$

where the sum is over all future payment dates of the coupon bond. If not all the relevant zero-coupon bonds are traded, we cannot justify the relation (10.1) as a result of the no-arbitrage principle. Still, it is a valuable relation. Suppose that an investor has determined (from private or macro economic information) a discount function showing the value *she* attributes to payments at different future points in time. Then she can value all sure cash flows in a consistent way by substituting that discount function into (10.1).

The information incorporated in prices of the many different bonds is usually better understood when transforming the bond prices into interest rates. Interest rates are always quoted on an annual basis, i.e. as some percentage per year. However, to apply and assess the magnitude of an interest rate, we also need to know the compounding frequency of that rate. More frequent compounding of a given interest rate per year results in higher "effective" interest rates. Furthermore, we need to know at which time the interest rate is set or observed and for which period of time the interest rate applies. Spot rates applies to a period beginning at the time the rate is set, whereas forward rates applies to a future period of time. The precise definitions follow below.

Annual compounding

Given the price B_t^T at time t on a zero-coupon bond maturing at time T, the relevant discount rate between time t and time T is the yield on the zero-coupon bond, the so-called **zero-coupon rate** or **spot rate** for date T. That $\hat{y}_t^T$ is the annually compounded zero-coupon rate means that

$$B_t^T = (1+\hat{y}_t^T)^{-(T-t)} \qquad \Leftrightarrow \qquad \hat{y}_t^T = \left(B_t^T\right)^{-1/(T-t)} - 1. \tag{10.2}$$

The zero-coupon rates as a function of maturity is called the **zero-coupon yield curve** or simply the **yield curve**. It is one way to express the term structure of interest rates.

While a zero-coupon or spot rate reflects the price on a loan between today and a given future date, a **forward rate** reflects the price on a loan between two future dates. The annually compounded relevant forward rate at time t for the period between time T and time S is denoted by $\hat{f}_t^{T,S}$. Here, we have $t \leq T < S$. This is the rate, which is appropriate at time t for discounting between time T and S. We can think of discounting from time S back to time t by first discounting from time S to time T and then discounting from time T to time t. We must therefore have that

$$\left(1+\hat{y}_t^S\right)^{-(S-t)} = \left(1+\hat{y}_t^T\right)^{-(T-t)} \left(1+\hat{f}_t^{T,S}\right)^{-(S-T)}, \tag{10.3}$$

from which we find that

$$\hat{f}_t^{T,S} = \frac{(1+\hat{y}_t^T)^{-(T-t)/(S-T)}}{(1+\hat{y}_t^S)^{-(S-t)/(S-T)}} - 1.$$

We can also link forward rates to bond prices:

$$B_t^S = B_t^T \left(1+\hat{f}_t^{T,S}\right)^{-(S-T)} \qquad \Leftrightarrow \qquad \hat{f}_t^{T,S} = \left(\frac{B_t^T}{B_t^S}\right)^{1/(S-T)} - 1. \tag{10.4}$$

Note that since $B_t^t = 1$, we have

$$\hat{f}_t^{t,S} = \left(\frac{B_t^t}{B_t^S}\right)^{1/(S-t)} - 1 = \left(B_t^S\right)^{-1/(S-t)} - 1 = \hat{y}_t^S,$$

i.e. the forward rate for a period starting today equals the zero-coupon rate or spot rate for the same period.

Compounding over other discrete periods – LIBOR rates

In practice, many interest rates are quoted using semi-annually, quarterly, or monthly compounding. An interest rate or R per year compounded m times a year, corresponds to a discount factor of $(1+R/m)^{-m}$ over a year. The annually compounded interest rate that corresponds to an interest rate of R compounded m times a year is $(1+R/m)^m - 1$. This is sometimes called the "effective" interest rate corresponding to the nominal interest rate R. Interest rates are set for loans with various maturities and currencies at the international money markets, the most commonly used being the LIBOR rates that are fixed in London. Traditionally, these rates are quoted using a compounding period equal to the maturity of the interest rate. If, for example, the three-month interest rate is $l_t^{t+0.25}$ per year, it means that

$$B_t^{t+0.25} = \frac{1}{1+0.25\, l_t^{t+0.25}} \qquad \Leftrightarrow \qquad l_t^{t+0.25} = \frac{1}{0.25}\left(\frac{1}{B_t^{t+0.25}} - 1\right).$$

More generally, the relations are

$$B_t^T = \frac{1}{1 + l_t^T(T-t)} \qquad \Leftrightarrow \qquad l_t^T = \frac{1}{T-t}\left(\frac{1}{B_t^T} - 1\right). \tag{10.5}$$

Similarly, discretely compounded forward rates can be computed as

$$L_t^{T,S} = \frac{1}{S-T}\left(\frac{B_t^T}{B_t^S} - 1\right). \tag{10.6}$$

Continuous compounding

Increasing the compounding frequency m, the effective annual return of one dollar invested at the interest rate R per year increases to e^R, due to the mathematical result saying that

$$\lim_{m\to\infty}\left(1 + \frac{R}{m}\right)^m = e^R.$$

A continuously compounded interest rate R is equivalent to an annually compounded interest rate of $e^R - 1$ (which is bigger than R). Similarly, the zero-coupon bond price B_t^T is related to the continuously compounded zero-coupon rate y_t^T by

$$B_t^T = e^{-y_t^T(T-t)} \qquad \Leftrightarrow \qquad y_t^T = -\frac{1}{T-t}\ln B_t^T. \tag{10.7}$$

The function $T \mapsto y_t^T$ is also a zero-coupon yield curve that contains exactly the same information as the discount function $T \mapsto B_t^T$ and also the same information as the annually compounded yield curve $T \mapsto \hat{y}_t^T$. The relation is $y_t^T = \ln(1 + \hat{y}_t^T)$.

If $f_t^{T,S}$ denotes the continuously compounded forward rate prevailing at time t for the period between T and S, we must have that $B_t^S = B_t^T e^{-f_t^{T,S}(S-T)}$, in analogy with (10.4). Consequently,

$$f_t^{T,S} = -\frac{\ln B_t^S - \ln B_t^T}{S-T} \tag{10.8}$$

and hence

$$f_t^{T,S} = \frac{y_t^S(S-t) - y_t^T(T-t)}{S-T}. \tag{10.9}$$

Analytical studies of the term structure of interest rates often focus on forward rates for future periods of infinitesimal length. The forward rate for an infinitesimal period starting at time T is simply referred to as the forward rate for time T and is defined as $f_t^T = \lim_{S\to T} f_t^{T,S}$. The function $T \mapsto f_t^T$ is called the **term structure of forward rates**. Assuming differentiability of the discount function, we get

$$f_t^T = -\frac{\partial \ln B_t^T}{\partial T} = -\frac{\partial B_t^T/\partial T}{B_t^T} \qquad \Leftrightarrow \qquad B_t^T = e^{-\int_t^T f_t^u\, du}. \tag{10.10}$$

Applying (10.9), the relation between the infinitesimal forward rate and the spot rates can be written as

$$f_t^T = \frac{\partial[y_t^T(T-t)]}{\partial T} = y_t^T + \frac{\partial y_t^T}{\partial T}(T-t) \tag{10.11}$$

under the assumption of a differentiable term structure of spot rates $T \mapsto y_t^T$. The forward rate reflects the slope of the zero-coupon yield curve. In particular, the forward rate f_t^T and the zero-coupon rate y_t^T will coincide if and only if the zero-coupon yield curve has a horizontal tangent at T. Conversely,

$$y_t^T = \frac{1}{T-t}\int_t^T f_t^u\, du, \tag{10.12}$$

i.e. the zero-coupon rate is an average of the forward rates.

It is important to realize that discount factors, spot rates, and forward rates (with any compounding frequency) are perfectly equivalent ways of expressing the same information. If a complete yield curve of, say, quarterly compounded spot rates is given, we can compute the discount function and spot rates and forward rates for any given period and with any given compounding frequency. If a complete term structure of forward rates is known, we can compute discount functions and spot rates, etc. Academics frequently apply continuous compounding since the mathematics involved in many relevant computations is more elegant when exponentials are used.

There are even more ways of representing the term structure of interest rates. Since most bonds are bullet bonds, many traders and analysts are used to thinking in terms of yields of bullet bonds rather than in terms of discount factors or zero-coupon rates. The **par yield** for a given maturity is the coupon rate that causes a bullet bond of the given maturity to have a price equal to its face value. Again we have to fix the coupon period of the bond. U.S. treasury bonds typically have semi-annual coupons which are therefore often used when computing par yields. Given a discount function $T \mapsto B_t^T$, the n-year par yield is the value of c satisfying

$$\sum_{i=1}^{2n} \left(\frac{c}{2}\right) B_t^{t+0.5i} + B_t^{t+n} = 1 \;\Rightarrow\quad c = \frac{2\left(1 - B_t^{t+n}\right)}{\sum_{i=1}^{2n} B_t^{t+0.5i}}.$$

It reflects the current market interest rate for an n-year bullet bond. The par yield is closely related to the so-called swap rate, which is a key concept in the swap markets.

10.3 Real interest rates and aggregate consumption

In order to study the link between interest rates and aggregate consumption, we assume the existence of a representative individual maximizing the expected time-additive utility $\mathrm{E}[\int_0^T e^{-\delta t} u(c_t)\, dt]$. As discussed in Chapter 7, a representative individual will exist in a complete market. The parameter δ is the subjective time preference rate with higher δ representing a more impatient individual. c_t is the consumption rate of the individual, which is then also the aggregate consumption level in the economy. In terms of the utility and time preference of the representative individual the state-price deflator is therefore characterized by

$$\zeta_t = e^{-\delta t} \frac{u'(c_t)}{u'(c_0)}.$$

Let us take a continuous-time framework and assume that $c = (c_t)$ follows a stochastic process of the form

$$dc_t = c_t \left[\mu_{ct}\, dt + \boldsymbol{\sigma}_{ct}^\top\, d\boldsymbol{z}_t\right],$$

where $\boldsymbol{z} = (\boldsymbol{z}_t)$ is a (possibly multi-dimensional) standard Brownian motion. Then we have shown in Chapter 8 that the equilibrium continuously compounded short-term interest rate is given by

$$r_t = \delta + \gamma(c_t)\mu_{ct} - \frac{1}{2}\eta(c_t)\|\boldsymbol{\sigma}_{ct}\|^2, \tag{10.13}$$

and that

$$\boldsymbol{\lambda}_t = \gamma(c_t)\boldsymbol{\sigma}_{ct} \tag{10.14}$$

defines a market price of risk process. Here $\gamma(c_t) \equiv -c_t u''(c_t)/u'(c_t)$ is the relative risk aversion and $\eta(c_t) \equiv c_t^2 u'''(c_t)/u'(c_t)$, which is positive under the very plausible assumption of decreasing absolute risk aversion. For notational simplicity we leave out the f superscript on the short-term interest rate in this chapter.

Equation (10.13) gives the interest rate at which the market for short-term borrowing and lending will clear. The equation relates the equilibrium short-term interest rate to the time preference rate and the expected growth rate μ_{ct} and the variance rate $\|\boldsymbol{\sigma}_{ct}\|^2$ of aggregate consumption growth over the next instant. We can observe the following relations:

- There is a positive relation between the time preference rate and the equilibrium interest rate. The intuition behind this is that when the individuals of the economy are impatient and has a high demand for current consumption, the equilibrium interest rate must be high in order to encourage the individuals to save now and postpone consumption.

- The multiplier of μ_{ct} in (10.13) is the relative risk aversion of the representative individual, which is positive. Hence, there is a positive relation between the expected growth in aggregate consumption and the equilibrium interest rate. This can be explained as follows: We expect higher future consumption and hence lower future marginal utility, so postponed payments due to saving have lower value. Consequently, a higher return on saving is needed to maintain market clearing.

- If u''' is positive, there will be a negative relation between the variance of aggregate consumption and the equilibrium interest rate. If the representative individual has decreasing absolute risk aversion, which is certainly a reasonable assumption, u''' has to be positive. The intuition is that the greater the uncertainty about future consumption, the more will the individuals appreciate the sure payments from the risk-free asset and hence the lower a return is necessary to clear the market for borrowing and lending.

In the special case of constant relative risk aversion, $u(c) = c^{1-\gamma}/(1-\gamma)$, Equation (10.13) simplifies to

$$r_t = \delta + \gamma\mu_{ct} - \frac{1}{2}\gamma(1+\gamma)\|\boldsymbol{\sigma}_{ct}\|^2. \tag{10.15}$$

In particular, we see that if the drift and variance rates of aggregate consumption are constant, i.e. aggregate consumption follows a geometric Brownian motion, then the short-term interest rate will be constant over time. In that case the time t price of the zero-coupon bond maturing at time s is

$$B_t^s = \mathrm{E}_t\left[\frac{\zeta_s}{\zeta_t}\right] = \mathrm{E}_t\left[\exp\left\{-r(s-t) - \frac{1}{2}\|\boldsymbol{\lambda}\|^2(s-t) - \boldsymbol{\lambda}^\top(\boldsymbol{z}_s - \boldsymbol{z}_t)\right\}\right] = e^{-r(s-t)}$$

and the corresponding continuous compounded yield is $y_t^s = r$. Consequently, the yield curve will be flat and constant over time. This is clearly an unrealistic case. To obtain interesting models we must either allow for variations in the expectation and the variance of aggregate consumption growth or allow for non-constant relative risk aversion (or both).

What can we say in general about the relation between the equilibrium yield curve and the expectations and uncertainty about future aggregate consumption?.[1] The equilibrium time t price

[1] The presentation is adapted from Breeden (1986).

of a zero-coupon bond paying one consumption unit at time $T \geq t$ is given by

$$B_t^T = \mathrm{E}_t\left[\frac{\zeta_T}{\zeta_t}\right] = e^{-\delta(T-t)}\frac{\mathrm{E}_t\left[u'(c_T)\right]}{u'(c_t)}, \tag{10.16}$$

where c_T is the uncertain future aggregate consumption level. We can write the left-hand side of the equation above in terms of the yield y_t^T of the bond as

$$B_t^T = e^{-y_t^T(T-t)} \approx 1 - y_t^T(T-t),$$

using a first order Taylor expansion. Turning to the right-hand side of the equation, we will use a second-order Taylor expansion of $u'(c_T)$ around c_t:

$$u'(c_T) \approx u'(c_t) + u''(c_t)(c_T - c_t) + \frac{1}{2}u'''(c_t)(c_T - c_t)^2.$$

This approximation is reasonable when c_T stays relatively close to c_t, which is the case for fairly low and smooth consumption growth and fairly short time horizons. Applying the approximation, the right-hand side of (10.16) becomes

$$\begin{aligned} e^{-\delta(T-t)}\frac{\mathrm{E}_t\left[u'(c_T)\right]}{u'(c_t)} &\approx e^{-\delta(T-t)}\left(1 + \frac{u''(c_t)}{u'(c_t)}\mathrm{E}_t[c_T - c_t] + \frac{1}{2}\frac{u'''(c_t)}{u'(c_t)}\mathrm{Var}_t[c_T - c_t]\right) \\ &\approx 1 - \delta(T-t) + e^{-\delta(T-t)}\frac{c_t u''(c_t)}{u'(c_t)}\mathrm{E}_t\left[\frac{c_T}{c_t} - 1\right] \\ &\quad + \frac{1}{2}e^{-\delta(T-t)}c_t^2\frac{u'''(c_t)}{u'(c_t)}\mathrm{Var}_t\left[\frac{c_T}{c_t}\right], \end{aligned}$$

where we have used the approximations $e^{-\delta(T-t)} \approx 1-\delta(T-t)$ and $(\mathrm{E}_t[c_T - c_t])^2 \approx 0$. Substituting the approximations of both sides into (10.16) and rearranging, we find the following approximate expression for the zero-coupon yield:

$$y_t^T \approx \delta + e^{-\delta(T-t)}\left(\frac{-c_t u''(c_t)}{u'(c_t)}\right)\frac{\mathrm{E}_t\left[c_T/c_t - 1\right]}{T-t} - \frac{1}{2}e^{-\delta(T-t)}c_t^2\frac{u'''(c_t)}{u'(c_t)}\frac{\mathrm{Var}_t\left[c_T/c_t\right]}{T-t}. \tag{10.17}$$

Again assuming $u' > 0$, $u'' < 0$, and $u''' > 0$, we can state the following conclusions. The equilibrium yield is increasing in the subjective rate of time preference. The equilibrium yield for the period $[t, T]$ is positively related to the expected growth rate of aggregate consumption over the period and negatively related to the uncertainty about the growth rate of consumption over the period. The intuition for these results is the same as for short-term interest rate discussed above. We see that the shape of the equilibrium time t yield curve $T \mapsto y_t^T$ is determined by how expectations and variances of consumption growth rates depend on the length of the forecast period. For example, if the economy is expected to enter a short period of high growth rates, real short-term interest rates tend to be high and the yield curve downward-sloping.

10.4 Real interest rates and aggregate production

In order to study the relation between interest rates and production, we will look at a slightly simplified version of the general equilibrium model of Cox, Ingersoll, and Ross (1985a).

Consider an economy with a single physical good that can be used either for consumption or investment. All values are expressed in units of this good. The instantaneous rate of return on an investment in the production of the good is

$$\frac{d\eta_t}{\eta_t} = g(X_t)\,dt + \xi(X_t)\,dz_{1t}, \tag{10.18}$$

where z_1 is a standard one-dimensional Brownian motion and g and ξ are well-behaved real-valued functions (given by Mother Nature) of some state variable X_t. We assume that $\xi(x)$ is non-negative for all values of X. The above dynamics means that η_0 goods invested in the production process at time 0 will grow to η_t goods at time t if the output of the production process is continuously reinvested in this period. We can interpret g as the expected real growth rate of production in the economy and the volatility ξ (assumed positive for all X) as a measure of the uncertainty about the growth rate of production in the economy. The production process has constant returns to scale in the sense that the distribution of the rate of return is independent of the scale of the investment. There is free entry to the production process. We can think of individuals investing in production directly by forming their own firm or indirectly by investing in stocks of production firms. For simplicity we take the first interpretation. All producers, individuals and firms, act competitively so that firms have zero profits and just passes production returns on to their owners. All individuals and firms act as price takers.

We assume that the state variable is a one-dimensional diffusion with dynamics

$$dX_t = m(X_t)\,dt + v_1(X_t)\,dz_{1t} + v_2(X_t)\,dz_{2t}, \tag{10.19}$$

where z_2 is another standard one-dimensional Brownian motion independent of z_1, and m, v_1, and v_2 are well-behaved real-valued functions. The instantaneous variance rate of the state variable is $v_1(x)^2 + v_2(x)^2$, the covariance rate of the state variable and the real growth rate is $\xi(x)v_1(x)$ so that the correlation between the state and the growth rate is $v_1(x)/\sqrt{v_1(x)^2 + v_2(x)^2}$. Unless $v_2 \equiv 0$, the state variable is imperfectly correlated with the real production returns. If v_1 is positive [negative], then the state variable is positively [negatively] correlated with the growth rate of production in the economy. Since the state determines the expected returns and the variance of returns on real investments, we may think of X_t as a productivity or technology variable.

In addition to the investment in the production process, we assume that the individuals have access to a financial asset with a price P_t with dynamics of the form

$$\frac{dP_t}{P_t} = \mu_t\,dt + \sigma_{1t}\,dz_{1t} + \sigma_{2t}\,dz_{2t}. \tag{10.20}$$

As a part of the equilibrium we will determine the relation between the expected return μ_t and the sensitivity coefficients σ_{1t} and σ_{2t}. Finally, the individuals can borrow and lend funds at an instantaneously risk-free interest rate r_t, which is also determined in equilibrium. The market is therefore complete. Other financial assets affected by z_1 and z_2 may be traded, but they will be redundant. We will get the same equilibrium relation between expected returns and sensitivity coefficients for these other assets as for the one modeled explicitly. For simplicity we stick to the case with a single financial asset.

If an individual at each time t consumes at a rate of $c_t \geq 0$, invests a fraction α_t of his wealth in the production process, invests a fraction π_t of wealth in the financial asset, and invests the remaining fraction $1 - \alpha_t - \pi_t$ of wealth in the risk-free asset, his wealth W_t will evolve as

$$\begin{aligned} dW_t = {} & \{r_t W_t + W_t\alpha_t\,(g(X_t) - r_t) + W_t\pi_t\,(\mu_t - r_t) - c_t\}\,dt \\ & + W_t\alpha_t\xi(X_t)\,dz_{1t} + W_t\pi_t\sigma_{1t}\,dz_{1t} + W_t\pi_t\sigma_{2t}\,dz_{2t}. \end{aligned} \tag{10.21}$$

Since a negative real investment is physically impossible, we should restrict α_t to the non-negative numbers. However, we will assume that this constraint is not binding.

Let us look at an individual maximizing expected utility of future consumption. The indirect utility function is defined as

$$J(W,x,t) = \sup_{(\alpha_s,\pi_s,c_s)_{s\in[t,T]}} \mathrm{E}_t\left[\int_t^T e^{-\delta(s-t)}u(c_s)\,ds\right],$$

i.e. the maximal expected utility the individual can obtain given his current wealth and the current value of the state variable. The dynamic programming technique of Section 6.5.2 lead to the Hamilton-Jacobi-Bellman equation

$$\begin{aligned}\delta J = \sup_{\alpha,\pi,c}\Big\{ & u(c) + \frac{\partial J}{\partial t} + J_W\left(rW + \alpha W(g-r) + \pi W(\mu - r) - c\right) \\ & + \frac{1}{2}J_{WW}W^2\left([\alpha\xi + \pi\sigma_1]^2 + \pi^2\sigma_2^2\right) + J_x m \\ & + \frac{1}{2}J_{xx}(v_1^2+v_2^2) + J_{Wx}Wv_1(\alpha\xi + \pi\sigma_1)\Big\}\end{aligned}$$

The first-order conditions for α and π imply that

$$\alpha^* = \frac{-J_W}{WJ_{WW}}\left[(g-r)\frac{\sigma_1^2+\sigma_2^2}{\xi^2\sigma_2^2} - (\mu-r)\frac{\sigma_1}{\xi\sigma_2^2}\right] + \frac{-J_{Wx}}{WJ_{WW}}\frac{\sigma_2 v_1 - \sigma_1 v_2}{\xi\sigma_2}, \tag{10.22}$$

$$\pi^* = \frac{-J_W}{WJ_{WW}}\left[-\frac{\sigma_1}{\xi\sigma_2^2}(g-r) + \frac{1}{\sigma_2^2}(\mu-r)\right] + \frac{-J_{Wx}}{WJ_{WW}}\frac{v_2}{\sigma_2}. \tag{10.23}$$

In equilibrium, prices and interest rates are such that (a) all individuals act optimally and (b) all markets clear. In particular, summing up the positions of all individuals in the financial asset we should get zero, and the total amount borrowed by individuals on a short-term basis should equal the total amount lend by individuals. Since the available production apparatus is to be held by some investors, summing the optimal α's over investors we should get 1. Since we have assumed a complete market, we can construct a representative individual, i.e. an individual with a given utility function so that the equilibrium interest rates and price processes are the same in the single individual economy as in the larger multi-individual economy. (Alternatively, we may think of the case where all individuals in the economy are identical so that they will have the same indirect utility function and always make the same consumption and investment choice.)

In an equilibrium, we have $\pi^* = 0$ for a representative individual, and hence (10.23) implies that

$$\mu - r = \frac{\sigma_1}{\xi}(g-r) - \frac{\left(\frac{-J_{Wx}}{WJ_{WW}}\right)}{\left(\frac{-J_W}{WJ_{WW}}\right)}\sigma_2 v_2. \tag{10.24}$$

Substituting this into the expression for α^* and using the fact that $\alpha^* = 1$ in equilibrium, we get that

$$\begin{aligned}1 &= \left(\frac{-J_W}{WJ_{WW}}\right)\left[(g-r)\frac{\sigma_1^2+\sigma_2^2}{\xi^2\sigma_2^2} - \frac{\sigma_1}{\xi}\frac{\sigma_1}{\xi\sigma_2^2}(g-r) + \frac{\left(\frac{-J_{Wx}}{WJ_{WW}}\right)}{\left(\frac{-J_W}{WJ_{WW}}\right)}\sigma_2 v_2\frac{\sigma_1}{\xi\sigma_2^2}\right] \\ &\quad + \left(\frac{-J_{Wx}}{WJ_{WW}}\right)\frac{\sigma_2 v_1 - \sigma_1 v_2}{\xi\sigma_2} \\ &= \left(\frac{-J_W}{WJ_{WW}}\right)\frac{g-r}{\xi^2} + \left(\frac{-J_{Wx}}{WJ_{WW}}\right)\frac{v_1}{\xi}.\end{aligned}$$

Consequently, the equilibrium short-term interest rate can be written as

$$r = g - \left(\frac{-WJ_{WW}}{J_W}\right)\xi^2 + \frac{J_{Wx}}{J_W}\xi v_1. \tag{10.25}$$

This equation ties the equilibrium real short-term interest rate to the production side of the economy. Let us address each of the three right-hand side terms:

- The equilibrium real interest rate r is positively related to the expected real growth rate g of the economy. The intuition is that for higher expected growth rates, the productive investments are more attractive relative to the risk-free investment, so to maintain market clearing the interest rate has to be higher as well.

- The term $-WJ_{WW}/J_W$ is the relative risk aversion of the representative individual's indirect utility. This is assumed to be positive. Hence, we see that the equilibrium real interest rate r is negatively related to the uncertainty about the growth rate of the economy, represented by the instantaneous variance ξ^2. For a higher uncertainty, the safe returns of a risk-free investment is relatively more attractive, so to establish market clearing the interest rate has to decrease.

- The last term in (10.25) is due to the presence of the state variable. The covariance rate of the state variable and the real growth rate of the economy is equal to ξv_1. Suppose that high values of the state variable represent good states of the economy, where the wealth of the individual is high. Then the marginal utility J_W will be decreasing in X, i.e. $J_{Wx} < 0$. If instantaneous changes in the state variable and the growth rate of the economy are positively correlated, we see from (10.22) that the hedge demand of the productive investment is decreasing, and hence the demand for depositing money at the short rate increasing, in the magnitude of the correlation (both J_{Wx} and J_{WW} are negative). To maintain market clearing, the interest rate must be decreasing in the magnitude of the correlation as reflected by (10.25).

We see from (10.24) that the market prices of risk are given by

$$\lambda_1 = \frac{g-r}{\xi}, \quad \lambda_2 = -\frac{\left(\frac{-J_{Wx}}{WJ_{WW}}\right)}{\left(\frac{-J_W}{WJ_{WW}}\right)} v_2 = -\frac{J_{Wx}}{J_W} v_2. \tag{10.26}$$

Applying the relation

$$g - r = \left(\frac{-WJ_{WW}}{J_W}\right)\xi^2 - \frac{J_{Wx}}{J_W}\xi v_1,$$

we can rewrite λ_1 as

$$\lambda_1 = \left(\frac{-WJ_{WW}}{J_W}\right)\xi - \frac{J_{Wx}}{J_W} v_1. \tag{10.27}$$

10.5 Equilibrium interest rate models

10.5.1 The Vasicek model

A classic but still widely used model of interest rate dynamics and the pricing of bonds and interest rate derivatives is the model proposed by Vasicek (1977). The basic assumptions of the model is that the continuously compounded short-term interest rate r_t has dynamics

$$dr_t = \kappa\left(\bar{r} - r_t\right)\, dt + \sigma_r\, dz_t, \tag{10.28}$$

where κ, $\bar{r}$, and σ_r are positive constants, and that the market price of risk associated with the shock z is a constant λ.

Before we study the consequences of these assumptions, let us see how they can be supported by a consumption-based equilibrium model. Following Goldstein and Zapatero (1996) assume that aggregate consumption evolves as

$$dc_t = c_t \left[\mu_{ct}\, dt + \sigma_c\, dz_t\right],$$

where z is a one-dimensional standard Brownian motion, σ_C is a constant, and the expected consumption growth rate μ_{ct} follows the process

$$d\mu_{ct} = \kappa\left(\bar{\mu}_c - \mu_{ct}\right) dt + \theta\, dz_t.$$

The representative individual is assumed to have a constant relative risk aversion of γ. It follows from (10.15) that the equilibrium real short-term interest rate is

$$r_t = \delta + \gamma\mu_{ct} - \frac{1}{2}\gamma(1+\gamma)\sigma_C^2$$

with dynamics $dr_t = \gamma d\mu_{ct}$, which gives (10.28) with $\sigma_r = \gamma\theta$ and $\bar{r} = \gamma\bar{\mu}_c + \delta - \frac{1}{2}\gamma(1+\gamma)\sigma_c^2$. The market price of risk is $\lambda = \gamma\sigma_c$, a constant.

The process (10.28) is a so-called Ornstein-Uhlenbeck process. An Ornstein-Uhlenbeck process exhibits *mean reversion* in the sense that the drift is positive when $r_t < \bar{r}$ and negative when $x_t > \bar{r}$. The process is therefore always pulled towards a long-term level of $\bar{r}$. However, the random shock to the process through the term $\sigma_r\, dz_t$ may cause the process to move further away from $\bar{r}$. The parameter κ controls the size of the expected adjustment towards the long-term level and is often referred to as the mean reversion parameter or the speed of adjustment.

To determine the distribution of the future value of the short-term interest rate define a new process y_t as some function of r_t such that $y = (y_t)_{t\geq 0}$ is a generalized Brownian motion. It turns out that this is satisfied for $y_t = g(r_t, t)$, where $g(r,t) = e^{\kappa t}r$. From Itô's Lemma we get

$$\begin{aligned} dy_t &= \left[\frac{\partial g}{\partial t}(r_t,t) + \frac{\partial g}{\partial r}(r_t,t)\kappa\left(\bar{r} - r_t\right) + \frac{1}{2}\frac{\partial^2 g}{\partial r^2}(r_t,t)\sigma_r^2\right] dt + \frac{\partial g}{\partial r}(r_t,t)\sigma_r\, dz_t \\ &= \left[\kappa e^{\kappa t}r_t + \kappa e^{\kappa t}\left(\bar{r} - r_t\right)\right] dt + e^{\kappa t}\sigma_r\, dz_t \\ &= \kappa\bar{r}e^{\kappa t}\, dt + \sigma_r e^{\kappa t}\, dz_t. \end{aligned}$$

This implies that

$$y_{t'} = y_t + \kappa\bar{r}\int_t^{t'} e^{\kappa u}\, du + \int_t^{t'} \sigma_r e^{\kappa u}\, dz_u.$$

After substitution of the definition of y_t and $y_{t'}$ and a multiplication by $e^{-\kappa t'}$, we arrive at the expression

$$\begin{aligned} r_{t'} &= e^{-\kappa(t'-t)}r_t + \kappa\bar{r}\int_t^{t'} e^{-\kappa(t'-u)}\, du + \int_t^{t'} \sigma_r e^{-\kappa(t'-u)}\, dz_u \\ &= e^{-\kappa(t'-t)}r_t + \bar{r}\left(1 - e^{-\kappa(t'-t)}\right) + \int_t^{t'} \sigma_r e^{-\kappa(t'-u)}\, dz_u. \end{aligned} \tag{10.29}$$

This holds for all $t' > t \geq 0$. In particular, we get that the solution to the stochastic differential equation (10.28) can be written as

$$r_t = e^{-\kappa t}r_0 + \bar{r}\left(1 - e^{-\kappa t}\right) + \int_0^t \sigma_r e^{-\kappa(t-u)}\, dz_u. \tag{10.30}$$

According to Theorem 2.3, the integral $\int_t^{t'} \sigma_r e^{-\kappa(t'-u)}\,dz_u$ is normally distributed with mean zero and variance $\int_t^{t'} \sigma_r^2 e^{-2\kappa(t'-u)}\,du = \frac{\sigma_r^2}{2\kappa}\left(1-e^{-2\kappa(t'-t)}\right)$. We can thus conclude that $r_{t'}$ (given r_t) is normally distributed, with mean and variance given by

$$\mathrm{E}_t[r_{t'}] = e^{-\kappa(t'-t)} r_t + \bar{r}\left(1-e^{-\kappa(t'-t)}\right), \tag{10.31}$$

$$\mathrm{Var}_t[r_{t'}] = \frac{\sigma_r^2}{2\kappa}\left(1-e^{-2\kappa(t'-t)}\right). \tag{10.32}$$

The value space of an Ornstein-Uhlenbeck process is $\mathbb{R}$. For $t' \to \infty$, the mean approaches $\bar{r}$, and the variance approaches $\sigma_r^2/(2\kappa)$. For $\kappa \to \infty$, the mean approaches $\bar{r}$, and the variance approaches 0. For $\kappa \to 0$, the mean approaches the current value r_t, and the variance approaches $\sigma_r^2(t'-t)$. The distance between the level of the process and the long-term level is expected to be halved over a period of $t'-t = (\ln 2)/\kappa$, since $\mathrm{E}_t[r_{t'}] - \bar{r} = \frac{1}{2}(r_t - \bar{r})$ implies that $e^{-\kappa(t'-t)} = \frac{1}{2}$ and, hence, $t'-t = (\ln 2)/\kappa$.

The effect of the different parameters can also be evaluated by looking at the paths of the process, which can be simulated by

$$r_{t_i} = r_{t_{i-1}} + \kappa[\bar{r} - r_{t_{i-1}}](t_i - t_{i-1}) + \sigma_r \varepsilon_i \sqrt{t_i - t_{i-1}},$$

where $\varepsilon_i \sim N(0,1)$. Figure 10.1 shows a single path for different combinations of r_0, κ, $\bar{r}$, and σ_r. In each sub-figure one of the parameters is varied and the others fixed. The base values of the parameters are $r_0 = 0.08$, $\bar{r} = 0.08$, $\kappa = \ln 2 \approx 0.69$, and $\sigma_r = 0.03$. All paths are computed using the same sequence of random numbers $\varepsilon_1, \ldots, \varepsilon_n$ and are therefore directly comparable. None of the paths shown involve negative values of the process, but other paths will, see e.g. Figure 10.2. As a matter of fact, it can be shown that an Ornstein-Uhlenbeck process with probability one will sooner or later become negative.

What are the implications of the Vasicek assumptions for bond prices and the yield curve? The time t price of a zero-coupon bond maturing at time s is given by

$$B_t^s = \mathrm{E}_t\left[\frac{\zeta_s}{\zeta_t}\right] = \mathrm{E}_t\left[\exp\left\{-\int_t^s r_u\,du - \frac{1}{2}\int_t^s \lambda^2\,du - \int_t^s \lambda\,dz_u\right\}\right].$$

In order to compute this expectation, first use (10.29) to find that

$$\int_t^s r_u\,du = \int_t^s e^{-\kappa(u-t)} r_t\,du + \int_t^s \bar{r}\left(1-e^{-\kappa(u-t)}\right)du + \int_t^s \int_t^u \sigma_r e^{-\kappa(u-v)}\,dz_v\,du.$$

Interchange the order of integration in the double integral (this follows from the so-called Fubini Theorem of stochastic calculus)

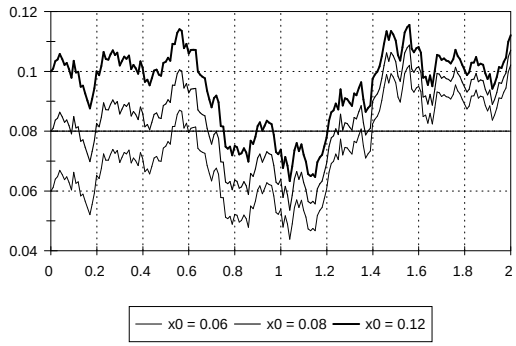
$$\int_t^s \left[\int_t^u \sigma_r e^{-\kappa(u-v)}\,dz_v\right]du = \int_t^s \left[\int_v^s \sigma_r e^{-\kappa(u-v)}\,du\right]dz_v.$$

Further note that $\int_t^s e^{-\kappa(u-t)}\,du = b(s-t)$, where we have introduced the function

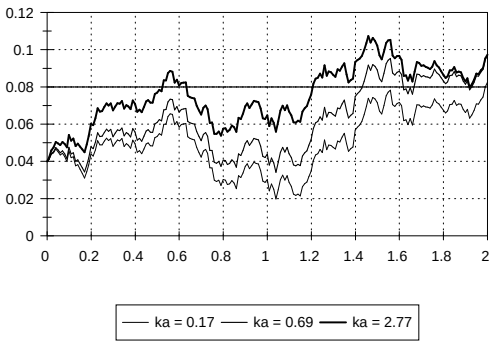
$$b(\tau) = \frac{1}{\kappa}\left(1-e^{-\kappa\tau}\right), \tag{10.33}$$

Also note that

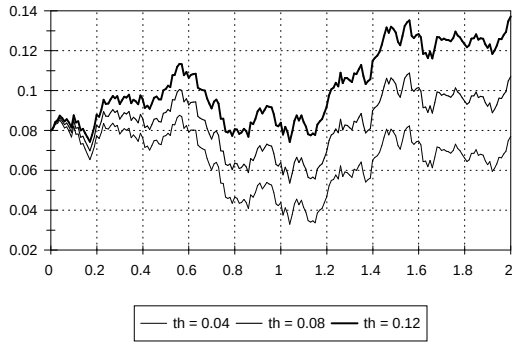
$$\int_t^s b(s-u)\,du = \frac{1}{\kappa}(s-t-b(s-t)), \qquad \int_t^s b(s-u)^2\,du = \frac{1}{\kappa^2}(s-t-b(s-t)) - \frac{1}{2\kappa}b(s-t)^2.$$



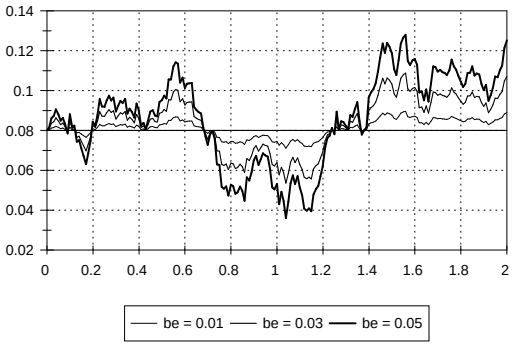
(a) Different initial values r_0



(b) Different κ-values; $r_0 = 0.04$



(c) Different $\bar{r}$-values



(d) Different σ_r-values

Figure 10.1: Simulated paths for an Ornstein-Uhlenbeck process. The basic parameter values are $r_0 = \bar{r} = 0.08$, $\kappa = \ln 2 \approx 0.69$, and $\sigma_r = 0.03$.

It now follows that

$$\int_t^s r_u \, du = r_t b(s-t) + \bar{r}\,(s - t - b(s-t)) + \int_t^s \sigma_r b(s-u) \, dz_u,$$

and using Theorem 2.3 and results in Appendix B we obtain

$$\begin{aligned} B_t^s &= e^{-r_t b(s-t) - \bar{r}(s-t-b(s-t)) - \frac{1}{2}\lambda^2 (s-t)} \, \mathrm{E}_t \left[e^{-\int_t^s (\lambda + \sigma_r b(s-u)) \, dz_u} \right] \\ &= e^{-r_t b(s-t) - \bar{r}(s-t-b(s-t)) - \frac{1}{2}\lambda^2 (s-t)} e^{\frac{1}{2}\int_t^s (\lambda + \sigma_r b(s-u))^2 \, du} \\ &= e^{-r_t b(s-t) - \bar{r}(s-t-b(s-t)) - \frac{1}{2}\lambda^2 (s-t)} e^{\frac{1}{2}\lambda^2 (s-t) + \lambda\sigma_r \int_t^s b(s-u) \, du + \frac{1}{2}\sigma_r^2 \int_t^s b(s-u)^2 \, du} \\ &= e^{-a(s-t) - b(s-t) r_t}, \end{aligned} \tag{10.34}$$

where

$$a(\tau) = y_\infty (\tau - b(\tau)) + \frac{\sigma_r^2}{4\kappa} b(\tau)^2 \tag{10.35}$$

and

$$y_\infty = \bar{r} - \frac{\lambda\sigma_r}{\kappa} - \frac{\sigma_r^2}{2\kappa^2}.$$

The continuously compounded yield of the zero-coupon bond maturing at time s is

$$y_t^s = -\frac{\ln B_t^s}{s-t} = \frac{a(s-t)}{s-t} + \frac{b(s-t)}{s-t} r_t, \tag{10.36}$$

which is affine in the current short rate r_t. A model with this property is called an affine term structure model. It can be shown that $y_t^s \to y_\infty$ for $s \to \infty$, which explains the notation. The asymptotic long yield is a constant in the Vasicek model. Concerning the shape of the yield curve $s \to y_t^s$ it can be shown that

(i) if $r_t < y_\infty - \frac{\sigma_r^2}{4\kappa^2}$, the yield curve is increasing;

(ii) if $r_t > y_\infty + \frac{\sigma_r^2}{2\kappa^2}$, the yield curve is decreasing;

(iii) for intermediate values of r_t, the yield curve is humped, i.e. increasing in s up to some maturity s^* and then decreasing for longer maturities.

Within the Vasicek model it is possible to find closed-form expressions for the prices of many other interesting assets, such as forwards and futures on bonds, Eurodollar futures, and European options on bonds; see Chapter 12.

There are many other affine term structure models and they can basically all be supported by a consumption-based asset pricing model in the same way as the Vasicek model. Assume that the expected growth rate and the variance rate of aggregate consumption are affine in some state variables, i.e.

$$\mu_{ct} = a_0 + \sum_{i=1}^n a_i X_{it}, \qquad \|\boldsymbol{\sigma}_{ct}\|^2 = b_0 + \sum_{i=1}^n b_i X_{it},$$

then the equilibrium short rate will be

$$r_t = \left(\delta + \gamma a_0 - \frac{1}{2}\gamma(1+\gamma)b_0\right) + \gamma \sum_{i=1}^n \left(a_i - \frac{1}{2}(1+\gamma)b_i\right) X_{it}.$$

Of course, we should have $b_0 + \sum_{i=1}^n b_i X_{it} \geq 0$ for all values of the state variables. The market price of risk is $\lambda_t = \gamma \sigma_{ct}$. If the state variables X_i follow processes of the affine type, we have an affine term structure model.

For other term structure models developed with the consumption-based approach, see e.g. Bakshi and Chen (1997).

10.5.2 The Cox-Ingersoll-Ross model

Another widely used model of interest rate dynamics was suggested by Cox, Ingersoll, and Ross (1985b). They assume that the short-term interest rate r_t follows a so-called square-root process

$$dr_t = \kappa\left(\bar{r} - r_t\right) dt + \sigma_r \sqrt{r_t}\, d\bar{z}_t, \tag{10.37}$$

where κ, $\bar{r}$, and σ_r are positive constants. Further they assume that the associated market price of risk is $\lambda_t = \lambda\sqrt{r_t}/\sigma_r$, where the λ on the right-hand side is a constant.

They derive their model as a special case of their general equilibrium model with production which we have reviewed in Section 10.4. The representative individual is assumed to have a logarithmic utility so that the relative risk aversion of the direct utility function is 1. In addition,

the individual is assumed to have an infinite time horizon, which implies that the indirect utility function will be independent of time. It can be shown that under these assumptions the indirect utility function of the individual is of the form $J(W,x) = A\ln W + B(x)$. In particular, $J_{Wx} = 0$ and the relative risk aversion of the indirect utility function is also 1. It follows from (10.25) that the equilibrium real short-term interest rate is equal to

$$r(x_t) = g(x_t) - \xi(x_t)^2.$$

The authors further assume that the expected rate of return and the variance rate of the return on the productive investment are both proportional to the state, i.e.

$$g(x) = k_1 x, \qquad \xi(x)^2 = k_2 x,$$

where $k_1 > k_2$. Then the equilibrium short-rate becomes $r(x) = (k_1 - k_2)x \equiv kx$. Assume now that the state variable follows a square-root process

$$\begin{aligned} dx_t &= \kappa\,(\bar{x} - x_t)\;dt + \rho\sigma_x\sqrt{x_t}\,dz_{1t} + \sqrt{1-\rho^2}\sigma_x\sqrt{x_t}\,dz_{2t} \\ &= \kappa\,(\bar{x} - x_t)\;dt + \sigma_x\sqrt{x_t}\,d\bar{z}_t, \end{aligned}$$

where $\bar{z}$ is a standard Brownian motion with correlation ρ with the standard Brownian motion z_1 and correlation $\sqrt{1-\rho^2}$ with z_2. Then the dynamics of the real short rate is $dr_t = k\,dx_t$, which yields

$$dr_t = \kappa\,(\bar{r} - r_t)\;dt + \sigma_r\sqrt{r_t}\,d\bar{z}_t, \tag{10.38}$$

where $\bar{r} = k\bar{x}$ and $\sigma_r = \sqrt{k}\sigma_x$. The market prices of risk given in (10.26) and (10.27) simplify to

$$\lambda_1 = \xi(x) = \sqrt{k_2 x} = \sqrt{k_2/k}\,\sqrt{r}, \qquad \lambda_2 = 0.$$

The market price of risk associated with the combined shock $\bar{z}$ is $\rho\lambda_1 + \sqrt{1-\rho^2}\lambda_2$, which is proportional to $\sqrt{r}$. These conclusions support (10.37) and the associated form of the market price of risk.

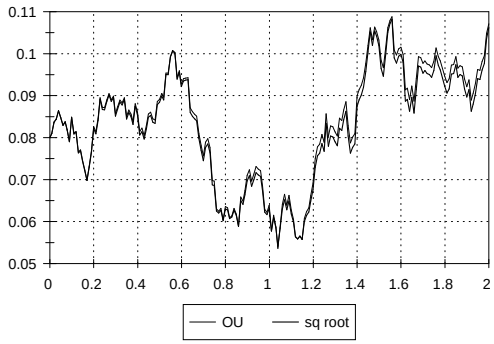
Before we discuss the implications for bond prices and the yield curve, let us look at the properties of the square-root process. The only difference to the Ornstein-Uhlenbeck process is the square-root term in the volatility. The variance rate is now $\sigma_r^2 r_t$ which is proportional to the level of the process. A square root process also exhibits mean reversion. A square root process can only take on non-negative values. To see this, note that if the value should become zero, then the drift is positive and the volatility zero, and therefore the value of the process will with certainty become positive immediately after (zero is a so-called reflecting barrier). It can be shown that if $2\kappa\bar{r} \geq \sigma_r^2$, the positive drift at low values of the process is so big relative to the volatility that the process cannot even reach zero, but stays strictly positive.[2] Hence, the value space for a square root process is either $\mathcal{S} = [0,\infty)$ or $\mathcal{S} = (0,\infty)$.

Paths for the square root process can be simulated by successively calculating

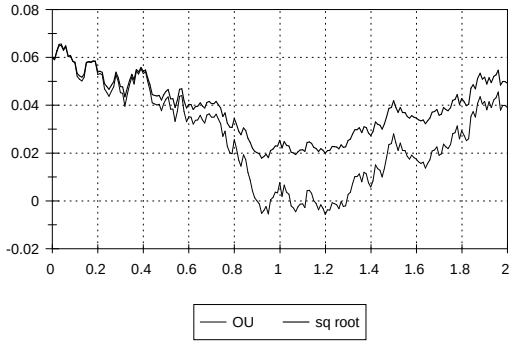
$$r_{t_i} = r_{t_{i-1}} + \kappa[\bar{r} - r_{t_{i-1}}](t_i - t_{i-1}) + \sigma_r\sqrt{r_{t_{i-1}}}\varepsilon_i\sqrt{t_i - t_{i-1}}.$$

Variations in the different parameters will have similar effects as for the Ornstein-Uhlenbeck process, which is illustrated in Figure 10.1. Instead, let us compare the paths for a square root process

[2] To show this, the results of Karlin and Taylor (1981, p. 226ff) can be applied.



(a) Initial value $r_0 = 0.08$, same random numbers as in Figure 10.1



(b) Initial value $r_0 = 0.06$, different random numbers

Figure 10.2: A comparison of simulated paths for an Ornstein-Uhlenbeck process and a square root process. For both processes, the parameters $\bar{r} = 0.08$ and $\kappa = \ln 2 \approx 0.69$ are used, while σ_r is set to 0.03 for the Ornstein-Uhlenbeck process and to $0.03/\sqrt{0.08} \approx 0.1061$ for the square root process.

and an Ornstein-Uhlenbeck process using the same drift parameters κ and $\bar{r}$, but where the σ_r-parameter for the Ornstein-Uhlenbeck process is set equal to the σ_r-parameter for the square root process multiplied by the square root of $\bar{r}$, which ensures that the processes will have the same variance rate at the long-term level. Figure 10.2 compares two pairs of paths of the processes. In part (a), the initial value is set equal to the long-term level, and the two paths continue to be very close to each other. In part (b), the initial value is lower than the long-term level, so that the variance rates of the two processes differ from the beginning. For the given sequence of random numbers, the Ornstein-Uhlenbeck process becomes negative, while the square root process of course stays positive. In this case there is a clear difference between the paths of the two processes.

Since a square root process cannot become negative, the future values of the process cannot be normally distributed. In order to find the actual distribution, let us try the same trick as for the Ornstein-Uhlenbeck process, that is we look at $y_t = e^{\kappa t} r_t$. By Itô's Lemma,

$$\begin{aligned} dy_t &= \kappa e^{\kappa t} r_t \, dt + \kappa e^{\kappa t} (\bar{r} - \kappa r_t) \, dt + e^{\kappa t} \sigma_r \sqrt{r_t} \, dz_t \\ &= \kappa \bar{r} e^{\kappa t} \, dt + \sigma_r e^{\kappa t} \sqrt{r_t} \, dz_t, \end{aligned}$$

so that

$$y_{t'} = y_t + \kappa \bar{r} \int_t^{t'} e^{\kappa u} \, du + \int_t^{t'} \sigma_r e^{\kappa u} \sqrt{r_u} \, dz_u.$$

Computing the ordinary integral and substituting the definition of y, we get

$$r_{t'} = r_t e^{-\kappa(t'-t)} + \bar{r} \left(1 - e^{-\kappa(t'-t)}\right) + \sigma_r \int_t^{t'} e^{-\kappa(t'-u)} \sqrt{r_u} \, dz_u. \tag{10.39}$$

Since r enters the stochastic integral we cannot immediately determine the distribution of $r_{t'}$ given r_t from this equation. We can, however, use it to obtain the mean and variance of $r_{t'}$. Due to the fact that the stochastic integral has mean zero, cf. Theorem 2.2, we easily get

$$\mathrm{E}_t[r_{t'}] = e^{-\kappa(t'-t)} r_t + \bar{r} \left(1 - e^{-\kappa(t'-t)}\right) = \bar{r} + (r_t - \bar{r}) \, e^{-\kappa(t'-t)}. \tag{10.40}$$

To compute the variance the second equation of Theorem 2.2 can be applied, which will eventually lead to

$$\mathrm{Var}_t[r_{t'}] = \frac{\sigma_r^2 r_t}{\kappa}\left(e^{-\kappa(t'-t)} - e^{-2\kappa(t'-t)}\right) + \frac{\sigma_r^2 \bar{r}}{2\kappa}\left(1 - e^{-\kappa(t'-t)}\right)^2. \tag{10.41}$$

Note that the mean is identical to the mean for an Ornstein-Uhlenbeck process, whereas the variance is more complicated for the square root process.

It can be shown that, given the value r_t, the value $r_{t'}$ with $t' > t$ is non-centrally χ^2-distributed. More precisely, the probability density function for $r_{t'}$ is

$$f_{r_{t'}|r_t}(x) = f_{\chi^2_{a,b}}(2cx),$$

where

$$\begin{aligned} c &= \frac{2\kappa}{\sigma_r^2\left(1 - e^{-\kappa(t'-t)}\right)}, \\ b &= c r_t e^{-\kappa(t'-t)}, \\ a &= \frac{4\kappa\bar{r}}{\sigma_r^2}, \end{aligned}$$

and where $f_{\chi^2_{a,b}}(\cdot)$ denotes the probability density function for a non-centrally χ^2-distributed random variable with a degrees of freedom and non-centrality parameter b.

It can be shown that the price of a zero-coupon bond maturing at time s is given by

$$B_t^s = e^{-a(s-t)-b(s-t)r_t}, \tag{10.42}$$

where

$$b(\tau) = \frac{2(e^{\nu\tau} - 1)}{(\nu + \hat{\kappa})(e^{\nu\tau} - 1) + 2\nu}, \tag{10.43}$$

$$a(\tau) = -\frac{2\kappa\bar{r}}{\sigma_r^2}\left(\ln(2\nu) + \frac{1}{2}(\hat{\kappa} + \nu)\tau - \ln\left[(\nu + \hat{\kappa})(e^{\nu\tau} - 1) + 2\nu\right]\right), \tag{10.44}$$

and $\hat{\kappa} = \kappa + \lambda$ and $\nu = \sqrt{\hat{\kappa}^2 + 2\sigma_r^2}$. As in the Vasicek model, the yields are affine in the current short rate,

$$y_t^s = \frac{a(s-t)}{s-t} + \frac{b(s-t)}{s-t} r_t.$$

It can be shown that the asymptotic long yield is

$$y_\infty \equiv \lim_{s\to\infty} y_t^s = \frac{2\kappa\bar{r}}{\hat{\kappa} + \nu},$$

and [see Kan (1992)] that the yield curve can have the following shapes:

(i) if $\kappa + \lambda > 0$, the yield curve is decreasing for $r_t \geq \kappa\bar{r}/(\kappa + \lambda)$ and increasing for $0 \leq r_t \leq \kappa\bar{r}/\nu$. For $\kappa\bar{r}/\nu < r_t < \kappa\bar{r}/(\kappa + \lambda)$, the yield curve is humped, i.e. first increasing, then decreasing.

(ii) if $\kappa + \lambda \leq 0$, the yield curve is increasing for $0 \leq r_t \leq \kappa\bar{r}/\nu$ and humped for $r_t > \kappa\bar{r}/\nu$.

Also in this model closed-form expressions can be derived for many popular interest rate related assets; see Chapter 12.

Longstaff and Schwartz (1992) study a two-factor version of the model. They assume that the production returns are given by

$$\frac{d\eta_t}{\eta_t} = g(X_{1t}, X_{2t})\, dt + \xi(X_{2t})\, dz_{1t},$$

where

$$g(X_1, X_2) = k_1 X_1 + k_2 X_2, \qquad \xi(X_2)^2 = k_3 X_2,$$

so that the state variable X_2 affects both expected returns and uncertainty of production, while the state variable X_1 only affects the expected return. With log utility the short rate is again equal to the expected return minus the variance,

$$r(X_1, X_2) = g(X_1, X_2) - \xi(X_2)^2 = k_1 X_1 + (k_2 - k_3) X_2.$$

The state variables are assumed to follow independent square-root processes,

$$\begin{aligned} dX_{1t} &= (\varphi_1 - \kappa_1 X_{1t})\ dt + \beta_1 \sqrt{X_{1t}}\, dz_{2t}, \\ dX_{2t} &= (\varphi_2 - \kappa_2 X_{2t})\ dt + \beta_2 \sqrt{X_{2t}}\, dz_{3t}, \end{aligned}$$

where z_2 are independent of z_1 and z_3, but z_1 and z_3 may be correlated. The market prices of risk associated with the Brownian motions are

$$\lambda_1(X_2) = \xi(X_2) = \sqrt{k_2}\sqrt{X_2}, \quad \lambda_2 = \lambda_3 = 0.$$

10.6 Real and nominal interest rates and term structures

In the following we shall first derive some generally valid relations between real rates, nominal rates, and inflation and investigate the differences between real and nominal bonds. Then we will discuss two different types of models in which we can say more about real and nominal rates. The first setting follows the neoclassical tradition in assuming that monetary holdings do not affect the preferences of the individuals so that the presence of money has no effects on real rates and real asset returns. Hence, the relations derived earlier in this chapter still applies. However, several empirical findings indicate that the existence of money does have real effects. For example, real stock returns are negatively correlated with inflation and positively correlated with money growth. Also, assets that are positively correlated with inflation have a lower expected return.[3] In the second setting we consider below, money is allowed to have real effects. Economies with this property are called *monetary economies*.

10.6.1 Real and nominal asset pricing

The relations between interest rates or yields and aggregate consumption and production obtained above apply to real interest rates or yields. The real short-term interest rate is the rate of return over the next instant of an asset that is risk-free in real terms, i.e. provides a certain purchasing power. A real yield of maturity s is derived from the real price of a real zero-coupon bond maturing at time s, i.e. a bond paying one consumption unit at time s. In reality most deposit arrangements and traded bonds are nominal in the sense that the dividends they promise are pre-specified units of some currency, not some units of consumption goods. In this section we search for the link between real and nominal interest rates and yields.

[3] Such results are reported by, e.g., Fama (1981), Fama and Gibbons (1982), Chen, Roll, and Ross (1986), and Marshall (1992).

Recall the results we derived on real and nominal pricing in Section 4.4. Let us stick to the continuous-time framework. Let F_t denote the price of the good in currency units at time t (or think of F_t as the value of the Consumer Price Index at time t.) A nominal dividend of $\tilde{D}_t$ corresponds to a real dividend of $D_t = \tilde{D}_t/F_t$ and a nominal price of $\tilde{P}_t$ corresponds to a real price of $P_t = \tilde{P}_t/F_t$. We have seen that a nominal state-price deflator $\tilde{\zeta} = (\tilde{\zeta}_t)$ is related to a real state-price deflator $\zeta = (\zeta_t)$ via the equation

$$\tilde{\zeta}_t = \frac{\zeta_t}{F_t}, \quad \text{all } t \in [0, T]. \tag{10.45}$$

The nominal state-price deflator links nominal dividends to nominal prices in the same way that a real state-price deflator links real dividends to real prices.

The real return on a nominally risk-free asset is generally stochastic (and conversely). The dynamics of the real state-price deflator is

$$d\zeta_t = -\zeta_t \left[r_t \, dt + \boldsymbol{\lambda}_t^\top \, d\boldsymbol{z}_t \right],$$

where $r = (r_t)$ is the short-term real interest rate and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_t)$ is the market price of risk. The dynamics of the price of the consumption good is written as

$$dF_t = F_t \left[\mu_{\varphi t} \, dt + \boldsymbol{\sigma}_{\varphi t}^\top \, d\boldsymbol{z}_t \right]. \tag{10.46}$$

We interpret $\varphi_{t+dt} = dF_t/F_t$ as the realized inflation rate over the next instant, $\mu_{\varphi t} = \mathrm{E}_t[\varphi_{t+dt}]$ as the expected inflation rate, and $\boldsymbol{\sigma}_{\varphi t}$ as the sensitivity vector of the inflation rate. Then we have shown that the nominal and the real short-term interest rates are related as follows

$$\tilde{r}_t = r_t + \mu_{\varphi t} - \|\boldsymbol{\sigma}_{\varphi t}\|^2 - \boldsymbol{\sigma}_{\varphi t}^\top \boldsymbol{\lambda}_t, \tag{10.47}$$

i.e. the nominal short rate is equal to the real short rate plus the expected inflation rate minus the variance of the inflation rate minus a risk premium. The presence of the last two terms invalidates the Fisher relation, which says that the nominal interest rate is equal to the sum of the real interest rate and the expected inflation rate.

An application of Itô's Lemma (Exercise 10.2) shows that the dynamics of the nominal state-price deflator is

$$d\tilde{\zeta}_t = -\tilde{\zeta}_t \left[\tilde{r}_t \, dt + \tilde{\boldsymbol{\lambda}}_t^\top \, d\boldsymbol{z}_t \right], \tag{10.48}$$

where $\tilde{\boldsymbol{\lambda}}_t = \boldsymbol{\lambda}_t + \boldsymbol{\sigma}_{\varphi t}$ is the nominal market price of risk.

The time t real price of a real zero-coupon bond maturing at time s is

$$B_t^s = \mathrm{E}_t \left[\frac{\zeta_s}{\zeta_t} \right] = \mathrm{E}_t \left[\exp \left\{ -\int_t^s r_u \, du - \frac{1}{2} \int_t^s \|\boldsymbol{\lambda}_u\|^2 \, du - \int_t^s \boldsymbol{\lambda}_u^\top \, d\boldsymbol{z}_u \right\} \right].$$

The time t nominal price of a nominal zero-coupon bond maturing at T is

$$\tilde{B}_t^T = \mathrm{E}_t \left[\frac{\tilde{\zeta}_T}{\tilde{\zeta}_t} \right] = \mathrm{E}_t \left[\exp \left\{ -\int_t^s \tilde{r}_u \, du - \frac{1}{2} \int_t^s \|\tilde{\boldsymbol{\lambda}}_u\|^2 \, du - \int_t^s \tilde{\boldsymbol{\lambda}}_u^\top \, d\boldsymbol{z}_u \right\} \right].$$

Clearly the prices of nominal bonds are related to the nominal short rate and the nominal market price of risk in exactly the same way as the prices of real bonds are related to the real short rate and the real market price of risk. Models that are based on specific exogenous assumptions about the

short rate dynamics and the market price of risk can be applied both to real term structures and to nominal term structures—but not simultaneously for both real and nominal term structures. This is indeed the case for most popular term structure models. However the equilibrium arguments that some authors offer in support of a particular term structure model, cf. Section 10.5, typically apply to real interest rates and real market prices of risk. The same arguments cannot generally support similar assumptions on nominal rates and market price of risk. Nevertheless, these models are often applied on nominal bonds and term structures.

Above we derived an equilibrium relation between real and nominal short-term interest rates. What can we say about the relation between longer-term real and nominal interest rates? Applying the well-known relation $\operatorname{Cov}[x, y] = \operatorname{E}[xy] - \operatorname{E}[x]\operatorname{E}[y]$, we can write

$$\begin{aligned} \tilde{B}_t^T &= \operatorname{E}_t\left[\frac{\zeta_T}{\zeta_t}\frac{F_t}{F_T}\right] \\ &= \operatorname{E}_t\left[\frac{\zeta_T}{\zeta_t}\right]\operatorname{E}_t\left[\frac{F_t}{F_T}\right] + \operatorname{Cov}_t\left[\frac{\zeta_T}{\zeta_t}, \frac{F_t}{F_T}\right] \\ &= B_t^T \operatorname{E}_t\left[\frac{F_t}{F_T}\right] + \operatorname{Cov}_t\left[\frac{\zeta_T}{\zeta_t}, \frac{F_t}{F_T}\right]. \end{aligned} \tag{10.49}$$

From the dynamics of the state-price deflator and the price index, we get

$$\begin{aligned} \frac{\zeta_T}{\zeta_t} &= \exp\left\{-\int_t^T \left(r_s + \frac{1}{2}\|\boldsymbol{\lambda}_s\|^2\right) ds - \int_t^T \boldsymbol{\lambda}_s^\top \, d\boldsymbol{z}_s\right\}, \\ \frac{F_t}{F_T} &= \exp\left\{-\int_t^T \left(\mu_{\varphi s} - \frac{1}{2}\|\boldsymbol{\sigma}_{\varphi s}\|^2\right) ds - \int_t^T \boldsymbol{\sigma}_{\varphi s}^\top \, d\boldsymbol{z}_s\right\}, \end{aligned}$$

which can be substituted into the above relation between prices on real and nominal bonds. However, the covariance-term on the right-hand side can only be explicitly computed under very special assumptions about the variations over time in r, $\boldsymbol{\lambda}$, μ_φ, and $\boldsymbol{\sigma}_\varphi$.

10.6.2 No real effects of inflation

In this subsection we will take as given some process for the consumer price index and assume that monetary holdings do not affect the utility of the individuals directly. As before the aggregate consumption level is assumed to follow the process

$$dc_t = c_t\left[\mu_{ct}\, dt + \boldsymbol{\sigma}_{ct}^\top \, d\boldsymbol{z}_t\right]$$

so that the dynamics of the real state-price density is

$$d\zeta_t = -\zeta_t\left[r_t\, dt + \boldsymbol{\lambda}_t^\top \, d\boldsymbol{z}_t\right].$$

The short-term real rate is given by

$$r_t = \delta + \frac{-c_t u''(c_t)}{u'(c_t)}\mu_{ct} - \frac{1}{2}c_t^2\frac{u'''(c_t)}{u'(c_t)}\|\boldsymbol{\sigma}_{ct}\|^2 \tag{10.50}$$

and the market price of risk vector is given by

$$\boldsymbol{\lambda}_t = \left(-\frac{c_t u''(c_t)}{u'(c_t)}\right)\boldsymbol{\sigma}_{ct}. \tag{10.51}$$

By substituting the expression (10.51) for λ_t into (4.61), we can write the short-term nominal rate as

$$\tilde{r}_t = r_t + \mu_{\varphi t} - \|\boldsymbol{\sigma}_{\varphi t}\|^2 - \left(-\frac{c_t u''(c_t)}{u'(c_t)}\right)\boldsymbol{\sigma}_{\varphi t}^\top \boldsymbol{\sigma}_{ct}.$$

In the special case where the representative individual has constant relative risk aversion, i.e. $u(c) = c^{1-\gamma}/(1-\gamma)$, and both the aggregate consumption and the price index follow geometric Brownian motions, we get constant rates

$$r = \delta + \gamma\mu_c - \frac{1}{2}\gamma(1+\gamma)\|\boldsymbol{\sigma}_c\|^2, \tag{10.52}$$

$$\tilde{r} = r + \mu_\varphi - \|\boldsymbol{\sigma}_\varphi\|^2 - \gamma\boldsymbol{\sigma}_\varphi^\top \boldsymbol{\sigma}_c. \tag{10.53}$$

Breeden (1986) considers the relations between interest rates, inflation, and aggregate consumption and production in an economy with multiple consumption goods. In general the presence of several consumption goods complicates the analysis considerably. Breeden shows that the equilibrium nominal short rate will depend on both an inflation rate computed using the average weights of the different consumption goods and an inflation rate computed using the marginal weights of the different goods, which are determined by the optimal allocation to the different goods of an extra dollar of total consumption expenditure. The average and the marginal consumption weights will generally be different since the representative individual may shift to other consumption goods as his wealth increases. However, in the special (probably unrealistic) case of Cobb-Douglas type utility function, the relative expenditure weights of the different consumption goods will be constant. For that case Breeden obtains results similar to our one-good conclusions.

10.6.3 A model with real effects of money

In the next model we consider, cash holdings enter the direct utility function of the individual(s). This may be rationalized by the fact that cash holdings facilitate frequent consumption transactions. In such a model the price of the consumption good is determined as a part of the equilibrium of the economy, in contrast to the models studied above where we took an exogenous process for the consumer price index. We follow the set-up of Bakshi and Chen (1996) closely.

The general model

We assume the existence of a representative individual who chooses a consumption process $c = (c_t)$ and a cash process $\tilde{M} = (\tilde{M}_t)$, where $\tilde{M}_t$ is the dollar amount held at time t. As before, let F_t be the unit dollar price of the consumption good. Assume that the representative individual has an infinite time horizon, no endowment stream, and an additively time-separable utility of consumption and the real value of the monetary holdings, i.e. $M_t = \tilde{M}_t/F_t$. At time t the individual has the opportunity to invest in a nominally risk-free bank account with a nominal rate of return of $\tilde{r}_t$. When the individual chooses to hold $\tilde{M}_t$ dollars in cash over the period $[t, t+dt]$, she therefore gives up a dollar return of $\tilde{M}_t\tilde{r}_t\,dt$, which is equivalent to a consumption of $\tilde{M}_t\tilde{r}_t\,dt/F_t$ units of the good. Given a (real) state-price deflator $\zeta = (\zeta_t)$, the total cost of choosing c and M is thus $\mathrm{E}\left[\int_0^\infty \zeta_t(c_t + \tilde{M}_t\tilde{r}_t/F_t)\,dt\right]$. In sum, the optimization problem of the individual can be written as

follows:

$$\sup_{(c_t,\tilde{M}_t)} \mathrm{E}\left[\int_0^\infty e^{-\delta t} u\left(c_t, \tilde{M}_t/F_t\right)\, dt\right]$$
$$\text{s.t. } \mathrm{E}\left[\int_0^\infty \zeta_t \left(c_t + \frac{\tilde{M}_t}{F_t}\tilde{r}_t\right)\, dt\right] \leq W_0,$$

where W_0 is the initial (real) wealth of the individual.

The Lagrangian associated with the optimization problem is

$$\begin{aligned}\mathcal{L} &= \mathrm{E}\left[\int_0^\infty e^{-\delta t} u\left(c_t, \tilde{M}_t/F_t\right)\, dt\right] + \psi\left(W_0 - \mathrm{E}\left[\int_0^\infty \zeta_t\left(c_t + \frac{\tilde{M}_t}{F_t}\tilde{r}_t\right)\, dt\right]\right) \\ &= \psi W_0 + \mathrm{E}\left[\int_0^\infty \left(e^{-\delta t} u\left(c_t, \tilde{M}_t/F_t\right) - \psi\zeta_t\left(c_t + \frac{\tilde{M}_t}{F_t}\tilde{r}_t\right)\right)\, dt\right].\end{aligned}$$

If we maximize the integrand "state-by-state", we will also maximize the expectation. The first order conditions are

$$e^{-\delta t} u_c(c_t, \tilde{M}_t/F_t) = \psi\zeta_t, \tag{10.54}$$
$$e^{-\delta t} u_M(c_t, \tilde{M}_t/F_t) = \psi\zeta_t\tilde{r}_t, \tag{10.55}$$

where u_c and u_M are the first-order derivatives of u with respect to the first and second argument, respectively. The Lagrange multiplier ψ is set so that the budget condition holds as an equality. Again, we see that the state-price deflator is given in terms of the marginal utility with respect to consumption. Imposing the initial value $\zeta_0 = 1$ and recalling the definition of M_t, we have

$$\zeta_t = e^{-\delta t}\frac{u_c(c_t, M_t)}{u_c(c_0, M_0)}. \tag{10.56}$$

We can apply the state-price deflator to value all payment streams. For example, an investment of one dollar at time t in the nominal bank account generates a continuous payment stream at the rate of $\tilde{r}_s$ dollars to the end of all time. The corresponding real investment at time t is $1/F_t$ and the real dividend at time s is $\tilde{r}_s/F_s$. Hence, we have the relation

$$\frac{1}{F_t} = \mathrm{E}_t\left[\int_t^\infty \frac{\zeta_s}{\zeta_t}\frac{\tilde{r}_s}{F_s}\, ds\right],$$

or, equivalently,

$$\frac{1}{F_t} = \mathrm{E}_t\left[\int_t^\infty e^{-\delta(s-t)}\frac{u_c(c_s, M_s)}{u_c(c_t, M_t)}\frac{\tilde{r}_s}{F_s}\, ds\right]. \tag{10.57}$$

Substituting the first optimality condition (10.54) into the second (10.55), we see that the nominal short rate is given by

$$\tilde{r}_t = \frac{u_M(c_t, \tilde{M}_t/F_t)}{u_c(c_t, \tilde{M}_t/F_t)}. \tag{10.58}$$

The intuition behind this relation can be explained in the following way. If you have an extra dollar now you can either keep it in cash or invest it in the nominally risk-free bank account. If you keep it in cash your utility grows by $u_M(c_t, \tilde{M}_t/F_t)/F_t$. If you invest it in the bank account you will earn a dollar interest of $\tilde{r}_t$ that can be used for consuming $\tilde{r}_t/F_t$ extra units of consumption, which will increase your utility by $u_c(c_t, \tilde{M}_t/F_t)\tilde{r}_t/F_t$. At the optimum, these utility increments must be

identical. Combining (10.57) and (10.58), we get that the price index must satisfy the recursive relation

$$\frac{1}{F_t} = \mathrm{E}_t\left[\int_t^\infty e^{-\delta(s-t)}\frac{u_M(c_s,M_s)}{u_c(c_t,M_t)}\frac{1}{F_s}\,ds\right]. \tag{10.59}$$

Let us find expressions for the equilibrium real short rate and the market price of risk in this setting. As always, the real short rate equals minus the percentage drift of the state-price deflator, while the market price of risk equals minus the percentage sensitivity vector of the state-price deflator. In an equilibrium, the representative individual must consume the aggregate consumption and hold the total money supply in the economy. Suppose that the aggregate consumption and the money supply follow exogenous processes of the form

$$\begin{aligned} dc_t &= c_t\left[\mu_{ct}\,dt + \boldsymbol{\sigma}_{ct}^\top\,d\boldsymbol{z}_t\right],\\ d\tilde{M}_t &= \tilde{M}_t\left[\tilde{\mu}_{Mt}\,dt + \tilde{\boldsymbol{\sigma}}_{Mt}^\top\,d\boldsymbol{z}_t\right]. \end{aligned}$$

Assuming that the endogenously determined price index will follow a similar process,

$$dF_t = F_t\left[\mu_{\varphi t}\,dt + \boldsymbol{\sigma}_{\varphi t}^\top\,d\boldsymbol{z}_t\right],$$

the dynamics of $M_t = \tilde{M}_t/F_t$ will be

$$dM_t = M_t\left[\mu_{Mt}\,dt + \boldsymbol{\sigma}_{Mt}^\top\,d\boldsymbol{z}_t\right],$$

where

$$\mu_{Mt} = \tilde{\mu}_{Mt} - \mu_{\varphi t} + \|\boldsymbol{\sigma}_{\varphi t}\|^2 - \tilde{\boldsymbol{\sigma}}_{Mt}^\top\boldsymbol{\sigma}_{\varphi t}, \qquad \boldsymbol{\sigma}_{Mt} = \tilde{\boldsymbol{\sigma}}_{Mt} - \boldsymbol{\sigma}_{\varphi t}.$$

Given these equations and the relation (10.56), we can find the drift and the sensitivity vector of the state-price deflator by an application of Itô's Lemma (Exercise 10.5). We find that the equilibrium real short-term interest rate can be written as

$$\begin{aligned} r_t = \delta &+ \left(\frac{-c_t u_{cc}(c_t,M_t)}{u_c(c_t,M_t)}\right)\mu_{ct} + \left(\frac{-M_t u_{cM}(c_t,M_t)}{u_c(c_t,M_t)}\right)\mu_{Mt}\\ &- \frac{1}{2}\frac{c_t^2 u_{ccc}(c_t,M_t)}{u_c(c_t,M_t)}\|\boldsymbol{\sigma}_{ct}\|^2 - \frac{1}{2}\frac{M_t^2 u_{cMM}(c_t,M_t)}{u_c(c_t,M_t)}\|\boldsymbol{\sigma}_{Mt}\|^2 - \frac{c_t M_t u_{ccM}(c_t,M_t)}{u_c(c_t,M_t)}\boldsymbol{\sigma}_{ct}^\top\boldsymbol{\sigma}_{Mt}, \end{aligned} \tag{10.60}$$

while the market price of risk vector is

$$\begin{aligned} \lambda_t &= \left(-\frac{c_t u_{cc}(c_t,M_t)}{u_c(c_t,M_t)}\right)\boldsymbol{\sigma}_{ct} + \left(\frac{-M_t u_{cM}(c_t,M_t)}{u_c(c_t,M_t)}\right)\boldsymbol{\sigma}_{Mt}\\ &= \left(-\frac{c_t u_{cc}(c_t,M_t)}{u_c(c_t,M_t)}\right)\boldsymbol{\sigma}_{ct} + \left(\frac{-M_t u_{cM}(c_t,M_t)}{u_c(c_t,M_t)}\right)(\tilde{\boldsymbol{\sigma}}_{Mt} - \boldsymbol{\sigma}_{\varphi t}). \end{aligned} \tag{10.61}$$

With $u_{cM} < 0$, we see that assets that are positively correlated with the inflation rate will have a lower expected real return, other things equal. Intuitively such assets are useful for hedging inflation risk so that they do not have to offer as high an expected return.

The relation (4.61) is also valid in the present setting. Substituting the expression (10.61) for the market price of risk into (4.61), we obtain

$$\tilde{r}_t - r_t - \mu_{\varphi t} + \|\boldsymbol{\sigma}_{\varphi t}\|^2 = -\left(-\frac{c_t u''(c_t)}{u'(c_t)}\right)\boldsymbol{\sigma}_{\varphi t}^\top\boldsymbol{\sigma}_{ct} - \left(\frac{-M_t u_{cM}(c_t,M_t)}{u_c(c_t,M_t)}\right)\boldsymbol{\sigma}_{\varphi t}^\top\boldsymbol{\sigma}_{Mt}. \tag{10.62}$$

An example

To obtain more concrete results, we must specify the utility function and the exogenous processes c and $\tilde{M}$. Assume a utility function of the Cobb-Douglas type,

$$u(c, M) = \frac{\left(c^{\varphi} M^{1-\varphi}\right)^{1-\gamma}}{1-\gamma},$$

where φ is a constant between zero and one, and γ is a positive constant. The limiting case for $\gamma = 1$ is log utility,

$$u(c, M) = \varphi \ln c + (1-\varphi) \ln M.$$

By inserting the relevant derivatives into (10.60), we see that the real short rate becomes

$$\begin{aligned} r_t = \delta &+ [1-\varphi(1-\gamma)]\mu_{ct} - (1-\varphi)(1-\gamma)\mu_{Mt} - \frac{1}{2}[1-\varphi(1-\gamma)][2-\varphi(1-\gamma)]\|\boldsymbol{\sigma}_{ct}\|^2 \\ &+ \frac{1}{2}(1-\varphi)(1-\gamma)[1-(1-\varphi)(1-\gamma)]\|\boldsymbol{\sigma}_{Mt}\|^2 + (1-\varphi)(1-\gamma)[1-\varphi(1-\gamma)]\boldsymbol{\sigma}_{ct}^{\top}\boldsymbol{\sigma}_{Mt}, \end{aligned} \tag{10.63}$$

which for $\gamma = 1$ simplifies to

$$r_t = \delta + \mu_{ct} - \|\boldsymbol{\sigma}_{ct}\|^2. \tag{10.64}$$

We see that with log utility, the real short rate will be constant if aggregate consumption $c = (c_t)$ follows a geometric Brownian motion. From (10.58), the nominal short rate is

$$\tilde{r}_t = \frac{1-\varphi}{\varphi}\frac{c_t}{M_t}. \tag{10.65}$$

The ratio c_t/M_t is called the *velocity of money*. If the velocity of money is constant, the nominal short rate will be constant. Since $M_t = \tilde{M}_t/F_t$ and F_t is endogenously determined, the velocity of money will also be endogenously determined.

We have to determine the price level in the economy, which is given only recursively in (10.59). This is possible under the assumption that both C and $\tilde{M}$ follow geometric Brownian motions. We conjecture that $F_t = k\tilde{M}_t/c_t$ for some constant k. From (10.59), we get

$$\frac{1}{k} = \frac{1-\varphi}{\varphi}\int_t^{\infty} e^{-\delta(s-t)}\,\mathrm{E}_t\left[\left(\frac{c_s}{c_t}\right)^{1-\gamma}\left(\frac{\tilde{M}_s}{\tilde{M}_t}\right)^{-1}\right] ds.$$

Inserting the relations

$$\begin{aligned} \frac{c_s}{c_t} &= \exp\left\{\left(\mu_c - \frac{1}{2}\|\boldsymbol{\sigma}_c\|^2\right)(s-t) + \boldsymbol{\sigma}_c^{\top}(\boldsymbol{z}_s - \boldsymbol{z}_t)\right\}, \\ \frac{\tilde{M}_s}{\tilde{M}_t} &= \exp\left\{\left(\tilde{\mu}_M - \frac{1}{2}\|\tilde{\boldsymbol{\sigma}}_M\|^2\right)(s-t) + \tilde{\boldsymbol{\sigma}}_M^{\top}(\boldsymbol{z}_s - \boldsymbol{z}_t)\right\}, \end{aligned}$$

and applying a standard rule for expectations of lognormal variables, we get

$$\begin{aligned} \frac{1}{k} = \frac{1-\varphi}{\varphi}\int_t^{\infty} \exp\Big\{\Big(&-\delta + (1-\gamma)(\mu_c - \frac{1}{2}\|\boldsymbol{\sigma}_c\|^2) - \tilde{\mu}_M + \|\tilde{\boldsymbol{\sigma}}_M\|^2 \\ &+ \frac{1}{2}(1-\gamma)^2\|\tilde{\boldsymbol{\sigma}}_c\|^2 - (1-\gamma)\boldsymbol{\sigma}_c^{\top}\tilde{\boldsymbol{\sigma}}_M\Big)(s-t)\Big\}\, ds, \end{aligned}$$

which implies that the conjecture is true with

$$k = \frac{\varphi}{1-\varphi}\left(\delta - (1-\gamma)(\mu_c - \frac{1}{2}\|\boldsymbol{\sigma}_c\|^2) + \tilde{\mu}_M - \|\tilde{\boldsymbol{\sigma}}_M\|^2 - \frac{1}{2}(1-\gamma)^2\|\boldsymbol{\sigma}_c\|^2 + (1-\gamma)\boldsymbol{\sigma}_c^{\top}\tilde{\boldsymbol{\sigma}}_M\right).$$

From an application of Itô's Lemma, it follows that the price index also follows a geometric Brownian motion

$$dF_t = F_t \left[\mu_\varphi \, dt + \boldsymbol{\sigma}_\varphi^\top \, d\boldsymbol{z}_t\right], \tag{10.66}$$

where

$$\mu_\varphi = \tilde{\mu}_M - \mu_c + \|\boldsymbol{\sigma}_c\|^2 - \tilde{\boldsymbol{\sigma}}_M^\top \boldsymbol{\sigma}_c, \qquad \boldsymbol{\sigma}_\varphi = \tilde{\boldsymbol{\sigma}}_M - \boldsymbol{\sigma}_c.$$

With $F_t = k\tilde{M}_t/c_t$, we have $M_t = c_t/k$, so that the velocity of money $c_t/M_t = k$ is constant, and the nominal short rate becomes

$$\tilde{r}_t = \frac{1-\varphi}{\varphi}k = \delta - (1-\gamma)(\mu_c - \frac{1}{2}\|\boldsymbol{\sigma}_c\|^2) + \tilde{\mu}_M - \|\tilde{\boldsymbol{\sigma}}_M\|^2 - \frac{1}{2}(1-\gamma)^2\|\boldsymbol{\sigma}_c\|^2 + (1-\gamma)\boldsymbol{\sigma}_c^\top \tilde{\boldsymbol{\sigma}}_M, \tag{10.67}$$

which is also a constant. With log utility, the nominal rate simplifies to $\delta + \tilde{\mu}_M - \|\tilde{\boldsymbol{\sigma}}_M\|^2$. In order to obtain the real short rate in the non-log case, we have to determine μ_{Mt} and $\boldsymbol{\sigma}_{Mt}$ and plug into (10.63). We get $\mu_{Mt} = \mu_c + \frac{1}{2}\|\boldsymbol{\sigma}_c\|^2 + \tilde{\boldsymbol{\sigma}}_M^\top \boldsymbol{\sigma}_c$ and $\boldsymbol{\sigma}_{Mt} = \boldsymbol{\sigma}_c$ and hence

$$r_t = \delta + \gamma\mu_c - \gamma\|\boldsymbol{\sigma}_c\|^2 \left[\frac{1}{2}(1+\gamma) + \varphi(1-\gamma)\right], \tag{10.68}$$

which is also a constant. In comparison with (10.52) for the case where money has no real effects, the last term in the equation above is new.

Another example

Bakshi and Chen (1996) also study another model specification in which both nominal and real short rates are time-varying, but evolve independently of each other. To obtain stochastic interest rates we have to specify more general processes for aggregate consumption and money supply than the geometric Brownian motions used above. They assume log-utility ($\gamma = 1$) in which case we have already seen that

$$r_t = \delta + \mu_{ct} - \|\boldsymbol{\sigma}_{ct}\|^2, \qquad \tilde{r}_t = \frac{1-\varphi}{\varphi}\frac{c_t}{M_t} = \frac{1-\varphi}{\varphi}\frac{c_t F_t}{\tilde{M}_t}.$$

The dynamics of aggregate consumption is assumed to be

$$dc_t = c_t \left[(\alpha_c + \kappa_c X_t)\, dt + \sigma_c \sqrt{X_t}\, dz_{1t}\right],$$

where X can be interpreted as a technology variable and is assumed to follow the process

$$dX_t = \kappa_x(\theta_x - X_t)\, dt + \sigma_x \sqrt{X_t}\, dz_{1t}.$$

The money supply is assumed to be $\tilde{M}_t = \tilde{M}_0 e^{\mu_M^* t} g_t/g_0$, where

$$dg_t = g_t \left[\kappa_g(\theta_g - g_t)\, dt + \sigma_g \sqrt{g_t}\left(\rho_{CM}\, dz_{1t} + \sqrt{1-\rho_{CM}^2}\, dz_{2t}\right)\right],$$

and where z_1 and z_2 are independent one-dimensional Brownian motions. Following the same basic procedure as in the previous model specification, the authors show that the real short rate is

$$r_t = \delta + \alpha_c + (\kappa_c - \sigma_c^2)X_t, \tag{10.69}$$

while the nominal short rate is

$$\tilde{r}_t = \frac{(\delta + \mu_M^*)(\delta + \mu_M^* + \kappa_g\theta_g)}{\delta + \mu_M^* + (\kappa_g + \sigma_g^2)g_t}. \tag{10.70}$$

Both rates are time-varying. The real rate is driven by the technology variable X, while the nominal rate is driven by the monetary shock process g. In this set-up, shocks to the real economy have opposite effects of the same magnitude on real rates and inflation so that nominal rates are unaffected.

The real price of a real zero-coupon bond maturing at time T is of the form

$$B_t^T = e^{-a(T-t)-b(T-t)x},$$

while the nominal price of a nominal zero-coupon bond maturing at T is

$$\tilde{B}_t^T = \frac{\tilde{a}(T-t)+\tilde{b}(T-t)g_t}{\delta+\mu_M^*+(\kappa_g+\sigma_g^2)g_t},$$

where a, b, $\tilde{a}$, and $\tilde{b}$ are deterministic functions of time for which Bakshi and Chen provide closed-form expressions.

In the very special case where these processes are uncorrelated, i.e. $\rho_{CM} = 0$, the real and nominal term structures of interest rates are independent of each other! Although this is an extreme result, it does point out that real and nominal term structures in general may have quite different properties.

10.7 The expectation hypothesis

The **expectation hypothesis** relates the current interest rates and yields to expected future interest rates or returns. This basic issue was discussed already by Fisher (1896) and further developed and concretized by Hicks (1939) and Lutz (1940). The original motivation of the hypothesis is that when lenders (bond investors) and borrowers (bond issuers) decide between long-term or short-term bonds, they will compare the price or yield of a long-term bond to the expected price or return on a roll-over strategy in short-term bonds. Hence, long-term rates and expected future short-term rates will be linked. Of course, a cornerstone of modern finance theory is that, when comparing different strategies, investors will also take the risks into account. So even before going into the specifics of the hypothesis you should really be quite skeptical, at least when it comes to very strict interpretations of the expectation hypothesis.

The vague idea that current yields and interest rates are linked to expected future rates and returns can be concretized in a number of ways. Below we will present and evaluate a number of versions. This analysis follows Cox, Ingersoll, and Ross (1981a) quite closely. We find that some versions are equivalent, some versions inconsistent. We end up concluding that none of the variants of the expectations hypothesis are consistent with any realistic behavior of interest rates. Hence, the analysis of the shape of the yield curve and models of term structure dynamics should not be based on this hypothesis. Hence, it is surprising, maybe even disappointing, that empirical tests of the expectation hypothesis have generated such a huge literature in the past and that the hypothesis still seems to be widely accepted among economists.

10.7.1 Versions of the pure expectation hypothesis

The first version of the pure expectation hypothesis that we will discuss says that prices in the bond markets are set so that the expected gross returns on all self-financing trading strategies

over a given period are identical. In particular, the expected gross return from buying at time t a zero-coupon bond maturing at time T and reselling it at time $t' \leq T$, which is given by $\mathrm{E}_t[B_{t'}^T/B_t^T]$, will be independent of the maturity date T of the bond (but generally not independent of t'). Let us refer to this as the *gross return* pure expectation hypothesis.

This version of the hypothesis is consistent with pricing in a world of risk-neutral investors. If we have a representative individual with time-additive expected utility, we know that zero-coupon bond prices satisfy

$$B_t^T = \mathrm{E}_t\left[e^{-\delta(t'-t)}\frac{u'(c_{t'})}{u'(c_t)}B_{t'}^T\right],$$

where u is the instantaneous utility function, δ is the time preference rate, and C denotes aggregate consumption. If the representative individual is risk-neutral, his marginal utility is constant, which implies that

$$\mathrm{E}_t\left[\frac{B_{t'}^T}{B_t^T}\right] = e^{\delta(t'-t)}, \tag{10.71}$$

which is clearly independent of T. Clearly, the assumption of risk-neutrality is not very attractive. There is also another serious problem with this hypothesis. As is to be shown in Exercise 10.4, it cannot hold when interest rates are uncertain.

A slight variation of the above is to align all expected continuously compounded returns, i.e. $\frac{1}{t'-t}\mathrm{E}_t[\ln\left(B_{t'}^T/B_t^T\right)]$ for all T. In particular with $T = t'$, the expected continuously compounded rate of return is known to be equal to the zero-coupon yield for maturity t', which we denote by $y_t^{t'} = -\frac{1}{t'-t}\ln B_t^{t'}$. We can therefore formulate the hypothesis as

$$\frac{1}{t'-t}\mathrm{E}_t\left[\ln\left(\frac{B_{t'}^T}{B_t^T}\right)\right] = y_t^{t'}, \text{ all } T \geq t'.$$

Let us refer to this as the *rate of return* pure expectation hypothesis. For $t' \to t$, the right-hand side approaches the current short rate r_t, while the left-hand side approaches the absolute drift rate of $\ln B_t^T$.

An alternative specification of the pure expectation hypothesis claims that the expected return over the next time period is the same for all investments in bonds and deposits. In other words there is no difference between expected returns on long-maturity and short-maturity bonds. In the continuous-time limit we consider returns over the next instant. The risk-free return over $[t, t+dt]$ is $r_t\,dt$, so for any zero-coupon bond, the hypothesis claims that

$$\mathrm{E}_t\left[\frac{dB_t^T}{B_t^T}\right] = r_t\,dt, \quad \text{for all } T > t, \tag{10.72}$$

or, equivalently,[4] that

$$B_t^T = \mathrm{E}_t\left[e^{-\int_t^T r_s\,ds}\right], \quad \text{for all } T > t.$$

This is the *local* pure expectations hypothesis.

[4]Here and later we use that, under suitable regularity conditions, the relative drift rate of an Itô process $X = (X_t)$ is given by the process $\mu = (\mu_t)$ if and only if $X_t = \mathrm{E}_t[X_T \exp\{-\int_t^T \mu_s\,ds\}]$. Suppose first that the relative drift rate is given by μ so that $dX_t = X_t[\mu_t\,dt + \boldsymbol{\sigma}_t^\top\,d\boldsymbol{z}_t]$. Then an application of Itô's Lemma reveals that the process $X_t \exp\{-\int_0^t \mu_s\,ds\}$ is a martingale so that $X_t \exp\{-\int_0^t \mu_s\,ds\} = \mathrm{E}_t[X_T \exp\{-\int_0^T \mu_s\,ds\}]$ and hence $X_t = \mathrm{E}_t[X_T \exp\{-\int_t^T \mu_s\,ds\}]$.

The absolute drift of X is the limit of $\frac{1}{\Delta t}\mathrm{E}_t[X_{t+\Delta t} - X_t]$ as $\Delta t \to 0$. If $X_t = \mathrm{E}_t[X_T \exp\{-\int_t^T \mu_s\,ds\}]$ for all t,

Another interpretation says that the return from holding a zero-coupon bond to maturity should equal the expected return from rolling over short-term bonds over the same time period, i.e.

$$\frac{1}{B_t^T} = \mathrm{E}_t\left[e^{\int_t^T r_s\,ds}\right], \quad \text{for all } T > t \tag{10.73}$$

or, equivalently,

$$B_t^T = \left(\mathrm{E}_t\left[e^{\int_t^T r_s\,ds}\right]\right)^{-1}, \quad \text{for all } T > t.$$

This is the *return-to-maturity* pure expectation hypothesis.

A related claim is that the yield on any zero-coupon bond should equal the "expected yield" on a roll-over strategy in short bonds. Since an investment of one at time t in the bank account generates $e^{\int_t^T r_s\,ds}$ at time T, the ex-post realized yield is $\frac{1}{T-t}\int_t^T r_s\,ds$. Hence, this *yield-to-maturity* pure expectation hypothesis says that

$$y_t^T = -\frac{1}{T-t}\ln B_t^T = \mathrm{E}_t\left[\frac{1}{T-t}\int_t^T r_s\,ds\right], \tag{10.74}$$

or, equivalently,

$$B_t^T = e^{-\mathrm{E}_t\left[\int_t^T r_s\,ds\right]}, \quad \text{for all } T > t.$$

Finally, the *unbiased* pure expectation hypothesis states that the forward rate for time T prevailing at time $t < T$ is equal to the time t expectation of the short rate at time T, i.e. that forward rates are unbiased estimates of future spot rates. In symbols,

$$f_t^T = \mathrm{E}_t[r_T], \quad \text{for all } T > t.$$

This implies that

$$-\ln B_t^T = \int_t^T f_t^s\,ds = \int_t^T \mathrm{E}_t[r_s]\,ds = \mathrm{E}_t\left[\int_t^T r_s\,ds\right],$$

from which we see that the unbiased version of the pure expectation hypothesis is indistinguishable from the yield-to-maturity version.

We will first show that *the different versions are inconsistent* when future rates are uncertain. This follows from an application of Jensen's inequality which states that if X is a random variable and f is a convex function, i.e. $f'' > 0$, then $\mathrm{E}[f(X)] > f(\mathrm{E}[X])$. Since $f(x) = e^x$ is a convex function, we have $\mathrm{E}[e^X] > e^{\mathrm{E}[X]}$ for any random variable X. In particular for $X = \int_t^T r_s\,ds$, we get

$$\mathrm{E}_t\left[e^{\int_t^T r_s\,ds}\right] > e^{\mathrm{E}_t\left[\int_t^T r_s\,ds\right]} \quad \Rightarrow \quad e^{-\mathrm{E}_t\left[\int_t^T r_s\,ds\right]} > \left(\mathrm{E}_t\left[e^{\int_t^T r_s\,ds}\right]\right)^{-1}.$$

---

then

$$\begin{aligned}
\frac{1}{\Delta t}\mathrm{E}_t[X_{t+\Delta t} - X_t] &= \frac{1}{\Delta t}\mathrm{E}_t\left[\left(\mathrm{E}_{t+\Delta t}\left[X_T e^{-\int_{t+\Delta t}^T \mu_s\,ds}\right]\right) - \left(\mathrm{E}_t\left[X_T e^{-\int_t^T \mu_s\,ds}\right]\right)\right] \\
&= \frac{1}{\Delta t}\mathrm{E}_t\left[X_T e^{-\int_{t+\Delta t}^T \mu_s\,ds} - X_T e^{-\int_t^T \mu_s\,ds}\right] \\
&= \mathrm{E}_t\left[X_T e^{-\int_t^T \mu_s\,ds}\frac{e^{\int_t^{t+\Delta t}\mu_s\,ds} - 1}{\Delta t}\right] \\
&\to \mu_t\,\mathrm{E}_t\left[X_T e^{-\int_t^T \mu_s\,ds}\right] = \mu_t X_t,
\end{aligned}$$

i.e. the relative drift rate equals μ_t.

This shows that the bond price according to the yield-to-maturity version is strictly greater than the bond price according to the return-to-maturity version. For $X = -\int_t^T r_s\,ds$, we get

$$\mathrm{E}_t\left[e^{-\int_t^T r_s\,ds}\right] > e^{\mathrm{E}_t\left[-\int_t^T r_s\,ds\right]} = e^{-\mathrm{E}_t\left[\int_t^T r_s\,ds\right]},$$

hence the bond price according to the local version of the hypothesis is strictly greater than the bond price according to the yield-to-maturity version. We can conclude that at most one of the versions of the local, return-to-maturity, and yield-to-maturity pure expectations hypothesis can hold.

10.7.2 The pure expectation hypothesis and equilibrium

Next, let us see whether the different versions can be consistent with any equilibrium. Assume that interest rates and bond prices are generated by a d-dimensional standard Brownian motion $\boldsymbol{z}$. Assuming absence of arbitrage there exists a market price of risk process λ so that for any maturity T, the zero-coupon bond price dynamics is of the form

$$dB_t^T = B_t^T\left[\left(r_t + \left(\boldsymbol{\sigma}_t^T\right)^\top \boldsymbol{\lambda}_t\right)dt + \left(\boldsymbol{\sigma}_t^T\right)^\top d\boldsymbol{z}_t\right], \tag{10.75}$$

where $\boldsymbol{\sigma}_t^T$ denotes the d-dimensional sensitivity vector of the bond price. Recall that the same $\boldsymbol{\lambda}_t$ applies to all zero-coupon bonds so that $\boldsymbol{\lambda}_t$ is independent of the maturity of the bond. Comparing with (10.72), we see that the local expectation hypothesis will hold if and only if $\left(\boldsymbol{\sigma}_t^T\right)^\top \boldsymbol{\lambda}_t = 0$ for all T. This is true if either investors are risk-neutral or interest rate risk is uncorrelated with aggregate consumption. Neither of these conditions hold in real life.

To evaluate the return-to-maturity version, first note that an application of Itô's Lemma on (10.75) show that

$$d\left(\frac{1}{B_t^T}\right) = \frac{1}{B_t^T}\left[\left(-r_t - \left(\boldsymbol{\sigma}_t^T\right)^\top \boldsymbol{\lambda}_t + \|\boldsymbol{\sigma}_t^T\|^2\right)dt - \left(\boldsymbol{\sigma}_t^T\right)^\top d\boldsymbol{z}_t\right].$$

On the other hand, according to the hypothesis (10.73) the relative drift of $1/B_t^T$ equals $-r_t$; cf. a previous footnote. To match the two expressions for the drift, we must have

$$\left(\boldsymbol{\sigma}_t^T\right)^\top \boldsymbol{\lambda}_t = \|\boldsymbol{\sigma}_t^T\|^2, \quad \text{for all } T. \tag{10.76}$$

Is this possible? Cox, Ingersoll, and Ross (1981a) conclude that it is impossible. If the exogenous shock $\boldsymbol{z}$ and therefore $\boldsymbol{\sigma}_t^T$ and $\boldsymbol{\lambda}_t$ are one-dimensional, they are right, since $\boldsymbol{\lambda}_t$ must then equal $\boldsymbol{\sigma}_t^T$, and this must hold for all T. Since $\boldsymbol{\lambda}_t$ is independent of T and the volatility $\boldsymbol{\sigma}_t^T$ approaches zero for $T \to t$, this can only hold if $\boldsymbol{\lambda}_t \equiv 0$ (risk-neutral investors) or $\boldsymbol{\sigma}_t^T \equiv 0$ (deterministic interest rates). However, as pointed out by McCulloch (1993) and Fisher and Gilles (1998), in multi-dimensional cases the key condition (10.76) may indeed hold, at least in very special cases. Let φ be a d-dimensional function with the property that $\|\varphi(\tau)\|^2$ is independent of τ. Define $\boldsymbol{\lambda}_t = 2\varphi(0)$ and $\boldsymbol{\sigma}_t^T = \varphi(0) - \varphi(T-t)$. Then (10.76) is indeed satisfied. However, all such functions φ seem to generate very strange bond price dynamics. The examples given in the two papers mentioned above are

$$\varphi(\tau) = k\begin{pmatrix}\sqrt{2e^{-\tau} - e^{-2\tau}}\\ 1 - e^{-\tau}\end{pmatrix}, \qquad \varphi(\tau) = k_1\begin{pmatrix}\cos(k_2\tau)\\ \sin(k_2\tau)\end{pmatrix},$$

where k, k_1, k_2 are constants.

As discussed above, the rate or return version implies that the absolute drift rate of the log-bond price equals the short rate. We can see from (10.74) that the same is true for the yield-to-maturity version and hence the unbiased version.[5] On the other hand Itô's Lemma and (10.75) imply that

$$d\left(\ln B_t^T\right) = \left(r_t + (\boldsymbol{\sigma}_t^T)^\top \boldsymbol{\lambda}_t - \frac{1}{2}\|\boldsymbol{\sigma}_t^T\|^2\right) dt + \left(\boldsymbol{\sigma}_t^T\right)^\top d\boldsymbol{z}_t. \tag{10.77}$$

Hence, these versions of the hypothesis will hold if and only if

$$(\boldsymbol{\sigma}_t^T)^\top \boldsymbol{\lambda}_t = \frac{1}{2}\|\boldsymbol{\sigma}_t^T\|^2, \quad \text{for all } T.$$

Again, it is possible that the condition holds. Just let φ and $\boldsymbol{\sigma}_t^T$ be as for the return-to-maturity hypothesis and let $\boldsymbol{\lambda}_t = \varphi(0)$. But such specifications are not likely to represent real life term structures.

The conclusion to be drawn from this analysis is that neither of the different versions of the pure expectation hypothesis seem to be consistent with any reasonable description of the term structure of interest rates.

10.7.3 The weak expectation hypothesis

Above we looked at versions of the *pure* expectation hypothesis that all aligns an expected return or yield with a current interest rate or yield. However, as pointed out by Campbell (1986), there is also a *weak* expectation hypothesis that allows for a difference between the relevant expected return/yield and the current rate/yield, but restricts this difference to be constant over time.

The local weak expectation hypothesis says that

$$\mathrm{E}_t\left[\frac{dB_t^T}{B_t^T}\right] = (r_t + g(T-t))\, dt$$

for some deterministic function g. In the pure version g is identically zero. For a given time-to-maturity there is a constant "instantaneous holding term premium". Comparing with (10.75), we see that this hypothesis will hold when the market price of risk $\boldsymbol{\lambda}_t$ is constant and the bond price sensitivity vector $\boldsymbol{\sigma}_t^T$ is a deterministic function of time-to-maturity. These conditions are satisfied in the Vasicek (1977) model and in other models of the Gaussian class.

Similarly, the weak yield-to-maturity expectation hypothesis says that

$$f_t^T = \mathrm{E}_t[r_T] + h(T-t)$$

for some deterministic function h with $h(0) = 0$, i.e. that there is a constant "instantaneous forward term premium". The pure version requires h to be identically equal to zero. It can be shown that this condition implies that the drift of $\ln B_t^T$ equals $r_t + h(T-t)$.[6] Comparing with (10.77), we

[5] According to the yield-to-maturity hypothesis

$$\frac{1}{\Delta t}\mathrm{E}_t\left[\ln B_{t+\Delta t}^T - \ln B_t^T\right] = \frac{1}{\Delta t}\mathrm{E}_t\left[-\mathrm{E}_{t+\Delta t}\left[\int_{t+\Delta t}^T r_s\, ds\right] + \mathrm{E}_t\left[\int_t^T r_s\, ds\right]\right] = \frac{1}{\Delta t}\mathrm{E}_t\left[\int_t^{t+\Delta t} r_s\, ds\right],$$

which approaches r_t as $\Delta t \to 0$. This means that the absolute drift of $\ln B_t$ equals r_t.

[6] From the weak yield-to-maturity hypothesis, it follows that $-\ln B_t^T = \int_t^T (\mathrm{E}_t[r_s] + h(s-t))\, ds$. Hence,

$$\begin{aligned}\frac{1}{\Delta t}\mathrm{E}_t\left[\ln B_{t+\Delta t}^T - \ln B_t^T\right] &= \frac{1}{\Delta t}\mathrm{E}_t\left[-\int_{t+\Delta t}^T (\mathrm{E}_{t+\Delta t}[r_s] + h(s-(t+\Delta t)))\, ds + \int_t^T (\mathrm{E}_t[r_s] + h(s-t))\, ds\right] \\ &= \frac{1}{\Delta t}\mathrm{E}_t\left[\int_t^{t+\Delta t} r_s\, ds\right] - \frac{1}{\Delta t}\left(\int_{t+\Delta t}^T h(s-(t+\Delta t))\, ds - \int_t^T h(s-t)\, ds\right).\end{aligned}$$

see that also this hypothesis will hold when λ_t is constant and σ_t^T is a deterministic function of $T-t$ as is the case in the Gaussian models.

The class of Gaussian models have several unrealistic properties. For example, such models allow negative interest rates and requires bond and interest rate volatilities to be independent of the level of interest rates. So far, the validity of even weak versions of the expectation hypothesis has not been shown in more realistic term structure models.

10.8 Liquidity preference, market segmentation, and preferred habitats

Another traditional explanation of the shape of the yield curve is given by the **liquidity preference hypothesis** introduced by Hicks (1939). He realized that the expectation hypothesis basically ignores investors' aversion towards risk and argued that expected returns on long-term bonds should exceed the expected returns on short-term bonds to compensate for the higher price fluctuations of long-term bonds. According to this view the yield curve should tend to be increasing. Note that the word "liquidity" in the name of the hypothesis is not used in the usual sense of the word. Short-term bonds are not necessarily more liquid than long-term bonds. A better name would be "the maturity preference hypothesis".

In contrast the **market segmentation hypothesis** introduced by Culbertson (1957) claims that investors will typically prefer to invest in bonds with time-to-maturity in a certain interval, a maturity segment, perhaps in an attempt to match liabilities with similar maturities. For example, a pension fund with liabilities due in 20-30 years can reduce risk by investing in bonds of similar maturity. On the other hand, central banks typically operate in the short end of the market. Hence, separated market segments can exist without any relation between the bond prices and the interest rates in different maturity segments. If this is really the case, we cannot expect to see continuous or smooth yield curves and discount functions across the different segments.

A more realistic version of this hypothesis is the **preferred habitats hypothesis** put forward by Modigliani and Sutch (1966). An investor may prefer bonds with a certain maturity, but should be willing to move away from that maturity if she is sufficiently compensated in terms of a higher yield.[7] The different segments are therefore not completely independent of each other, and yields and discount factors should depend on maturity in a smooth way.

It is really not possible to quantify the market segmentation or the preferred habitats hypothesis without setting up an economy with individuals having different favorite maturities. The resulting equilibrium yield curve will depend heavily on the degree of risk aversion of the various individuals as illustrated by an analysis of Cox, Ingersoll, and Ross (1981a).

The limit of $\frac{1}{\Delta t}\left(\int_{t+\Delta t}^T h(s-(t+\Delta t))\,ds - \int_t^T h(s-t)\,ds\right)$ as $\Delta t \to 0$ is exactly the derivative of $\int_t^T h(s-t)\,ds$ with respect to t. Applying Leibnitz' rule and $h(0)=0$, this derivative equals $-\int_t^T h'(s-t)\,ds = -h(T-t)$. In sum, the drift rate of $\ln B_t^T$ becomes $r_t + h(T-t)$ according to the hypothesis.

[7] In a sense the liquidity preference hypothesis simply says that all investors prefer short bonds.

10.9 Concluding remarks

For models of the equilibrium term structure of interest rates with investor heterogeneity or more general utility functions than studied in this chapter, see, e.g., Duffie and Epstein (1992a), Wang (1996), Riedel (2000, 2004), Wachter (2006). The effects of central banks on the term structure are discussed and modeled by, e.g., Babbs and Webber (1994), Balduzzi, Bertola, and Foresi (1997), and Piazzesi (2001).

10.10 Exercises

EXERCISE 10.1 Show that if there is no arbitrage and the short rate can never go negative, then the discount function is non-increasing and all forward rates are non-negative.

EXERCISE 10.2 Show Equation (10.48).

EXERCISE 10.3 The term premium at time t for the future period $[t', T]$ is the current forward rate for that period minus the expected spot rate, i.e. $f_t^{t',T} - \mathrm{E}_t[y_{t'}^T]$. This exercise will give a link between the term premium and a state-price deflator $\zeta = (\zeta_t)$.

(a) Show that

$$B_t^T = B_t^{t'} \, \mathrm{E}_t \left[B_{t'}^T \right] + \mathrm{Cov}_t \left[\frac{\zeta_{t'}}{\zeta_t}, \frac{\zeta_T}{\zeta_{t'}} \right]$$

for any $t \leq t' \leq T$.

(b) Using the above result, show that

$$\mathrm{E}_t \left[e^{-y_{t'}^T (T-t')} \right] - e^{-f_t^{t',T}(T-t')} = -\frac{1}{B_t^{t'}} \mathrm{Cov}_t \left[\frac{\zeta_{t'}}{\zeta_t}, \frac{\zeta_T}{\zeta_{t'}} \right].$$

Using the previous result and the approximation $e^x \approx 1 + x$, show that

$$f_t^{t',T} - \mathrm{E}_t[y_{t'}^T] \approx -\frac{1}{(T-t')B_t^{t'}} \mathrm{Cov}_t \left[\frac{\zeta_{t'}}{\zeta_t}, \frac{\zeta_T}{\zeta_{t'}} \right].$$

EXERCISE 10.4 The purpose of this exercise is to show that the claim of the gross return pure expectation hypothesis is inconsistent with interest rate uncertainty. In the following we consider time points $t_0 < t_1 < t_2$.

(a) Show that if the hypothesis holds, then

$$\frac{1}{B_{t_0}^{t_1}} = \frac{1}{B_{t_0}^{t_2}} \, \mathrm{E}_{t_0} \left[B_{t_1}^{t_2} \right].$$

Hint: Compare two investment strategies over the period $[t_0, t_1]$. The first strategy is to buy at time t_0 zero-coupon bonds maturing at time t_1. The second strategy is to buy at time t_0 zero-coupon bonds maturing at time t_2 and to sell them again at time t_1.

(b) Show that if the hypothesis holds, then

$$\frac{1}{B_{t_0}^{t_2}} = \frac{1}{B_{t_0}^{t_1}} \mathrm{E}_{t_0}\left[\frac{1}{B_{t_1}^{t_2}}\right].$$

(c) Show from the two previous questions that the hypothesis implies that

$$\mathrm{E}_{t_0}\left[\frac{1}{B_{t_1}^{t_2}}\right] = \frac{1}{\mathrm{E}_{t_0}\left[B_{t_1}^{t_2}\right]}. \qquad (*)$$

(d) Show that (*) can only hold under full certainty. *Hint: Use Jensen's inequality.*

EXERCISE 10.5 Show (10.60) and (10.61).

EXERCISE 10.6 Go through the derivations in the subsection with the heading "An example" in Section 10.6.3.

EXERCISE 10.7 Constantinides (1992) develops the so-called SAINTS model of the nominal term structure of interest rates by specifying exogenously the nominal state-price deflator $\tilde{\zeta}$. In a slightly simplified version, his assumption is that

$$\tilde{\zeta}_t = ke^{-gt+(X_t-\alpha)^2},$$

where k, g, and α are constants, and $X = (X_t)$ follows the Ornstein-Uhlenbeck process

$$dX_t = -\kappa X_t\, dt + \sigma\, dz_t,$$

where κ and σ are positive constants with $\sigma^2 < \kappa$ and $z = (z_t)$ is a standard one-dimensional Brownian motion.

(a) Derive the dynamics of the nominal state-price deflator. Express the nominal short-term interest rate, $\tilde{r}_t$, and the nominal market price of risk, $\tilde{\lambda}_t$, in terms of the variable X_t.

(b) Find the dynamics of the nominal short rate.

(c) Find parameter constraints that ensure that the short rate stays positive? *Hint: The short rate is a quadratic function of X. Find the minimum value of this function.*

(d) What is the distribution of X_T given X_t?

(e) Let Y be a normally distributed random variable with mean μ and variance v^2. Show that

$$\mathrm{E}\left[e^{-\gamma Y^2}\right] = (1+2\gamma v^2)^{-1/2} \exp\left\{-\frac{\gamma\mu^2}{1+2\gamma v^2}\right\}.$$

(f) Use the results of the two previous questions to derive the time t price of a nominal zero-coupon bond with maturity T, i.e. $\tilde{B}_t^T$. It will be an exponential-quadratic function of X_t. What is the yield on this bond?

(g) Find the percentage volatility σ_t^T of the price of the zero-coupon bond maturing at T.

(h) The instantaneous expected excess rate of return on the zero-coupon bond maturing at T is often called the term premium for maturity T. Explain why the term premium is given by $\sigma_t^T \tilde{\lambda}_t$ and show that the term premium can be written as

$$4\sigma^2\alpha^2\left(1-F(T-t)\right)\left(\frac{X_t}{\alpha}-1\right)\left(\frac{X_t}{\alpha}-\frac{1-F(T-t)e^{\kappa(T-t)}}{1-F(T-t)}\right),$$

where

$$F(\tau)=\frac{1}{\frac{\sigma^2}{\kappa}+\left(1-\frac{\sigma^2}{\kappa}\right)e^{2\kappa\tau}}.$$

For which values of X_t will the term premium for maturity T be positive/negative? For a given state X_t, is it possible that the the term premium is positive for some maturities and negative for others?

EXERCISE 10.8 Assume a continuous-time economy where the state-price deflator $\zeta=(\zeta_t)$ has dynamics

$$d\zeta_t=-\zeta_t\left[r_t\,dt+\lambda\,dz_{1t}\right],$$

where $z_1=(z_{1t})$ is a (one-dimensional) standard Brownian motion, λ is a constant, and $r=(r_t)$ follows the Ornstein-Uhlenbeck process

$$dr_t=\kappa[\bar{r}-r_t]\,dt+\sigma_r\,dz_{1t}.$$

This is the Vasicek model so we know that the prices of zero-coupon bonds are given by (10.34) and the corresponding yields are given by (10.36).
Suppose you want to value a real uncertain cash flow of F_T coming at time T. Let $x_t=\mathrm{E}_t[F_T]$ and assume that

$$dx_t=x_t\left[\mu_x\,dt+\sigma_x\rho\,dz_{1t}+\sigma_x\sqrt{1-\rho^2}\,dz_{2t}\right],$$

where μ_x, σ_x, and ρ are constants, and where $z_2=(z_{2t})$ is another (one-dimensional) standard Brownian motion independent of z_1.

(a) Argue that $x=(x_t)$ must be a martingale and hence that $\mu_x=0$.

(b) Show that the time t value of the claim to the cash flow F_T is given by

$$V_t\equiv V(t,r_t,x_t)=x_te^{-A(T-t)-B(T-t)r_t}, \qquad (*)$$

where $B(\tau)=b(\tau)$ and

$$A(\tau)=a(\tau)+\rho\lambda\sigma_x\tau+\frac{\rho\sigma_x\sigma_r}{\kappa}\left(\tau-b(\tau)\right).$$

(c) Write the dynamics of $V = (V_t)$ as $dV_t = V_t[\mu_t^V\,dt + \sigma_{1t}^V\,dz_{1t} + \sigma_{2t}^V\,dz_{2t}]$. Use (*) to identify μ_t^V, σ_{1t}^V, and σ_{2t}^V. Verify that $\mu_t^V = r_t + \left(\boldsymbol{\sigma}_t^V\right)^\top \boldsymbol{\lambda}_t$, where $\boldsymbol{\sigma}^V = (\sigma_1^V, \sigma_2^V)^\top$ and $\boldsymbol{\lambda}$ is the market price of risk vector (the market price of risk associated with z_2 is zero! Why?).

(d) Define the risk-adjusted discount rate R_t for the cash flow by the relation $V_t = \mathrm{E}_t[F_T]e^{-R_t[T-t]}$. What is the difference between R_t and y_t^T? How does this difference depend on the cash flow payment date T?

EXERCISE 10.9 Consider an economy with complete financial markets and a representative agent with CRRA utility, $u(C) = \frac{C^{1-\gamma}}{1-\gamma}$, where $\gamma > 0$, and a time preference rate of δ. The aggregate consumption level C is assumed to follow the stochastic process

$$dc_t = c_t\left[\left(a_1X_t^2 + a_2X_t + a_3\right)\,dt + \sigma_c\,dz_t\right],$$

where $z = (z_t)$ is a standard one-dimensional Brownian motion under the real-life probability measure $\mathbb{P}$ and where a_1, a_2, a_3, σ_c are constants with $\sigma_c > 0$. Furthermore, $X = (X_t)$ is a stochastic process with dynamics

$$dX_t = -\kappa X_t\,dt + dz_t,$$

where κ is a positive constant.

(a) Show that the short-term interest rate is of the form $r_t = d_1X_t^2 + d_2X_t + d_3$ and determine the constants d_1, d_2, d_3.

(b) Find a parameter condition under which the short-term interest rate is always non-negative.

(c) Write up the dynamics of r_t.

(d) What is the market price of risk in this economy?

Suppose that the above applies to the real economy and that money has no effects on the real economy. The consumer price index F_t is supposed to have dynamics

$$dF_t = F_t\left[\mu_{\varphi t}\,dt + \rho_{CF}\sigma_{\varphi t}\,dz_t + \sqrt{1-\rho_{CF}^2}\sigma_{\varphi t}\,d\hat{z}_t\right],$$

where ρ_{CF} is a constant correlation coefficient and $\hat{z} = (\hat{z}_t)$ is another standard Brownian motion independent of z. Assume that $\mu_{\varphi t}$ and $\sigma_{\varphi t}$ are on the form

$$\mu_{\varphi t} = b_1X_t^2 + b_2X_t + b_3, \qquad \sigma_{\varphi t} = kX_t.$$

(e) Write up an expression for the nominal short-term interest rate, $\tilde{r}_t$.

Assume in the rest of the problem that $\gamma a_1 + b_1 = k^2$.

(f) Show that the nominal short rate $\tilde{r}_t$ is affine in X_t and express X_t as an affine function of $\tilde{r}_t$.

(g) Compute the nominal market price of risk $\tilde{\lambda}_t$.

(h) Determine the dynamics of the nominal short rate. The drift and volatility should be expressed in terms of $\tilde{r}_t$, not X_t.

Chapter 11

Risk-adjusted probabilities

11.1 Introduction

Chapter 4 illustrated how the general pricing mechanism in a financial market can be represented by a state-price deflator. However, the state-price deflator is not the only way to represent the pricing mechanism of a financial market. As indicated in a one-period framework in Section 4.5.1, one can equivalently represent the pricing mechanism by a risk-neutral probability measure and the risk-free return. This chapter explores and generalizes this idea and also outlines some applications of this alternative representation. The risk-neutral pricing technique is the standard approach in the valuation of derivative securities. The next chapter focuses on derivatives and will illustrate the use of risk-neutral valuation for derivative pricing.

Apparently, the idea of risk-neutral valuation stems from Arrow (1970) and Drèze (1971) and was further explored by Harrison and Kreps (1979).

The rest of this chapter is organized in the following way. Section 11.2 outlines how a general change of the probability measure is formalized. The risk-neutral probability measure is defined and studied in Section 11.3, while the so-called forward risk-adjusted probability measures are introduced in Section 11.4. Section 11.5 shows that an appropriate risk-adjusted probability measure can be defined for any given asset or trading strategy with a positive value. Section 11.6 demonstrates that the risk-adjusted probability measure associated with the so-called growth-optimal trading strategy is identical to the real-world probability measure.

11.2 Change of probability measure

Any financial model with uncertainty formally builds on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the state space (the set of possible realizations of all relevant uncertain objects), $\mathcal{F}$ is the set of events that can be assigned a probability, and $\mathbb{P}$ is a probability measure assigning probabilities to events. It is implicitly understood that $\mathbb{P}$ gives the true or real-world probabilities of events. The consumption and investment decisions of individuals will depend on the probabilities they associate with different events and, hence, the equilibrium asset prices will reflect those probabilities. However, as we will see in the following sections, for some purposes it will be interesting to consider other probability measures on the same set of events. We will use the term real-world probability

measure for $\mathbb{P}$ but in the literature $\mathbb{P}$ is also referred to as the true, the physical, or the empirical probability measure.

A word on notation. Whenever the expectation operator is written without a superscript, it means the expectation using the probability measure $\mathbb{P}$. The expectation under a different probability measure $\mathbb{Q}$ will be denoted by $\mathrm{E}^{\mathbb{Q}}$. Similarly for variances and covariances and for conditional moments.

The alternative probability measures we will consider will be equivalent to $\mathbb{P}$. Two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same set of events $\mathcal{F}$ are said to be **equivalent probability measures** if they assign probability zero to exactly the same events, i.e.

$$\mathbb{P}(F) = 0 \quad \Leftrightarrow \quad \mathbb{Q}(F) = 0.$$

The link between two equivalent probability measures $\mathbb{P}$ and $\mathbb{Q}$ can be represented by a random variable, which is typically denoted by $\frac{d\mathbb{Q}}{d\mathbb{P}}$ and referred to as the **Radon-Nikodym derivative** of $\mathbb{Q}$ with respect to $\mathbb{P}$. For any state $\omega \in \Omega$, the value of $\frac{d\mathbb{Q}}{d\mathbb{P}}$ shows what the $\mathbb{P}$-probability of ω should be multiplied by in order to get the $\mathbb{Q}$-probability of ω. In the special case of a finite state space $\Omega = \{1, 2, \ldots, S\}$, the probability measures $\mathbb{P}$ and $\mathbb{Q}$ are defined by the probabilities p_ω and q_ω, respectively, of the individual states $\omega = 1, 2, \ldots, S$. The Radon-Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$ is then captured by the S possible realizations

$$\frac{d\mathbb{Q}}{d\mathbb{P}}(\omega) = \frac{q_\omega}{p_\omega}, \quad \omega = 1, \ldots, S.$$

The Radon-Nikodym derivative $\frac{d\mathbb{Q}}{d\mathbb{P}}$ must be strictly positive on all events having a non-zero $\mathbb{P}$-probability. Furthermore, to ensure that the $\mathbb{Q}$-probabilities sum up to one, we must have $\mathrm{E}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] = 1$. For example, with a finite state space

$$\mathrm{E}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] = \sum_{\omega=1}^{S} p_\omega \frac{d\mathbb{Q}}{d\mathbb{P}}(\omega) = \sum_{\omega=1}^{S} p_\omega \frac{q_\omega}{p_\omega} = \sum_{\omega=1}^{S} q_\omega = 1.$$

The expected value under the measure $\mathbb{Q}$ of a random variable X is given by

$$\mathrm{E}^{\mathbb{Q}}[X] = \mathrm{E}\left[\frac{d\mathbb{Q}}{d\mathbb{P}} X\right]. \tag{11.1}$$

Again, this is easily demonstrated with a finite state space:

$$\mathrm{E}^{\mathbb{Q}}[X] = \sum_{\omega=1}^{S} q_\omega X(\omega) = \sum_{\omega=1}^{S} p_\omega \frac{q_\omega}{p_\omega} X(\omega) = \sum_{\omega=1}^{S} p_\omega \frac{d\mathbb{Q}}{d\mathbb{P}}(\omega) X(\omega) = \mathrm{E}\left[\frac{d\mathbb{Q}}{d\mathbb{P}} X\right].$$

In a multi-period model where all the uncertainty is resolved at time T, the Radon-Nikodym derivative $\frac{d\mathbb{Q}}{d\mathbb{P}}$ will be realized at time T, but usually not known before time T. Define the stochastic process $\xi = (\xi_t)_{t \in \mathcal{T}}$ by

$$\xi_t = \mathrm{E}_t\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right].$$

In particular, $\xi_T = \frac{d\mathbb{Q}}{d\mathbb{P}}$. The process ξ is called the change-of-measure process or the likelihood ratio process. Note that the process ξ is a $\mathbb{P}$-martingale since, for any $t < t' \leq T$, we have

$$\mathrm{E}_t\left[\xi_{t'}\right] = \mathrm{E}_t\left[\mathrm{E}_{t'}\left[\xi_T\right]\right] = \mathrm{E}_t\left[\xi_T\right] = \xi_t.$$

Here the first and the third equalities follow from the definition of ξ. The second equality follows from the Law of Iterated Expectations, Theorem 2.1.

In multi-period models we often work with conditional probabilities and the following result turns out to be very useful. Let $X = (X_t)_{t\in\mathcal{T}}$ be any stochastic process. Then we have

$$\mathrm{E}_t^{\mathbb{Q}}[X_{t'}] = \frac{\mathrm{E}_t[\xi_{t'}X_{t'}]}{\mathrm{E}_t[\xi_{t'}]} = \mathrm{E}_t\left[\frac{\xi_{t'}}{\xi_t}X_{t'}\right]. \tag{11.2}$$

This is called Bayes' Formula. For a proof, see Björk (2004, Prop. B.41).

A change of the probability measure can be handled very elegantly in continuous-time models where the underlying uncertainty is represented by a standard Brownian motion $z = (z_t)_{t\in[0,T]}$ (under the real-world probability measure $\mathbb{P}$), which is the case in all the continuous-time models considered in this book. Let $\lambda = (\lambda_t)_{t\in[0,T]}$ be any adapted and sufficiently well-behaved stochastic process.[1] Here, z and λ must have the same dimension. For notational simplicity, we assume in the following that they are one-dimensional, but the results generalize naturally to the multi-dimensional case. We can generate an equivalent probability measure $\mathbb{Q}^\lambda$ in the following way. Define the process $\xi^\lambda = (\xi_t^\lambda)_{t\in[0,T]}$ by

$$\xi_t^\lambda = \exp\left\{-\int_0^t \lambda_s\,dz_s - \frac{1}{2}\int_0^t \lambda_s^2\,ds\right\}. \tag{11.3}$$

Then $\xi_0^\lambda = 1$, ξ^λ is strictly positive, and an application of Itô's Lemma shows that $d\xi_t^\lambda = -\xi_t^\lambda\lambda_t\,dz_t$ so that ξ^λ is a $\mathbb{P}$-martingale (see Exercise 2.4) and $\mathrm{E}[\xi_T^\lambda] = \xi_0^\lambda = 1$. Consequently, an equivalent probability measure $\mathbb{Q}^\lambda$ can be defined by the Radon-Nikodym derivative

$$\frac{d\mathbb{Q}^\lambda}{d\mathbb{P}} = \xi_T^\lambda = \exp\left\{-\int_0^T \lambda_s\,dz_s - \frac{1}{2}\int_0^T \lambda_s^2\,ds\right\}.$$

From (11.2), we get that

$$\mathrm{E}_t^{\mathbb{Q}^\lambda}[X_{t'}] = \mathrm{E}_t\left[\frac{\xi_{t'}^\lambda}{\xi_t^\lambda}X_{t'}\right] = \mathrm{E}_t\left[X_{t'}\exp\left\{-\int_t^{t'} \lambda_s\,dz_s - \frac{1}{2}\int_t^{t'} \lambda_s^2\,ds\right\}\right] \tag{11.4}$$

for any stochastic process $X = (X_t)_{t\in[0,T]}$. A central result is Girsanov's Theorem:

Theorem 11.1 (Girsanov) *The process $z^\lambda = (z_t^\lambda)_{t\in[0,T]}$ defined by*

$$z_t^\lambda = z_t + \int_0^t \lambda_s\,ds, \quad 0 \le t \le T, \tag{11.5}$$

is a standard Brownian motion under the probability measure $\mathbb{Q}^\lambda$. In differential notation,

$$dz_t^\lambda = dz_t + \lambda_t\,dt.$$

This theorem has the attractive consequence that the effects on a stochastic process of changing the probability measure from $\mathbb{P}$ to some $\mathbb{Q}^\lambda$ are captured by a simple adjustment of the drift. If $X = (X_t)$ is an Itô-process with dynamics

$$dX_t = \mu_t\,dt + \sigma_t\,dz_t,$$

[1] Basically, λ must be square-integrable in the sense that $\int_0^T \lambda_t^2\,dt$ is finite with probability 1 and that λ satisfies Novikov's condition, i.e. the expectation $\mathrm{E}\left[\exp\left\{\frac{1}{2}\int_0^T \lambda_t^2\,dt\right\}\right]$ is finite.

then

$$dX_t = \mu_t\,dt + \sigma_t\left(dz_t^\lambda - \lambda_t\,dt\right) = (\mu_t - \sigma_t\lambda_t)\,dt + \sigma_t\,dz_t^\lambda. \tag{11.6}$$

Hence, $\mu - \sigma\lambda$ is the drift under the probability measure $\mathbb{Q}^\lambda$, which is different from the drift under the original measure $\mathbb{P}$ unless σ or λ are identically equal to zero. In contrast, the volatility remains the same as under the original measure. We will say that the equation (11.6) is the $\mathbb{Q}^\lambda$-dynamics of the process X.

In many financial models, the relevant change of measure is such that the distribution under $\mathbb{Q}^\lambda$ of the future value of the central processes is of the same class as under the original $\mathbb{P}$ measure, but with different moments. However, in general, a shift of probability measure may change not only some or all moments of future values, but also the distributional class.

11.3 Risk-neutral probabilities

11.3.1 Definition

A risk-neutral probability measure for a given financial market can only be defined if the investors at any point in time considered in the model and for any state can trade in an asset which provides a risk-free return until the next point in time where the investors can rebalance their portfolios. In a one-period economy, this is simply a one-period risk-free asset. As before, R^f denotes the gross return on that asset. In a discrete-time economy with trading at $t = 0, 1, 2, \ldots, T-1$, the assumptions is that investors can roll over in one-period risk-free investments. Investing one unit in a one-period risk-free investment at time t will give you R_t^f at time $t+1$. Reinvesting that in a risk-free manner over the next period will give you $R_{t,t+2}^f = R_t^f R_{t+1}^f$ at time $t+2$. Continuing that procedure, you end up with

$$R_{t,t+n}^f = R_t^f R_{t+1}^f R_{t+2}^f \ldots R_{t+n-1}^f = \prod_{m=0}^{n-1} R_{t+m}^f$$

at time $t+n$. Note that, in general, this return is not known before time $t+n-1$ and in particular not at time t where the investment strategy is initiated. An investment with a truly risk-free return between time t and $t+n$ is a zero-coupon bond maturing at time $t+n$. If B_t^{t+n} denotes the price of this bond at time t and the face value of the bond is normalized at 1, the gross risk-free return between t and $t+n$ is $1/B_t^{t+n}$.

In a continuous-time economy we can think of the limit of the above roll-over strategy. If r_t^f denotes the continuously compounded risk-free net rate of return at time t (the interest rate over the instant following time t), an investment of 1 in this roll-over strategy at time t will give you

$$R_{t,t'}^f = \exp\left\{\int_t^{t'} r_u^f\,du\right\}$$

at time t'.

Whether the model is formulated in discrete or in continuous time we refer to the roll-over strategy in short risk-free investments as the **bank account** and refer to $R_{t,s}^f$ as the gross return on the bank account between time t and time $s > t$. In the one-period model, the bank account is simply a one-period risk-free asset.

We can now give a unified definition of a risk-neutral probability measure. A probability measure $\mathbb{Q}$ is called a **risk-neutral probability measure** for a given financial market in which a bank account is traded if the following conditions are satisfied:

(i) $\mathbb{P}$ and $\mathbb{Q}$ are equivalent;

(ii) the Radon-Nikodym derivative $\frac{d\mathbb{Q}}{d\mathbb{P}}$ has finite variance;

(iii) the price of a future dividend equals the $\mathbb{Q}$-expectation of the ratio of the dividend to the gross return on the bank account between the pricing date and the dividend payment date.

When $\mathbb{Q}$ is a risk-neutral probability measure, we will refer to the $\mathbb{Q}$-expectation as the risk-neutral expectation.

The pricing condition (iii) is a bit vague at this point. In a one-period framework, it means that

$$P_i = \mathrm{E}^{\mathbb{Q}}\left[(R^f)^{-1} D_i\right] = (R^f)^{-1}\,\mathrm{E}^{\mathbb{Q}}\left[D_i\right], \tag{11.7}$$

and, consequently, $\mathrm{E}[R_i] = R^f$. In a discrete-time model, the pricing condition (iii) is

$$P_{it} = \mathrm{E}_t^{\mathbb{Q}}\left[\sum_{s=t+1}^{T} \frac{D_{is}}{R^f_{t,s}}\right], \tag{11.8}$$

which is equivalent to

$$P_{it} = \mathrm{E}_t^{\mathbb{Q}}\left[\frac{P_{it'}}{R^f_{t,t'}} + \sum_{s=t+1}^{t'} \frac{D_{is}}{R^f_{t,s}}\right], \quad t < t' \le T, \tag{11.9}$$

cf. Exercise 11.1. In particular, for $t' = t+1$ this reduces to

$$P_{it} = \frac{1}{R^f_t}\,\mathrm{E}_t^{\mathbb{Q}}\left[P_{i,t+1} + D_{i,t+1}\right] \tag{11.10}$$

so that $\mathrm{E}_t[R_{i,t+1}] = R^f_t$. In the continuous-time framework, the pricing condition (iii) is interpreted as

$$\begin{aligned} P_{it} &= \mathrm{E}_t^{\mathbb{Q}}\left[\int_t^T (R^f_{t,s})^{-1}\delta_{is}P_{is}\,ds + (R^f_{t,T})^{-1}D_{iT}\right] \\ &= \mathrm{E}_t^{\mathbb{Q}}\left[(R^f_{t,T})^{-1} e^{\int_t^T \delta_{is}\,ds} D_{iT}\right] = \mathrm{E}_t^{\mathbb{Q}}\left[e^{-\int_t^T (r^f_s - \delta_{is})\,ds} D_{iT}\right], \end{aligned} \tag{11.11}$$

from which the relation

$$P_{it} = \mathrm{E}_t^{\mathbb{Q}}\left[\int_t^{t'} (R^f_{t,s})^{-1}\delta_{is}P_{is}\,ds + (R^f_{t,t'})^{-1}P_{i,t'}\right] = \mathrm{E}_t^{\mathbb{Q}}\left[e^{-\int_t^{t'} (r^f_s - \delta_{is})\,ds} P_{i,t'}\right] \tag{11.12}$$

follows. This implies that

$$dP_{it} = P_{it}\left[(r^f_t - \delta_{it})\,dt + \boldsymbol{\sigma}_{it}^\top\,d\boldsymbol{z}_t^{\mathbb{Q}}\right] \tag{11.13}$$

so that the total instantaneous rate of return has a risk-neutral expectation equal to the risk-free rate. Therefore, in all of the modeling frameworks, the risk-neutral expected return on any asset over the next period equals the risk-free return over that period. The above considerations also hold for all trading strategies (as always, in the continuous-time framework some "wild" trading strategies must be ruled out).

We can see from the above equations that given the stochastic process for the risk-free return and a risk-neutral probability measure, we can price any dividend process. Therefore the risk-free return process and a risk-neutral probability measure jointly capture the market-wide pricing mechanism of the financial market.

A risk-neutral probability measure is sometimes called an **equivalent martingale measure**. Of course, the word *equivalent* refers to the equivalence of the risk-neutral probability measure and the real-world probability measure. The word *martingale* is used here since the risk-free discounted gains process of any asset will be a $\mathbb{Q}$-martingale. The risk-free discounted gains process of asset i is denoted by $\bar{G}_i = (\bar{G}_{it})_{t\in\mathcal{T}}$. In the discrete-time setting, it is defined as

$$\bar{G}_{it} = \frac{P_{it}}{R^f_{0,t}} + \sum_{s=1}^{t} \frac{D_{is}}{R^f_{0,s}}$$

Since $R^f_{0,s} = R^f_{0,t} R^f_{t,s}$ for all $t < s$, the pricing condition (11.9) can be rewritten as

$$\frac{P_{it}}{R^f_{0,t}} = \mathrm{E}^{\mathbb{Q}}_t \left[\frac{P_{it'}}{R^f_{0,t'}} + \sum_{s=t+1}^{t'} \frac{D_{is}}{R^f_{0,s}} \right], \quad t < t' \leq T,$$

which is equivalent to

$$\frac{P_{it}}{R^f_{0,t}} + \sum_{s=1}^{t} \frac{D_{is}}{R^f_{0,s}} = \mathrm{E}^{\mathbb{Q}}_t \left[\frac{P_{it'}}{R^f_{0,t'}} + \sum_{s=1}^{t'} \frac{D_{is}}{R^f_{0,s}} \right], \quad t < t' \leq T,$$

i.e. $\bar{G}_{it} = \mathrm{E}^{\mathbb{Q}}_t[\bar{G}_{i,t'}]$ so that $\bar{G}_i$ indeed is a $\mathbb{Q}$-martingale. In the continuous-time setting, the discounted gains process of asset i is defined as

$$\bar{G}_{it} = \frac{P_{it}}{R^f_{0,t}} + \int_0^t \frac{\delta_{is} P_{is}}{R^f_{0,s}}\, ds,$$

and again it can be shown that the pricing condition (11.12) is equivalent to $\bar{G}_i$ being a $\mathbb{Q}$-martingale.

11.3.2 Relation to state-price deflators

Since we can represent the general pricing mechanism of a financial market either by a state-price deflator or by a risk-neutral probability measure and the risk-free return process, it should come as no surprise that there is a close relation between these quantities.

First, consider a one-period economy with a risk-free asset. Given a state-price deflator ζ, the risk-free rate is $R^f = 1/\mathrm{E}[\zeta]$ and we can define the random variable

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = R^f \zeta,$$

which is a strictly positive random variable with

$$\mathrm{E}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] = R^f \,\mathrm{E}[\zeta] = 1.$$

Therefore, $\frac{d\mathbb{Q}}{d\mathbb{P}}$ defines a probability measure $\mathbb{Q}$, which is equivalent to $\mathbb{P}$. Since a state-price deflator has finite variance and R^f is a constant, $\frac{d\mathbb{Q}}{d\mathbb{P}}$ has finite variance. Furthermore, from (11.1) we get

$$\mathrm{E}^{\mathbb{Q}}\left[\frac{D_i}{R^f}\right] = \mathrm{E}\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\frac{D_i}{R^f}\right] = \mathrm{E}[\zeta D_i] = P_i.$$

Hence $\mathbb{Q}$ is indeed a risk-neutral probability measure. Conversely, if $\mathbb{Q}$ is a risk-neutral probability measure, a state-price deflator can be defined by

$$\zeta = (R^f)^{-1} \frac{d\mathbb{Q}}{d\mathbb{P}}.$$

Changing the probability measure is a reallocation of probability mass over the states. We can see that the risk-neutral measure allocates a higher probability to states ω for which $\zeta_\omega > (R^f)^{-1} = \mathrm{E}[\zeta]$, i.e. if the value of the state-price deflator for state ω is higher than average.

Example 11.1 Consider the same one-period economy as in the Examples 3.1 and 4.2. The real-world probabilities of the three states are $p_1 = 0.5$, $p_2 = p_3 = 0.25$, respectively. The state-price deflator is given by $\zeta_1 = 0.6$, $\zeta_2 = 0.8$, and $\zeta_3 = 1.2$. Since

$$\mathrm{E}[\zeta] = 0.5 \cdot 0.6 + 0.25 \cdot 0.8 + 0.25 \cdot 1.2 = 0.8,$$

the gross risk-free return is $R^f = 1/\mathrm{E}[\zeta] = 1.25$ corresponding to a 25% risk-free net rate of return. It follows that the risk-neutral probabilities are

$$q_1 = R^f \zeta_1 p_1 = 0.375, \quad q_2 = R^f \zeta_2 p_2 = 0.25, \quad q_3 = R^f \zeta_3 p_3 = 0.375.$$

The risk-neutral measure allocates a larger probability to state 3, the same probability to state 2, and a lower probability to state 1 than the real-world measure. □

In a multi-period model where all uncertainty is resolved at time T, the relation linking the state-price deflator $\zeta = (\zeta_t)_{t\in\mathcal{T}}$ and the risk-neutral probability measure is

$$\zeta_t = \frac{1}{R^f_{0,t}} \mathrm{E}_t \left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] = \frac{\xi_t}{R^f_{0,t}}, \tag{11.14}$$

where $\xi_t = \mathrm{E}_t^{\mathbb{P}}[\frac{d\mathbb{Q}}{d\mathbb{P}}]$ as before. Given a risk-neutral probability measure $\mathbb{Q}$ and the risk-free return process $(R^f_{0,t})_{t\in\mathcal{T}}$, this equation defines the state-price deflator. Conversely, given a state-price deflator ζ, the risk-free return process is

$$R^f_{0,t} = R^f_0 R^f_1 \dots R^f_{t-1} = \left(\mathrm{E}\left[\frac{\zeta_1}{\zeta_0}\right] \mathrm{E}_1\left[\frac{\zeta_2}{\zeta_1}\right] \dots \mathrm{E}_t\left[\frac{\zeta_t}{\zeta_{t-1}}\right]\right)^{-1}$$

and

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = R^f_{0,T} \zeta_T \tag{11.15}$$

defines a risk-neutral probability measure $\mathbb{Q}$.

Let us consider a discrete-time framework and verify that the pricing condition (11.9) in the definition of a risk-neutral probability measure is satisfied when ζ is a state-price deflator and $\mathbb{Q}$ is defined through the Radon-Nikodym derivative (11.15). First note that when ζ is a state-price deflator,

$$1 = \mathrm{E}_t\left[\frac{\zeta_T}{\zeta_t} R^f_{t,T}\right]$$

and hence

$$\xi_t = \mathrm{E}_t\left[\frac{d\mathbb{Q}}{d\mathbb{P}}\right] = \mathrm{E}_t\left[\zeta_T R^f_{0,T}\right] = \zeta_t R^f_{0,t} \mathrm{E}_t\left[\frac{\zeta_T}{\zeta_t} R^f_{t,T}\right] = \zeta_t R^f_{0,t},$$

so that

$$\xi_t = \zeta_t R^f_{0,t}, \quad t = 0, 1, \dots, T,$$

and, thus,

$$\frac{\xi_s}{\xi_t} = \frac{\zeta_s}{\zeta_t} R^f_{t,s}, \quad t < s \leq T. \tag{11.16}$$

We now have that

$$\begin{aligned}
\mathrm{E}^{\mathbb{Q}}_t \left[\frac{P_{it'}}{R^f_{t,t'}} + \sum_{s=t+1}^{t'} \frac{D_{is}}{R^f_{t,s}} \right] &= \mathrm{E}^{\mathbb{Q}}_t \left[\frac{P_{it'}}{R^f_{t,t'}} \right] + \sum_{s=t+1}^{t'} \mathrm{E}^{\mathbb{Q}}_t \left[\frac{D_{is}}{R^f_{t,s}} \right] \\
&= \mathrm{E}_t \left[\frac{\xi_{t'}}{\xi_t} \frac{P_{it'}}{R^f_{t,t'}} \right] + \sum_{s=t+1}^{t'} \mathrm{E}_t \left[\frac{\xi_s}{\xi_t} \frac{D_{is}}{R^f_{t,s}} \right] \\
&= \mathrm{E}_t \left[\frac{\xi_{t'}}{\xi_t} \frac{P_{it'}}{R^f_{t,t'}} + \sum_{s=t+1}^{t'} \frac{\xi_s}{\xi_t} \frac{D_{is}}{R^f_{t,s}} \right] \\
&= \mathrm{E}_t \left[\frac{\zeta_{t'}}{\zeta_t} P_{it'} + \sum_{s=t+1}^{t'} \frac{\zeta_s}{\zeta_t} D_{is} \right] \\
&= P_{it},
\end{aligned}$$

as was to be shown. Here the second equality is due to the relation (11.2), the fourth equality comes from inserting (11.16), and the final equality holds since ζ is a state-price deflator. Taking the above steps in the reverse order will show that if the pricing condition in the definition of a risk-neutral probability measure is satisfied, then the pricing condition in the definition of a state-price deflator is satisfied when ζ is defined as in (11.14). An analogous procedure works in the continuous-time setting.

Combining the relation between risk-neutral measures and state-price deflators with the results on the existence and uniqueness of state-price deflators derived in Section 4.3, we can make the following conclusions:

Theorem 11.2 *Assume that a bank account is traded. Prices admit no arbitrage if and only if a risk-neutral probability measure exists. An arbitrage-free market is complete if and only if there is a unique risk-neutral probability measure.*

In the continuous-time framework some technical conditions have to be added or the definition of arbitrage must be slightly adjusted. In that framework, we know from Chapter 4 that a state-price deflator is of the form

$$\zeta_t = \exp\left\{ -\int_0^t r^f_s \, ds - \frac{1}{2} \int_0^t \|\boldsymbol{\lambda}_s\|^2 \, ds - \int_0^t \boldsymbol{\lambda}_s^\top \, d\boldsymbol{z}_s \right\}, \tag{4.43}$$

where $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_t)$ is a market price of risk process so that

$$\boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r^f_t \mathbf{1} = \underline{\underline{\sigma}}_t \boldsymbol{\lambda}_t.$$

The corresponding risk-neutral probability measure is defined by

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = R^f_{0,T} \zeta_T = e^{\int_0^T r^f_s \, ds} \zeta_T = \exp\left\{ -\frac{1}{2} \int_0^T \|\boldsymbol{\lambda}_s\|^2 \, ds - \int_0^T (\boldsymbol{\lambda}_s)^\top \, d\boldsymbol{z}_s \right\} \tag{11.17}$$

and

$$\xi_t = \mathrm{E}_t \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \right] = \exp\left\{ -\frac{1}{2} \int_0^t \|\boldsymbol{\lambda}_s\|^2 \, ds - \int_0^t (\boldsymbol{\lambda}_s)^\top \, d\boldsymbol{z}_s \right\}. \tag{11.18}$$

It follows by the Girsanov Theorem 11.1 that the process $\boldsymbol{z}^{\mathbb{Q}} = (\boldsymbol{z}^{\mathbb{Q}})_{t\in[0,T]}$ defined by $\boldsymbol{z}_0^{\mathbb{Q}} = \mathbf{0}$ and

$$d\boldsymbol{z}_t^{\mathbb{Q}} = d\boldsymbol{z}_t + \boldsymbol{\lambda}_t\, dt \tag{11.19}$$

is a standard Brownian motion under the risk-neutral probability measure. We can then transform the dynamics of any process $X = (X_t)_{t\in[0,T]}$ as follows:

$$\begin{aligned} dX_t &= \mu_{Xt}\, dt + \boldsymbol{\sigma}_{Xt}^\top\, d\boldsymbol{z}_t \\ &= \mu_{Xt}\, dt + \boldsymbol{\sigma}_{Xt}^\top \left(d\boldsymbol{z}_t^{\mathbb{Q}} - \boldsymbol{\lambda}_t\, dt\right) \\ &= (\mu_{Xt} - \boldsymbol{\sigma}_{Xt}^\top \boldsymbol{\lambda}_t)\; dt + \boldsymbol{\sigma}_{Xt}^\top d\boldsymbol{z}_t^{\mathbb{Q}}. \end{aligned}$$

The instantaneous sensitivity is unchanged, but the product of the sensitivity vector and the market price of risk is subtracted from the drift. In particular, for a price process we get

$$\begin{aligned} dP_{it} &= P_{it}\left[\mu_{it}\, dt + \boldsymbol{\sigma}_{it}^\top\, d\boldsymbol{z}_t\right] \\ &= P_{it}\left[(\mu_{it} - \boldsymbol{\sigma}_{it}^\top \boldsymbol{\lambda}_t)\; dt + \boldsymbol{\sigma}_{it}^\top\, d\boldsymbol{z}_t^{\mathbb{Q}}\right] \\ &= P_{it}\left[\left(r_t^f - \delta_{it}\right)\; dt + \boldsymbol{\sigma}_{it}^\top\, d\boldsymbol{z}_t^{\mathbb{Q}}\right], \end{aligned} \tag{11.20}$$

where the last equality follows from the definition of a market price of risk. Again we see that the risk-neutral expectation of the total instantaneous rate of return is identical to the risk-free interest rate.

11.3.3 Valuation with risk-neutral probabilities

From the pricing condition in the definition of a risk-neutral probability measure, it is clear that the valuation of an asset requires knowledge of the joint risk-neutral probability distribution of the risk-free discount factor $(R_{t,s}^f)^{-1}$ and the asset dividend $D_{i,s}$. More precisely, we have to know the covariance under the risk-neutral probability measure of the two variables. For example, in the discrete-time setting, (11.8) implies that

$$P_{it} = \sum_{s=t+1}^{T} \left(\mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,s}^f\right)^{-1}\right] \mathrm{E}_t^{\mathbb{Q}}\left[D_{is}\right] + \mathrm{Cov}_t^{\mathbb{Q}}\left[\left(R_{t,s}^f\right)^{-1}, D_{is}\right]\right).$$

Note that $\mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,s}^f\right)^{-1}\right] = B_t^s$, the time t price of a zero-coupon bond maturing with a unit payment at time s. Therefore, we can rewrite the above equation as

$$P_{it} = \sum_{s=t+1}^{T} B_t^s \left(\mathrm{E}_t^{\mathbb{Q}}\left[D_{is}\right] + \frac{\mathrm{Cov}_t^{\mathbb{Q}}\left[\left(R_{t,s}^f\right)^{-1}, D_{is}\right]}{B_t^s}\right).$$

Example 11.2 Consider the two-period economy illustrated in Figures 2.1 and 2.2 and also studied in Exercise 4.8. There are six states. The real-world state probabilities and the assumed values of the state-price deflator at time 1 and time 2 are listed in left-most columns of Table 11.1. Figure 11.1 illustrates the economy as a two-period tree. Each state corresponds to a path through the tree. Two numbers are written along each branch. The left-most number is the value of the next-period deflator along that branch (ζ_1 over the first period and ζ_2/ζ_1 over the second period).

ω	p	ζ_1	ζ_2/ζ_1	ζ_2	R_1^f	$R_{0,2}^f$	$\frac{d\mathbb{Q}}{d\mathbb{P}}$	q
1	0.24	1.2	1	1.2	1.0714	1.1398	1.3678	0.3283
2	0.06	1.2	0.6667	0.8	1.0714	1.1398	0.9119	0.0547
3	0.04	1	1	1	1.0989	1.1690	1.1690	0.0468
4	0.16	1	0.9	0.9	1.0989	1.1690	1.0521	0.1683
5	0.2	1	0.9	0.9	1.0989	1.1690	1.0521	0.2104
6	0.3	0.6	0.9	0.54	1.1111	1.1820	0.6383	0.1915

Table 11.1: $\mathbb{P}$, $\mathbb{Q}$, R^f, and ζ in Example 11.2.

The right-most number is the conditional real-world probability of that branch. The conditional probabilities can be computed from the state probabilities, e.g. the conditional probability for the upward branch leaving the upper node at time 1 is the probability that state 1 is realized given that it is known that the true state is either 1 or 2, i.e. $0.24/(0.24 + 0.06) = 0.8$.

Before the risk-neutral probabilities can be computed, we have to find the gross return $R_{0,2}^f = R_0^f R_1^f$ on the bank account. We can identify this from the state-price deflator and the real-world probabilities. Over the first period the risk-free gross return is

$$R_0^f = \frac{1}{\mathrm{E}[\zeta_1]} = \frac{1}{0.3 \cdot 1.2 + 0.4 \cdot 1 + 0.3 \cdot 0.6} = \frac{1}{0.94} \approx 1.0638.$$

Over the second period the risk-free gross return depends on the information at time 1. In the upper node at time 1 the one-period gross risk-free return is

$$R_1^f = \frac{1}{\mathrm{E}_1[\zeta_2/\zeta_1]} = \frac{1}{0.8 \cdot 1 + 0.2 \cdot 0.6667} = \frac{1}{0.9333} \approx 1.0714.$$

Similarly, $R_1^f \approx 1.0989$ in the middle node and $R_1^f \approx 1.1111$ in the lower node at time 1. Now the risk-neutral probabilities can be computed as shown in the right-most part of Table 11.1.

Given the risk-neutral probabilities of each state, we can compute the conditional risk-neutral probabilities of transitions from one point in time to the next, exactly as for the real-world probabilities. Together with the one-period risk-free returns, the conditional risk-neutral probabilities contain all the necessary information to price a given dividend process by backwards recursions through the tree. This information is illustrated in Figure 11.2. The conditional risk-neutral probabilities can also be computed directly from the conditional real-world probabilities and the risk-free return and the state-price deflator for that transition. For example, the conditional risk-neutral probability of the upper branch leaving the upper node at time 1 equals $1.0714 \cdot 1 \cdot 0.8 \approx 0.8571$.

Let us compute the price process of an asset with the dividends written in the circles in Figure 11.3 (this is asset 2 from Exercise 4.8). The ex-dividend price in the upper node at time 1 is simply

$$P_1^u = \frac{1}{1.0714}\left(0.8571 \cdot 2 + 0.1429 \cdot 1\right) \approx 1.7333.$$

Similarly, the time 1 prices in the middle and lower node are $P_1^m = 2.27$ and $P_1^l = 2.7$, respectively. The time 0 price is then computed as

$$P_0 = \frac{1}{1.0638}\left(0.3830 \cdot [1.7333 + 2] + 0.4255 \cdot [2.27 + 3] + 0.1915 \cdot [2.7 + 4]\right) = 4.658.$$

□

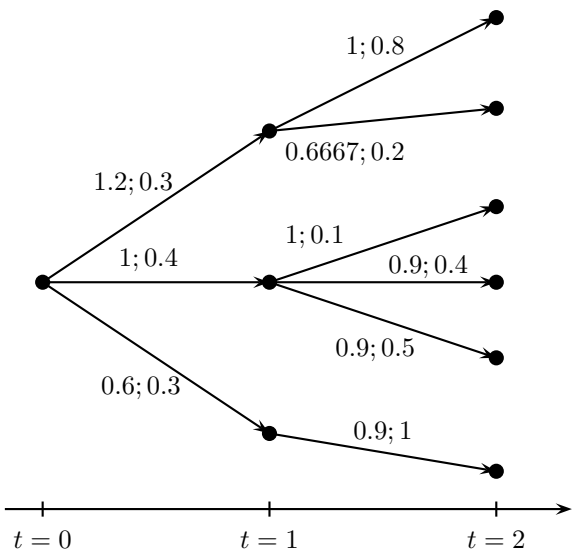


Figure 11.1: Real-world probabilities and the state-price deflator in Example 11.2.

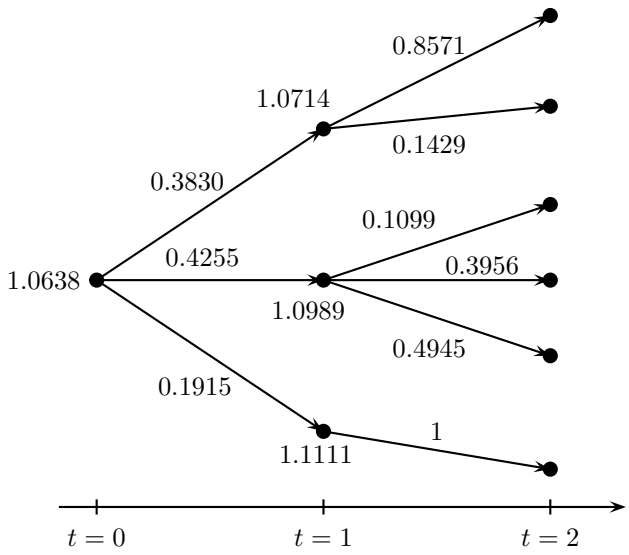


Figure 11.2: Risk-neutral probabilities and one-period risk-free returns in Example 11.2.

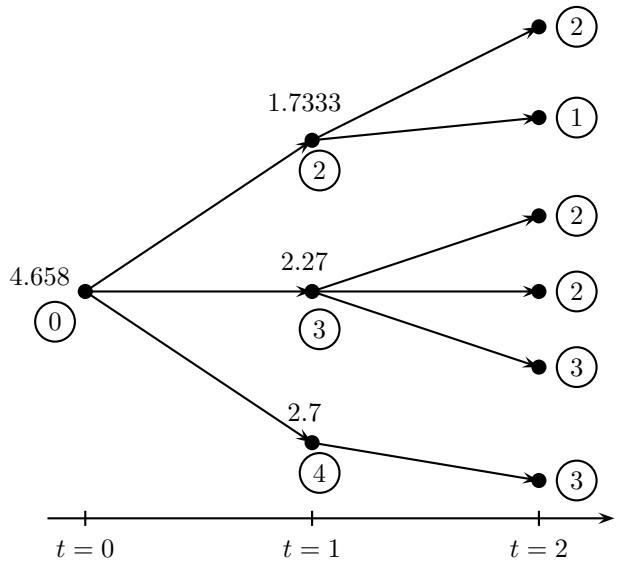


Figure 11.3: Risk-neutral valuation of a dividend process in Example 11.2.

11.4 Forward risk-adjusted probability measures

When valuing an asset with the risk-neutral valuation approach, we have to know the risk-neutral covariance between the risk-free discount factor $(R^f_{t,s})^{-1}$ and the asset dividend D_{is}. Except for simple cases, such covariances are hard to compute analytically. In this section we introduce an alternative probability measure where we do not need to deal with such covariances. The downside is that we have to use a separate probability measure for each payment date.

11.4.1 Definition

Let $s \in \mathcal{T}$ be a trading date and assume that zero-coupon bonds with a face value of 1 maturing at time s are traded. As before, the price at time $t \leq s$ of such a bond is denoted by B^s_t. A probability measure $\mathbb{Q}^s$ on $(\Omega, \mathcal{F}_s)$ is then called a **forward risk-adjusted probability measure** (or just a forward measure) for maturity s if the following conditions are satisfied:

(i) $\mathbb{P}$ and $\mathbb{Q}^s$ are equivalent;

(ii) the Radon-Nikodym derivative $\frac{d\mathbb{Q}^s}{d\mathbb{P}}$ has finite variance;

(iii) the time t price of a dividend paid at time $s \geq t$ equals the product of the zero-coupon bond price B^s_t and the $\mathbb{Q}^s$-expectation of the dividend.

The time t price of a discrete-time dividend process $D_i = (D_{is})$ is then

$$P_{it} = \sum_{s=t+1}^{T} B^s_t \, \mathrm{E}^{\mathbb{Q}^s}_t [D_{is}]. \tag{11.21}$$

No covariance or joint distribution is necessary, but a separate probability measure must be used for each payment date. If you trust the market valuation of bonds, you can observe B^s_t in the bond

market and you only have to find the expected dividend under the appropriate forward measure. If zero-coupon bonds are not traded, implicit zero-coupon bond prices can be derived or estimated from market prices of traded coupon bonds, see e.g. Munk (2005b, Ch. 2).

Apparently, forward measures were introduced by Jamshidian (1987) and Geman (1989). Some authors use the names forward neutral measure or forward martingale measure instead.

The word *forward* can be explained as follows. A forward (contract) on a given asset, say asset i, is a binding agreement between two parties stipulating that one party has to sell a unit of the asset to the other party at a given future point in time, say time s, for a price already set today. The (unique) delivery price that ensures that the present value of this contract equals zero is called the forward price of asset i with delivery at time s. If asset i is assumed to pay no dividends before time s, the forward price for delivery at time s can be shown to be P_{it}/B_t^s, i.e. the current price of the asset "discounted forward in time" using the zero-coupon bond price maturing at the delivery date. The $\mathbb{Q}^s$-measure is defined such that the $\mathbb{Q}^s$-expectation of the dividend equals the forward price of the asset with delivery at time s (in case of no intermediary dividends).

11.4.2 Relation to state-price deflators and risk-neutral measures

The time 0 price of a dividend payment of D_s at time s is given by both $\mathrm{E}[\zeta_s D_s]$ and

$$B_0^s \, \mathrm{E}^{\mathbb{Q}^s}[D_s] = B_0^s \, \mathrm{E}\left[\frac{d\mathbb{Q}^s}{d\mathbb{P}} D_s\right].$$

Therefore, a forward measure for maturity s is related to a state-price deflator through

$$B_0^s \frac{d\mathbb{Q}^s}{d\mathbb{P}} = \zeta_s \quad \Leftrightarrow \quad \frac{d\mathbb{Q}^s}{d\mathbb{P}} = \frac{\zeta_s}{B_0^s} = \frac{\zeta_s}{\mathrm{E}[\zeta_s]}. \tag{11.22}$$

The zero-coupon bond price and therefore the Radon-Nikodym derivative $\frac{d\mathbb{Q}^s}{d\mathbb{P}}$ only "makes sense" up to time s. Results on the existence and uniqueness of $\mathbb{Q}^s$ follow from the corresponding conclusions about state-price deflators.

In terms of a risk-neutral probability measure $\mathbb{Q}$, the time 0 value of the dividend D_s is $\mathrm{E}^{\mathbb{Q}}[(R_{0,s}^f)^{-1} D_s]$ and therefore a forward measure for maturity s is related to a risk-neutral probability measure through the equation

$$B_0^s \frac{d\mathbb{Q}^s}{d\mathbb{Q}} = \left(R_{0,s}^f\right)^{-1} \quad \Leftrightarrow \quad \frac{d\mathbb{Q}^s}{d\mathbb{Q}} = (B_0^s)^{-1} \left(R_{0,s}^f\right)^{-1} = \frac{\left(R_{0,s}^f\right)^{-1}}{\mathrm{E}^{\mathbb{Q}}\left[\left(R_{0,s}^f\right)^{-1}\right]}. \tag{11.23}$$

In a continuous-time framework, the last equality can be rewritten as

$$\frac{d\mathbb{Q}^s}{d\mathbb{Q}} = \frac{e^{-\int_0^s r_u^f \, du}}{\mathrm{E}^{\mathbb{Q}}\left[e^{-\int_0^s r_u^f \, du}\right]}. \tag{11.24}$$

If the future risk-free rates are non-random, we see that the forward measure for maturity s and the risk-neutral probability measure will assign identical probabilities to all events that are decidable at time s, i.e. $\mathbb{Q}^s = \mathbb{Q}$ on $\mathcal{F}_s$. In a one-period economy, $\mathbb{Q}$ and $\mathbb{Q}^1$ are always identical.

Assume a continuous-time setting and write the dynamics of the zero-coupon bond price maturing at time s as

$$dB_t^s = B_t^s \left[\left(r_t^f + (\boldsymbol{\sigma}_t^s)^\top \boldsymbol{\lambda}_t\right) dt + (\boldsymbol{\sigma}_t^s)^\top \, d\boldsymbol{z}_t\right]. \tag{11.25}$$

ω	p	ζ_2	$\frac{d\mathbb{Q}^2}{d\mathbb{P}}$	q^2
1	0.24	1.2	1.3921	0.3341
2	0.06	0.8	0.9281	0.0557
3	0.04	1	1.1601	0.0464
4	0.16	0.9	1.0441	0.1671
5	0.2	0.9	1.0441	0.2088
6	0.3	0.54	0.6265	0.1879

Table 11.2: The $\mathbb{Q}^2$-probabilities in Example 11.3.

This implies that

$$1 = B_s^s = B_0^s \exp\left\{\int_0^s \left(r_t^f + (\boldsymbol{\sigma}_t^s)^\top \boldsymbol{\lambda}_t + \frac{1}{2}\|\boldsymbol{\sigma}_t^s\|^2\right) dt - \int_0^s (\boldsymbol{\sigma}_t^s)^\top \, d\boldsymbol{z}_t\right\}.$$

The Radon-Nikodym derivative of the forward measure with respect to the real-world probability measure can now be written as

$$\frac{d\mathbb{Q}^s}{d\mathbb{P}} = \frac{\zeta_s}{B_0^s} = \exp\left\{-\frac{1}{2}\int_0^s \|\boldsymbol{\lambda}_t - \boldsymbol{\sigma}_t^s\|^2 - \int_0^s (\boldsymbol{\lambda}_t - \boldsymbol{\sigma}_t^s)^\top \, d\boldsymbol{z}_t\right\}. \tag{11.26}$$

According to the Girsanov Theorem 11.1 the process $\boldsymbol{z}^s = (\boldsymbol{z}^s)_{t\in[0,T]}$ defined by $\boldsymbol{z}_0^s = \mathbf{0}$ and

$$d\boldsymbol{z}_t^s = d\boldsymbol{z}_t + (\boldsymbol{\lambda}_t - \boldsymbol{\sigma}_t^s)\, dt \tag{11.27}$$

is a standard Brownian motion under the forward measure for maturity s. The dynamics of any process $X = (X_t)_{t\in[0,T]}$ is transformed via

$$dX_t = \mu_{Xt}\, dt + \boldsymbol{\sigma}_{Xt}^\top d\boldsymbol{z}_t = (\mu_{Xt} - \boldsymbol{\sigma}_{Xt}^\top (\boldsymbol{\lambda}_t - \boldsymbol{\sigma}_t^s))\, dt + \boldsymbol{\sigma}_{Xt}^\top d\boldsymbol{z}_t^s \tag{11.28}$$

and for a price process the analogue is

$$dP_{it} = P_{it}\left[\mu_{it}\, dt + \boldsymbol{\sigma}_{it}^\top d\boldsymbol{z}_t\right] = P_{it}\left[\left(r_t^f - \delta_{it} + \boldsymbol{\sigma}_{it}^\top \boldsymbol{\sigma}_t^s\right) dt + \boldsymbol{\sigma}_{it}^\top d\boldsymbol{z}_t^s\right]. \tag{11.29}$$

11.4.3 Valuation with forward measures

Example 11.3 Consider the same two-period economy as in Example 11.2. Let us find the forward measure for maturity at time 2, i.e. $\mathbb{Q}^2$. First, we must find the price of the zero-coupon bond maturing at time 2:

$$B_0^2 = \mathrm{E}[\zeta_2] = 0.862.$$

Now the forward probabilities of the states can be computed as $q_\omega^2 = \zeta_2(\omega)p_\omega/B_0^2$ yielding the numbers in Table 11.2. Note that the forward probabilities are different from the risk-neutral probabilities computed in Table 11.1.

Given the forward probabilities for maturity 2, it is easy to value a dividend received at time 2. The time 0 value of the time 2 dividend illustrated in Figure 11.3 is

$$B_0^2\, \mathrm{E}^{\mathbb{Q}^2}[D_2] = 0.862 \cdot \left(1 \cdot q_2^2 + 2 \cdot [q_1^2 + q_3^2 + q_4^2] + 3 \cdot [q_5^2 + q_6^2]\right) = 2.018.$$

The dividend received at time 1 is not valued using the forward measure for maturity 2 but with the forward measure for time 1, i.e. $\mathbb{Q}^1$. The forward measure at time 1 only assigns probabilities to the decidable events at time 1, i.e. the events $\{1,2\}$, $\{3,4,5\}$, $\{6\}$ and unions of these events. Since the one-period bond is the risk-free asset over the first period, the $\mathbb{Q}^1$-probabilities are identical to the risk-neutral probabilities of these events, which are depicted in Figure 11.2. Note that these are different from the $\mathbb{Q}^2$-probabilities of the same events, e.g. $q_1^2 + q_2^2 = 0.3898$ while $q_1 + q_2 = 0.3830$. The time 0 value of the time 1 dividend illustrated in Figure 11.3 is

$$B_0^1 \, \mathrm{E}^{\mathbb{Q}^1}[D_1] = (R_0^f)^{-1} \, \mathrm{E}^{\mathbb{Q}}[D_1] = 2.64$$

so that the total time 0 value of the asset is $2.64 + 2.018 = 4.658$ as found in Example 11.2. Exercise 11.3 has more on the forward measures in this example. □

11.5 General risk-adjusted probability measures

Consider an asset with a single dividend payment of D_s at time s. Using the risk-neutral probability measure $\mathbb{Q}$ the price at time $t < s$ of this asset is

$$P_t = \mathrm{E}_t^{\mathbb{Q}} \left[\left(R_{t,s}^f \right)^{-1} D_s \right].$$

If we invest 1 in the bank account at time 0 and roll-over at the short-term interest rate, the value at time t will be $P_t^f = R_{0,t}^f$. We can think of $P^f = (P_t^f)$ as the price process of the bank account. Since $R_{t,s}^f = R_{0,s}^f / R_{0,t}^f = P_s^f / P_t^f$, we can rewrite the above equation as

$$\frac{P_t}{P_t^f} = \mathrm{E}_t^{\mathbb{Q}} \left[\frac{D_s}{P_s^f} \right].$$

Both the current price and the future dividend of the asset is measured relative to the price of the bank account, i.e. the bank account is used as the numeraire. By the Law of Iterated Expectations (Theorem 2.1),

$$\frac{P_t}{P_t^f} = \mathrm{E}_t^{\mathbb{Q}} \left[\mathrm{E}_{t'}^{\mathbb{Q}} \left[\frac{D_s}{P_s^f} \right] \right] = \mathrm{E}_t^{\mathbb{Q}} \left[\frac{P_{t'}}{P_{t'}^f} \right], \quad t < t' \leq s,$$

so the relative price process (P_t / P_t^f) is a $\mathbb{Q}$-martingale.

Valuation with the forward measure for maturity s uses the zero-coupon bond maturing at s as the numeraire. The price of the bond B_t^s converges to its face value of 1 as $t \to s$. If we let $B_s^s = 1$ denote the cum-dividend price of the bond at maturity, the price P_t of the asset paying D_s at time $s > t$ will satisfy

$$\frac{P_t}{B_t^s} = \mathrm{E}_t^{\mathbb{Q}^s} \left[\frac{D_s}{B_s^s} \right]$$

from which it is clear that the zero-coupon bond is used as a numeraire. The relative price process (P_t / B_t^s) is a $\mathbb{Q}^s$-martingale.

In fact we can use any asset or trading strategy having a strictly positive price process as the numeraire and find an equivalent probability measure under which the relative price processes are martingales. Let $\boldsymbol{\theta}$ be a trading strategy with associated value process $V^{\boldsymbol{\theta}} = (V_t^{\boldsymbol{\theta}})_{t \in \mathcal{T}}$ (see Chapter 3 for the precise definition) so that $V_t^{\boldsymbol{\theta}} > 0$ with probability 1 for all $t \in \mathcal{T}$. Then $\mathbb{Q}^{\boldsymbol{\theta}}$ is said to be a risk-adjusted measure for $\boldsymbol{\theta}$ if

(i) $\mathbb{P}$ and $\mathbb{Q}^{\bar{\boldsymbol{\theta}}}$ are equivalent;

(ii) the Radon-Nikodym derivative $\frac{d\mathbb{Q}^{\bar{\boldsymbol{\theta}}}}{d\mathbb{P}}$ has finite variance;

(iii) the time t price of an asset paying a dividend of D_s at time $s > t$ is

$$P_t = V_t^{\boldsymbol{\theta}} \, \mathrm{E}_t^{\mathbb{Q}^{\boldsymbol{\theta}}} \left[\frac{D_s}{V_s^{\boldsymbol{\theta}}} \right]. \tag{11.30}$$

Clearly, the pricing condition (11.30) can be rewritten as

$$\frac{P_t}{V_t^{\boldsymbol{\theta}}} = \mathrm{E}_t^{\mathbb{Q}^{\boldsymbol{\theta}}} \left[\frac{D_s}{V_s^{\boldsymbol{\theta}}} \right],$$

and the relative price process $(P_t/V_t^{\boldsymbol{\theta}})$ is a $\mathbb{Q}^{\boldsymbol{\theta}}$-martingale. If the trading strategy $\boldsymbol{\theta}$ is self-financing, its gross return between t and s will be $R_{t,s}^{\boldsymbol{\theta}} = V_s^{\boldsymbol{\theta}}/V_t^{\boldsymbol{\theta}}$ so that the pricing condition can also be written as

$$P_t = \mathrm{E}_t^{\mathbb{Q}^{\boldsymbol{\theta}}} \left[\left(R_{t,s}^{\boldsymbol{\theta}} \right)^{-1} D_s \right].$$

For a full discrete dividend process $D_i = (D_{it})_{t=1,2,\ldots,T}$, the price will be

$$P_{it} = V_t^{\boldsymbol{\theta}} \, \mathrm{E}_t^{\mathbb{Q}^{\boldsymbol{\theta}}} \left[\sum_{s=t+1}^{T} \frac{D_{is}}{V_s^{\boldsymbol{\theta}}} \right].$$

Comparing the pricing expressions involving expectations under $\mathbb{P}$ and $\mathbb{Q}^{\boldsymbol{\theta}}$, we see that the Radon-Nikodym derivative $\frac{d\mathbb{Q}^{\boldsymbol{\theta}}}{d\mathbb{P}}$ is linked to a state-price deflator through the relations

$$\zeta_t = \frac{V_0^{\boldsymbol{\theta}}}{V_t^{\boldsymbol{\theta}}} \, \mathrm{E}_t \left[\frac{d\mathbb{Q}^{\boldsymbol{\theta}}}{d\mathbb{P}} \right], \quad \frac{d\mathbb{Q}^{\boldsymbol{\theta}}}{d\mathbb{P}} = \zeta_T \frac{V_T^{\boldsymbol{\theta}}}{V_0^{\boldsymbol{\theta}}}. \tag{11.31}$$

Now take a continuous-time framework and write the real-world dynamics of the value of the numeraire as

$$dV_t^{\boldsymbol{\theta}} = V_t^{\boldsymbol{\theta}} \left[\left(r_t^f + \left(\boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \right)^\top \boldsymbol{\lambda}_t \right) dt + \left(\boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \right)^\top d\boldsymbol{z}_t \right],$$

which implies that

$$\frac{V_T^{\boldsymbol{\theta}}}{V_0^{\boldsymbol{\theta}}} = \exp\{ \int_0^T \left(r_t^f + \left(\boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \right)^\top \boldsymbol{\lambda}_t - \frac{1}{2} \| \boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \|^2 \right) dt + \int_0^T \left(\boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \right)^\top d\boldsymbol{z}_t \}.$$

Hence,

$$\frac{d\mathbb{Q}^{\boldsymbol{\theta}}}{d\mathbb{P}} = \zeta_T \frac{V_T^{\boldsymbol{\theta}}}{V_0^{\boldsymbol{\theta}}} = \exp\{ -\frac{1}{2} \int_0^T \| \boldsymbol{\lambda}_t - \boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \|^2 \, dt - \int_0^T \left(\boldsymbol{\lambda}_t - \boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \right)^\top d\boldsymbol{z}_t \},$$

and it follows from the Girsanov Theorem that the process $\boldsymbol{z}^{\boldsymbol{\theta}}$ defined by

$$d\boldsymbol{z}_t^{\boldsymbol{\theta}} = d\boldsymbol{z}_t + \left(\boldsymbol{\lambda}_t - \boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \right) dt$$

is a standard Brownian motion under the measure $\mathbb{Q}^{\boldsymbol{\theta}}$. The dynamics of any process $X = (X_t)_{t \in [0,T]}$ is transformed via

$$dX_t = \mu_{Xt} \, dt + \boldsymbol{\sigma}_{Xt}^\top d\boldsymbol{z}_t = \left(\mu_{Xt} - \boldsymbol{\sigma}_{Xt}^\top \left(\boldsymbol{\lambda}_t - \boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \right) \right) dt + \boldsymbol{\sigma}_{Xt}^\top d\boldsymbol{z}_t^{\boldsymbol{\theta}}$$

and for a price process the analogue is

$$dP_{it} = P_{it} \left[\mu_{it} \, dt + \boldsymbol{\sigma}_{it}^\top d\boldsymbol{z}_t \right] = P_{it} \left[\left(r_t^f - \delta_{it} + \boldsymbol{\sigma}_{it}^\top \boldsymbol{\sigma}_t^{\boldsymbol{\theta}} \right) dt + \boldsymbol{\sigma}_{it}^\top d\boldsymbol{z}_t^{\boldsymbol{\theta}} \right].$$

In particular for the numeraire itself,

$$dV_t^{\boldsymbol{\theta}} = V_t^{\boldsymbol{\theta}} \left[\left(r_t^f + \|\boldsymbol{\sigma}_t^{\boldsymbol{\theta}}\|^2 \right) dt + \left(\boldsymbol{\sigma}_t^{\boldsymbol{\theta}}\right)^\top d\boldsymbol{z}_t^{\boldsymbol{\theta}} \right]. \tag{11.32}$$

Risk-adjusted probability measures are often used in the pricing of derivatives. If the payoff of the derivative is fully determined by some underlying asset, it is sometimes helpful to express the price of the derivative using the risk-adjusted probability measure with the underlying asset as the numeraire. Some examples will be given in Chapter 12.

11.6 Changing the numeraire without changing the measure

Now consider the following question: Is there a trading strategy $\boldsymbol{\theta}$ for which the associated risk-adjusted probability measure is identical to the real-world probability measure, i.e. $\mathbb{Q}^{\boldsymbol{\theta}} = \mathbb{P}$? The answer is affirmative. The so-called growth-optimal portfolio (or just GOP) strategy does the job.

Let us first consider a one-period setting. Here, the growth-optimal portfolio is defined as the portfolio maximizing the expected log-return (or expected log-growth rate of the invested amount) among all portfolios, i.e. it solves

$$\max_{\boldsymbol{\pi}} \mathrm{E}[\ln(\boldsymbol{\pi}^\top \boldsymbol{R})] \quad \text{s.t. } \boldsymbol{\pi}^\top \mathbf{1} = 1.$$

The Lagrangian for this problem is $\mathcal{L} = \mathrm{E}[\ln(\boldsymbol{\pi}^\top \boldsymbol{R})] + \nu(1 - \boldsymbol{\pi}^\top \mathbf{1})$, where ν is the Lagrange multiplier, with first-order condition

$$\mathrm{E}\left[\frac{1}{\boldsymbol{\pi}^\top \boldsymbol{R}} \boldsymbol{R} \right] = \nu \mathbf{1}.$$

We cannot solve explicitly for the portfolio $\boldsymbol{\pi}_{\mathrm{GOP}}$ satisfying this equation, but we can see that its gross return $R_{\mathrm{GOP}} = \boldsymbol{\pi}_{\mathrm{GOP}}^\top \boldsymbol{R}$ satisfies

$$\mathrm{E}\left[\frac{1}{R_{\mathrm{GOP}}} \boldsymbol{R} \right] = \nu \mathbf{1}.$$

Pre-multiplying by any portfolio $\boldsymbol{\pi}$ we get

$$\mathrm{E}\left[\frac{R^{\boldsymbol{\pi}}}{R_{\mathrm{GOP}}} \right] = \nu \boldsymbol{\pi}^\top \mathbf{1} = \nu.$$

In particular, with $\boldsymbol{\pi} = \boldsymbol{\pi}_{\mathrm{GOP}}$, we see that

$$\nu = \mathrm{E}\left[\frac{R_{\mathrm{GOP}}}{R_{\mathrm{GOP}}} \right] = \mathrm{E}[1] = 1.$$

We can thus conclude that for any asset i, we have

$$\mathrm{E}\left[\frac{R_i}{R_{\mathrm{GOP}}} \right] = 1 \quad \Leftrightarrow \quad P_i = \mathrm{E}\left[(R_{\mathrm{GOP}})^{-1} D_i \right]. \tag{11.33}$$

Note that the expectations are under the real-world probability measure.

If the state space is finite, $\Omega = \{1, 2, \ldots, S\}$, and the market is complete, it is possible to construct an Arrow-Debreu asset for any state $\omega \in \Omega$, i.e. an asset with a dividend of 1 if state ω is realized and a zero dividend otherwise. In this case, any portfolio $\boldsymbol{\pi}$ of the basic assets can be seen as a portfolio $\hat{\boldsymbol{\pi}}$ of the S Arrow-Debreu assets. If ψ_ω denotes the state price of state ω,

the gross return on the Arrow-Debreu asset for state ω will be a random variable $R^{\mathrm{AD}(\omega)}$, which if state s is realized takes on the value

$$R_s^{\mathrm{AD}(\omega)} = \begin{cases} \frac{1}{\psi_s} & \text{if } s = \omega, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{R}^{\mathrm{AD}} = \left(R^{\mathrm{AD}(1)}, \ldots, R^{\mathrm{AD}(S)}\right)^\top$ denote the random return vector of the S Arrow-Debreu assets. The gross return on a portfolio $\hat{\boldsymbol{\pi}}$ of Arrow-Debreu assets will be

$$R_s^{\hat{\pi}} = \hat{\boldsymbol{\pi}}^\top \boldsymbol{R}_s^{\mathrm{AD}} = \frac{\hat{\pi}_s}{\psi_s}, \quad s = 1, 2, \ldots, S.$$

If we use the first-order condition (11.33) for the Arrow-Debreu asset for state ω, we therefore get

$$1 = \mathrm{E}\left[\frac{R^{\mathrm{AD}(\omega)}}{R^{\hat{\boldsymbol{\pi}}}_{\mathrm{GOP}}}\right] = p_\omega \frac{1/\psi_\omega}{\hat{\pi}_\omega/\psi_\omega} = \frac{p_\omega}{\hat{\pi}_\omega},$$

where p_ω is the real-world probability of state ω. Therefore, in terms of the Arrow-Debreu assets, the GOP consists of p_ω units of the Arrow-Debreu asset for state ω for each $\omega = 1, 2, \ldots, S$.

In a multi-period setting the growth-optimal trading strategy is the trading strategy maximizing $\mathrm{E}[\ln V_T^{\boldsymbol{\pi}}]$ among all self-financing trading strategies. Hence, it also maximizes $\mathrm{E}[\ln(V_T^{\boldsymbol{\pi}}/V_0^{\boldsymbol{\pi}})]$, the expected log-growth rate between time 0 and time T. For now, focus on the discrete-time framework. Note that

$$\begin{aligned} \ln\left(\frac{V_T^{\boldsymbol{\pi}}}{V_0^{\boldsymbol{\pi}}}\right) &= \ln\left(\frac{V_1^{\boldsymbol{\pi}}}{V_0^{\boldsymbol{\pi}}}\frac{V_2^{\boldsymbol{\pi}}}{V_1^{\boldsymbol{\pi}}} \cdots \frac{V_T^{\boldsymbol{\pi}}}{V_{T-1}^{\boldsymbol{\pi}}}\right) \\ &= \ln\left(\frac{V_1^{\boldsymbol{\pi}}}{V_0^{\boldsymbol{\pi}}}\right) + \ln\left(\frac{V_2^{\boldsymbol{\pi}}}{V_1^{\boldsymbol{\pi}}}\right) + \cdots + \ln\left(\frac{V_T^{\boldsymbol{\pi}}}{V_{T-1}^{\boldsymbol{\pi}}}\right) \\ &= \ln R_1^{\boldsymbol{\pi}} + \ln R_2^{\boldsymbol{\pi}} + \cdots + \ln R_T^{\boldsymbol{\pi}}, \end{aligned}$$

where $R_{t+1}^{\boldsymbol{\pi}} = V_{t+1}^{\boldsymbol{\pi}}/V_t^{\boldsymbol{\pi}}$ is the gross return on the trading strategy between time t and time $t+1$, i.e. the gross return on the portfolio $\boldsymbol{\pi}_t$ chosen at time t. Therefore the growth-optimal trading strategy $\boldsymbol{\pi} = (\boldsymbol{\pi}_t)_{t\in\mathcal{T}}$ is such that each $\boldsymbol{\pi}_t$ maximizes $\mathrm{E}_t[\ln R_{t+1}^{\boldsymbol{\pi}_t}] = \mathrm{E}_t[\ln(\boldsymbol{\pi}_t^\top \boldsymbol{R}_{t+1})]$, where $\boldsymbol{R}_{t+1}$ is the vector of gross returns on all the basic assets between time t and time $t+1$. As in the one-period setting, the first-order condition implies that

$$\mathrm{E}_t\left[\frac{R_{i,t+1}}{R_{t+1}^{\boldsymbol{\pi}}}\right] = 1$$

for all assets i (and portfolios). Again, it is generally not possible to solve explicitly for the portfolio $\boldsymbol{\pi}_t$.

In the continuous-time framework assume that a bank account is traded with instantaneous risk-free rate of return r_t^f and let $\boldsymbol{\pi}_t$ denote the portfolio weights of the instantaneously risky assets. Then the dynamics of the value $V_t^{\boldsymbol{\pi}}$ of a self-financing trading strategy $\boldsymbol{\pi} = (\boldsymbol{\pi}_t)_{t\in[0,T]}$ is given by

$$dV_t^{\boldsymbol{\pi}} = V_t^{\boldsymbol{\pi}}\left[\left(r_t^f + \boldsymbol{\pi}_t^\top[\boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r_t^f \mathbf{1}]\right)\, dt + \boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}_t\, d\boldsymbol{z}_t\right], \tag{11.34}$$

cf. Sections 3.3.3 and 6.5.2. This implies that

$$V_T^{\boldsymbol{\pi}} = V_0^{\boldsymbol{\pi}} \exp\left\{\int_0^T \left(r_t^f + \boldsymbol{\pi}_t^\top[\boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r_t^f \mathbf{1}] - \frac{1}{2}\boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}_t \underline{\underline{\sigma}}_t^\top \boldsymbol{\pi}_t\right) dt + \int_0^T \boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}_t\, d\boldsymbol{z}_t\right\},$$

and thus

$$\ln\left(\frac{V_T^{\boldsymbol{\pi}}}{V_0^{\boldsymbol{\pi}}}\right) = \int_0^T \left(r_t^f + \boldsymbol{\pi}_t^\top[\boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r_t^f\mathbf{1}] - \frac{1}{2}\boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}_t \underline{\underline{\sigma}}_t^\top \boldsymbol{\pi}_t\right) dt + \int_0^T \boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}_t \, d\boldsymbol{z}_t.$$

If the process $\boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}_t$ is sufficiently nice, the stochastic integral in the above equation will have mean zero so that the growth-optimal trading strategy is maximizing the expectation of the first integral, which can be found by maximizing $\boldsymbol{\pi}_t^\top[\boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r_t^f\mathbf{1}] - \frac{1}{2}\boldsymbol{\pi}_t^\top \underline{\underline{\sigma}}_t \underline{\underline{\sigma}}_t^\top \boldsymbol{\pi}_t$ for each t and each state. The first-order condition implies that

$$\underline{\underline{\sigma}}_t \underline{\underline{\sigma}}_t^\top \boldsymbol{\pi}_t = \boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r_t^f\mathbf{1}, \tag{11.35}$$

which means that

$$\underline{\underline{\sigma}}_t^\top \boldsymbol{\pi}_t = \boldsymbol{\lambda}_t \tag{11.36}$$

for some market price of risk process $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_t)$. If $\underline{\underline{\sigma}}_t$ is a square, non-singular matrix, the unique GOP strategy is given by

$$\boldsymbol{\pi}_t = \left(\underline{\underline{\sigma}}_t^\top\right)^{-1} \boldsymbol{\lambda}_t = \left(\underline{\underline{\sigma}}_t \underline{\underline{\sigma}}_t^\top\right)^{-1} \left(\boldsymbol{\mu}_t + \boldsymbol{\delta}_t - r_t^f\mathbf{1}\right). \tag{11.37}$$

This shows that the GOP strategy is a combination of the instantaneously risk-free asset and the tangency portfolio of risky assets, introduced in Section 6.5.2. The GOP strategy is the optimal trading strategy for an individual with time-additive logarithmic utility. Substituting the expression for $\boldsymbol{\pi}_t$ back into the value dynamics, we see that the value of the GOP strategy evolves as

$$dV_t^{\boldsymbol{\pi}} = V_t^{\boldsymbol{\pi}} \left[\left(r_t^f + \|\boldsymbol{\lambda}_t\|^2\right) dt + \boldsymbol{\lambda}_t^\top \, d\boldsymbol{z}_t\right]. \tag{11.38}$$

The value process of the GOP strategy (and the real-world probability measure) contain sufficient information to price any specific dividend process. Since knowing the value process of the GOP strategy boils down to knowing the risk-free rate process and the market price of risk process, this is not a surprise. Also note that $\zeta_t = V_0^{\boldsymbol{\pi}}/V_t^{\boldsymbol{\pi}}$ defines a state-price deflator.

11.7 Concluding remarks

This chapter has introduced several risk-adjusted probability measures and discussed the application of these measures in the pricing of assets. Each risk-adjusted probability measure (in conjunction with the price process for the associated numeraire) is in a one-to-one relation with a state-price deflator. The models for state-price deflators developed in the previous chapters are therefore also models concretizing the risk-adjusted probability measures. For example, the consumption-based CAPM will define a risk-neutral probability measure in terms of the (aggregate) consumption of a given (representative) individual. Any full factor pricing model will also nail down the risk-neutral probability measure.

As discussed in Chapter 4, there is a distinction between *real* state-price deflators and *nominal* state-price deflators. Similarly, we can distinguish between a real risk-neutral probability measure and a nominal risk-neutral probability measure. The real (nominal, respectively) risk-neutral probability measure is defined with a bank account yielding the real (nominal, respectively) short-term risk-free interest rate as the numeraire. In the same manner real and nominal forward measures are defined with a real and nominal, respectively, zero-coupon bond as the numeraire.

11.8 Exercises

EXERCISE 11.1 Show that the Equations (11.8) and (11.9) are equivalent.

EXERCISE 11.2 Take a continuous-time framework and assume that $\zeta = (\zeta_t)_{t\in[0,T]}$ is a state-price deflator. What is the $\mathbb{Q}$-dynamics of ζ?

EXERCISE 11.3 Consider Example 11.3. Compute the conditional $\mathbb{Q}^2$-probabilities of the transitions over the second period of the tree. Compare with conditional $\mathbb{Q}$-probabilities illustrated in Figure 11.2 and explain why they are (not?) different.

EXERCISE 11.4 In the same two-period economy considered in Examples 11.2 and 11.3, compute the price of an asset giving a time 1 dividend of 0 in the upper or middle node and 1 in the lower node and a time 2 dividend of 3, 2, 3, 3, 4, or 5 from the top node and down (this is asset 3 in Exercise 4.8).

Chapter 12

Derivatives

12.1 Introduction

A derivative is an asset whose dividend(s) and price are derived from the price of another asset, the underlying asset, or the value of some other variable. The main types of derivatives are forwards, futures, options, and swaps. While a large number of different derivatives are traded in today's financial markets, most of them are variations of these four main types.

A forward is the simplest derivative. A forward contract is an agreement between two parties on a given transaction at a given future point in time and at a price that is already fixed when the agreement is made. For example, a forward on a bond is a contract where the parties agree to trade a given bond at a future point in time for a price which is already fixed today. This fixed price is usually set so that the value of the contract at the time of inception is equal to zero so that no money changes hand before the delivery date. Forward contracts are not traded or listed at financial exchanges but traded in quite organized over-the-counter (OTC) markets dominated by large financial institutions. For example, forwards on foreign exchange are quite common.

As a forward contract, a futures contract is an agreement upon a specified future transaction, e.g. a trade of a given security. The special feature of a future is that changes in its value are settled continuously throughout the life of the contract (usually once every trading day). This so-called *marking-to-market* ensures that the value of the contract (i.e. the value of the payments still to come) is zero immediately following a settlement. This procedure makes it practically possible to trade futures at organized exchanges, since there is no need to keep track of when the futures position was originally taken. Futures on many different assets or variables are traded at different exchanges around the world, including futures on stocks, bonds, interest rates, foreign exchange, oil, metals, frozen concentrate orange juice, live cattle, and the temperature in Las Vegas!

An option gives the holder the right to make some specified future transaction at terms that are already fixed. A call option gives the holder the right to buy a given security at a given price at or before a given date. Conversely, a put option gives the holder the right to sell a given security. If the option gives the right to make the transaction at only one given date, the option is said to be European-style. If the right can be exercised at any point in time up to some given date, the option is said to be American-style. Both European- and American-style options are traded. Options on stocks, bonds, foreign exchange, and many other assets, commodities, and variables

Instruments/	Futures		Options	
Location	Amount outstanding	Turnover	Amount outstanding	Turnover
All markets	17,662	213,455	31,330	75,023
Interest rate	17,024	202,064	28,335	63,548
Currency	84	1,565	37	120
Equity index	553	9,827	2,958	11,355
North America	9,778	122,516	18,120	49,278
Europe	5,534	77,737	12,975	19,693
Asia-Pacific	2,201	11,781	170	5,786
Other markets	149	1,421	66	266

Table 12.1: Derivatives traded on organized exchanges. All amounts are in billions of US dollars. The amount outstanding is of September 2004, while the turnover figures are for the third quarter of 2004. Source: Table 23A in BIS (2004).

are traded at many exchanges around the world and also on the OTC-markets. Also options on futures on some asset or variable are traded, i.e. a derivative on a derivative! In addition, many financial assets or contracts have "embedded" options. For example, many mortgage-backed bonds and corporate bonds are callable, in the sense that the issuer has the right to buy back the bond at a pre-specified price.

A swap is an exchange of two dividend streams between two parties. In a "plain vanilla" interest rate swap, two parties exchange a stream of fixed interest rate payments and a stream of floating interest rate payments. In a currency swap, streams of payments in different currencies are exchanged. Many exotic swaps with special features are widely used.

The markets for derivatives are of an enormous size. Table 12.1 provides some interesting statistics published by the Bank for International Settlements (BIS) on the size of derivatives markets at organized exchanges. The markets for interest rate derivatives are much larger than the markets for currency- or equity-linked derivatives. The option markets generally dominate futures markets measured by the amounts outstanding but ranked according to turnover futures markets are larger than options markets.

The BIS statistics also contain information about the size of OTC markets for derivatives. BIS estimates that in June 2004 the total amount outstanding on OTC derivative markets was 220,058 billions of US dollars, of which single-currency interest rate derivatives account for 164,626 billions, currency derivatives account for 26,997 billions, equity-linked derivatives for 4,521 billions, commodity contracts for 1,270 billions, while the remaining 22,644 billions cannot be split into any of these categories, cf. Table 19 in BIS (2004). Table 12.2 shows how the interest rate derivatives market can be disaggregated according to instrument and maturity. Approximately 38% of these OTC-traded interest rate derivatives are denominated in Euro, 35% in US dollars, 13% in yen, and 7% in pound sterling, cf. Table 21B in BIS (2004).

This chapter gives an introduction to frequently traded derivatives and their valuation. We will specify the payments of these derivatives, discuss the links between different derivatives, and we will also indicate what we can conclude about their prices from general asset pricing theory.

Contracts	total	Maturity in years		
		≤ 1	1–5	≥ 5
All interest rate	164,626	57,157	66,093	41,376
Forward rate agreements	13,144	49,397	56,042	35,275
Swaps	127,570			
Options	23,912	7,760	10,052	6,101

Table 12.2: Amounts outstanding (billions of US dollars) on OTC single-currency interest rate derivatives as of June 2004. Source: Tables 21A and 21C in BIS (2004).

Throughout the chapter we assume that prices are arbitrage-free and that a bank account and zero-coupon bonds of relevant maturities are traded so that we can define and work with the risk-neutral probability measures and forward measures introduced in Chapter 11. We will denote the continuously compounded risk-free short-term interest rate by r_t instead of r_t^f.

Section 12.2 deals with forwards and futures, Section 12.3 with options, and Section 12.4 with swaps and swaptions. Some features of American-style derivatives are discussed in Section 12.5.

12.2 Forwards and futures

12.2.1 General results on forward prices and futures prices

A forward with maturity date T and delivery price K provides a dividend of $P_T - K$ at time T, where P is the underlying variable, typically the price of an asset or a specific interest rate. If you plan to buy a unit of an asset at time T, you can lock in the effective purchase price with a forward on that asset. Conversely, if you plan to sell a unit of an asset, you can lock in the effective selling price by taking a short position in a forward on the asset, which will give a terminal dividend of $K - P_T$. Of course, forwards can also be used for speculation. If you believe in high values of P_T, you can take a long position in a forward. If you believe in low values of P_T, you can take a short position in a forward.

In terms of a risk-neutral probability measure $\mathbb{Q}$, the time t value of such a future payoff can be written as

$$\begin{aligned} V_t &= \mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1}(P_T - K)\right] \\ &= \mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1} P_T\right] - K\,\mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1}\right] \\ &= \mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1} P_T\right] - K B_t^T, \end{aligned}$$

where $B_t^T = \mathrm{E}_t^{\mathbb{Q}}[(R_{t,T}^f)^{-1}]$ is the price of the zero-coupon bond maturing at time T with a unit payment. Here $R_{t,T}^f$ is the gross return between time t and T on the bank account, i.e. a roll-over in short risk-free investments, cf. the discussion in Section 11.3.

For forwards contracted upon at time t, the delivery price K is typically set so that the value of the forward at time t is zero. This value of K is called the forward price at time t (for the delivery date T) and is denoted by F_t^T. We define the terminal forward price to be $F_T^T = P_T$, the only

reasonable price for immediate delivery. Solving the equation $V_t = 0$, we get that the forward price is given by

$$F_t^T = \frac{\mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1} P_T\right]}{B_t^T}.$$

If the underlying variable is the price of a traded asset with no payments in the period $[t, T]$, we have

$$\mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1} P_T\right] = P_t,$$

so that the forward price can be written as $F_t^T = P_t/B_t^T$. Applying a well-known property of covariances, we have that

$$\begin{aligned}\mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1} P_T\right] &= \mathrm{Cov}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1}, P_T\right] + \mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1}\right]\mathrm{E}_t^{\mathbb{Q}}[P_T] \\ &= \mathrm{Cov}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1}, P_T\right] + B_t^T\,\mathrm{E}_t^{\mathbb{Q}}[P_T]\end{aligned}$$

and therefore

$$F_t^T = \mathrm{E}_t^{\mathbb{Q}}[P_T] + \frac{\mathrm{Cov}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1}, P_T\right]}{B_t^T}. \tag{12.1}$$

We can also characterize the forward price in terms of the forward measure for maturity T. The forward price process for contracts with delivery date T is a $\mathbb{Q}^T$-martingale. This is clear from the following considerations. With B_t^T as the numeraire, we have that the forward price F_t^T is set so that

$$\frac{0}{B_t^T} = \mathrm{E}_t^{\mathbb{Q}^T}\left[\frac{P_T - F_t^T}{B_T^T}\right]$$

and hence

$$F_t^T = \mathrm{E}_t^{\mathbb{Q}^T}[P_T] = \mathrm{E}_t^{Q^T}[F_T^T],$$

which implies that the forward price F_t^T is a $\mathbb{Q}^T$-martingale.

We summarize our findings in the following theorem.

Theorem 12.1 *The forward price for delivery at time T is given by*

$$F_t^T = \frac{\mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1} P_T\right]}{B_t^T} = \mathrm{E}_t^{\mathbb{Q}^T}[P_T]. \tag{12.2}$$

If the underlying variable is the price of a traded asset with no payments in the period $[t, T]$, the forward price can be written as

$$F_t^T = \frac{P_t}{B_t^T}. \tag{12.3}$$

Note that when (12.3) holds, the forward price of an asset follows immediately from the spot price of the asset and the price of the zero-coupon bond maturing at the delivery date. No model for the price dynamics of the underlying asset is needed. This is because the forward is perfectly replicated by a portfolio of one unit of the underlying asset and a short position in K zero-coupon bonds maturing at the delivery date of the forward.

Consider now a futures contract with final settlement at time T. The marking-to-market at a given date involves the payment of the change in the so-called futures price of the contract relative

to the previous settlement date. Let Φ_t^T be the futures price at time t. The futures price at the settlement time is by definition equal to the price of the underlying security, $\Phi_T^T = P_T$. At maturity of the contract the futures thus gives a payoff equal to the difference between the price of the underlying asset at that date and the futures price at the previous settlement date. After the last settlement before maturity, the futures is therefore indistinguishable from the corresponding forward contract, so the values of the futures and the forward at that settlement date must be identical. At the next-to-last settlement date before maturity, the futures price is set to that value that ensures that the net present value of the upcoming settlement at the last settlement date before maturity (which depends on this futures price) *and* the final payoff is equal to zero. Similarly at earlier settlement dates. We assume that the futures is marked-to-market at every trading date considered in the model. In the discrete-time framework, the dividend from the futures at time $t+1$ is therefore $\Phi_{t+1}^T - \Phi_t^T$. In a continuous-time setting, the dividend over any infinitesimal interval $[t, t+dt]$ is $d\Phi_t^T$. The following theorem characterizes the futures price:

Theorem 12.2 *The futures price Φ_t^T is a martingale under the risk-neutral probability measure $\mathbb{Q}$. In particular,*

$$\Phi_t^T = \mathrm{E}_t^{\mathbb{Q}}\left[P_T\right]. \tag{12.4}$$

Proof: We give a proof in the discrete-time framework, a proof originally due to Cox, Ingersoll, and Ross (1981b). Then the continuous-time version of the result follows by taking a limit. For a proof based on the same idea but formulated directly in continuous time, see Duffie and Stanton (1992).

Consider a discrete-time setting in which positions can be changed and the futures contracts marked-to-market at times $t, t+\Delta t, t+2\Delta t, \ldots, t+N\Delta t \equiv T$. Let R_t^f denote the risk-free gross return between t and $t+\Delta t$ and let $R_{t,t+n\Delta t}^f = R_t^f R_{t+\Delta t}^f \ldots R_{t+(n-1)\Delta t}^f$. The idea is to set up a self-financing strategy that requires an initial investment at time t equal to the futures price Φ_t^T. Hence, at time t, Φ_t^T is invested in the bank account. In addition, R_t^f futures contracts are acquired (at a price of zero).

At time $t+\Delta t$, the deposit at the bank account has grown to $R_t^f \Phi_t^T$. The marking-to-market of the futures position yields a payoff of $R_t^f\left(\Phi_{t+\Delta t}^T - \Phi_t^T\right)$, which is deposited at the bank account, so that the balance of the account becomes $R_t^f \Phi_{t+\Delta t}^T$. The position in futures is increased (at no extra costs) to a total of $R_t^f R_{t+\Delta t}^f = R_{t,t+2\Delta t}^f$ contracts.

At time $t+2\Delta t$, the deposit has grown to $R_{t,t+2\Delta t}^f \Phi_{t+\Delta t}^T$, which together with the marking-to-market payment of $R_{t,t+2\Delta t}^f\left(\Phi_{t+2\Delta t}^T - \Phi_{t+\Delta t}^T\right)$ gives a total of $R_{t,t+2\Delta t}^f \Phi_{t+2\Delta t}^T$.

Continuing this way, the balance of the bank account at time $T = t+N\Delta t$ will be

$$R_{t,t+N\Delta t}^f \Phi_{t+N\Delta t}^T = R_{t,T}^f \Phi_T^T = R_{t,T}^f P_T.$$

This is true for any value of Δt and $\Delta t = 1$ covers our standard discrete-time framework and $\Delta t \to 0$ gives the continuous-time limit.

So a self-financing trading strategy with an initial time t investment of Φ_t^T will give a dividend of $R_{t,T}^f P_T$ at time T. On the other hand, we can value the time T dividend by multiplying by $(R_{t,T}^f)^{-1}$ and taking the risk-neutral expectation. Hence,

$$\Phi_t^T = \mathrm{E}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1} R_{t,T}^f P_T\right] = \mathrm{E}_t^{\mathbb{Q}}[P_T],$$

as was to be shown. □

Note that in order to compute the futures price of an asset we generally have to model the dynamics of the underlying spot price.

Comparing (12.4) with (11.12), we see that we can think of the futures price as the price of a traded asset with a continuous dividend given by the product of the current price and the short-term interest rate.

From (12.1) and (12.4) we get that the difference between the forward price F_t^T and the futures price Φ_t^T is given by

$$F_t^T - \Phi_t^T = \frac{\mathrm{Cov}_t^{\mathbb{Q}}\left[\left(R_{t,T}^f\right)^{-1}, P_T\right]}{B_t^T}. \tag{12.5}$$

The forward price and the futures price will only be identical if the two random variables P_T and $(R_{t,T}^f)^{-1}$ are uncorrelated under the risk-neutral probability measure. In particular, this is true if the short-term risk-free rate is constant or deterministic.

The forward price is larger [smaller] than the futures price if the variables $(R_{t,T}^f)^{-1}$ and P_T are positively [negatively] correlated under the risk-neutral probability measure. An intuitive, heuristic argument for this goes as follows. If the forward price and the futures price are identical, the total undiscounted payments from the futures contract will be equal to the terminal payment of the forward. Suppose the interest rate and the spot price of the underlying asset are positively correlated, which ought to be the case whenever $(R_{t,T}^f)^{-1}$ and P_T are negatively correlated. Then the marking-to-market payments of the futures tend to be positive when the interest rate is high and negative when the interest rate is low. So positive payments can be reinvested at a high interest rate, whereas negative payments can be financed at a low interest rate. With such a correlation, the futures contract is clearly more attractive than a forward contract when the futures price and the forward price are identical. To maintain a zero initial value of both contracts, the futures price has to be larger than the forward price. Conversely, if the sign of the correlation is reversed.

If the underlying asset has a constant or deterministic volatility and pays no dividends before time T, we can write the risk-neutral price dynamics as

$$dP_t = P_t\left[r_t\,dt + \sigma(t)\,dz_t^{\mathbb{Q}}\right],$$

where $z^{\mathbb{Q}} = (z_t^{\mathbb{Q}})$ is a standard Brownian motion under the risk-neutral measure $\mathbb{Q}$. If, furthermore, the short-term risk-free rate is constant or follows a Gaussian process as for example in the Vasicek model introduced in Section 10.5.1, the future values of P_T will be lognormally distributed under the risk-neutral measure. In that case, the futures price can be stated in closed form. In Exercise 12.6 you are asked to compute and compare the forward price and the futures price on a zero-coupon bond under the assumptions of the Vasicek model of interest rate dynamics introduced in Section 10.5.1.

12.2.2 Interest rates forwards and futures

Forward interest rates are rates for a future period relative to the time where the rate is set. Many participants in the financial markets may on occasion be interested in "locking in" an interest rate for a future period, either in order to hedge risk involved with varying interest rates or to speculate

in specific changes in interest rates. In the money markets the agents can lock in an interest rate by entering a forward rate agreement (FRA). Suppose the relevant future period is the time interval between T and S, where $S > T$. In principle, a forward rate agreement with a face value H and a contract rate of K involves two payments: a payment of $-H$ at time T and a payment of $H[1+(S-T)K]$ at time S. (Of course, the payments to the other part of the agreement are H at time T and $-H[1+(S-T)K]$ at time S.) In practice, the contract is typically settled at time T so that the two payments are replaced by a single payment of $B_T^S H[1+(S-T)K]-H$ at time T.

Usually the contract rate K is set so that the present value of the future payment(s) is zero at the time the contract is made. Suppose the contract is made at time $t<T$. Then the time t value of the two future payments of the contract is equal to $-HB_t^T + H[1+(S-T)K]B_t^S$. This is zero if and only if

$$K = \frac{1}{S-T}\left(\frac{B_t^T}{B_t^S} - 1\right) = L_t^{T,S},$$

cf. (10.6), i.e. when the contract rate equals the forward rate prevailing at time t for the period between T and S. For this contract rate, we can think of the forward rate agreement having a single payment at time T, which is given by

$$B_T^S H[1+(S-T)K] - H = H\left(\frac{1+(S-T)L_t^{T,S}}{1+(S-T)l_T^S} - 1\right) = \frac{(S-T)(L_t^{T,S} - l_T^S)H}{1+(S-T)l_T^S}. \tag{12.6}$$

The numerator is exactly the interest lost by lending out H from time T to time S at the forward rate given by the FRA rather than the realized spot rate. Of course, this amount may be negative, so that a gain is realized. The division by $1+(S-T)l_T^S$ corresponds to discounting the gain/loss from time S back to time T.

Interest rate futures trade with a very high volume at several international exchanges, e.g. CME (Chicago Mercantile Exchange), LIFFE (London International Financial Futures & Options Exchange), and MATIF (Marché à Terme International de France). The CME interest rate futures involve the three-month Eurodollar deposit rate and are called Eurodollar futures. The interest rate involved in the futures contracts traded at LIFFE and MATIF is the three-month LIBOR rate on the Euro currency. We shall simply refer to all these contracts as Eurodollar futures and refer to the underlying interest rate as the three-month LIBOR rate, whose value at time t we denote by $l_t^{t+0.25}$.

The price quotation of Eurodollar futures is a bit complicated since the amounts paid in the marking-to-market settlements are not exactly the changes in the quoted futures price. We must therefore distinguish between the *quoted* futures price, $\tilde{\mathcal{E}}_t^T$, and the *actual* futures price, $\mathcal{E}_t^T$, with the settlements being equal to changes in the actual futures price. At the maturity date of the contract, T, the quoted Eurodollar futures price is defined in terms of the prevailing three-month LIBOR rate according to the relation

$$\tilde{\mathcal{E}}_T^T = 100\left(1 - l_T^{T+0.25}\right), \tag{12.7}$$

which using (10.5) on page 222 can be rewritten as

$$\tilde{\mathcal{E}}_T^T = 100\left(1 - 4\left(\frac{1}{B_T^{T+0.25}} - 1\right)\right) = 500 - 400\,\frac{1}{B_T^{T+0.25}}.$$

Traders and analysts typically transform the Eurodollar futures price to an interest rate, the so-called **LIBOR futures rate**, which we denote by φ_t^T and define by

$$\varphi_t^T = 1 - \frac{\tilde{\mathcal{E}}_t^T}{100} \quad \Leftrightarrow \quad \tilde{\mathcal{E}}_t^T = 100\left(1 - \varphi_t^T\right).$$

It follows from (12.7) that the LIBOR futures rate converges to the three-month LIBOR spot rate, as the maturity of the futures contract approaches.

The actual Eurodollar futures price is given by

$$\mathcal{E}_t^T = 100 - 0.25(100 - \tilde{\mathcal{E}}_t^T) = \frac{1}{4}\left(300 + \tilde{\mathcal{E}}_t^T\right) = 100 - 25\varphi_t^T$$

per 100 dollars of nominal value. It is the change in the actual futures price which is exchanged in the marking-to-market settlements. At the CME the nominal value of the Eurodollar futures is 1 million dollars. A quoted futures price of $\tilde{\mathcal{E}}_t^T = 94.47$ corresponds to a LIBOR futures rate of 5.53% and an actual futures price of

$$\frac{1\,000\,000}{100} \cdot [100 - 25 \cdot 0.0553] = 986\,175.$$

If the quoted futures price increases to 94.48 the next day, corresponding to a drop in the LIBOR futures rate of one basis point (0.01 percentage points), the actual futures price becomes

$$\frac{1\,000\,000}{100} \cdot [100 - 25 \cdot 0.0552] = 986\,200.$$

An investor with a long position will therefore receive $986\,200 - 986\,175 = 25$ dollars at the settlement at the end of that day.

If we simply sum up the individual settlements without discounting them to the terminal date, the total gain on a long position in a Eurodollar futures contract from t to expiration at T is given by

$$\mathcal{E}_T^T - \mathcal{E}_t^T = \left(100 - 25\varphi_T^T\right) - \left(100 - 25\varphi_t^T\right) = -25\left(\varphi_T^T - \varphi_t^T\right)$$

per 100 dollars of nominal value, i.e. the total gain on a contract with nominal value H is equal to $-0.25\left(\varphi_T^T - \varphi_t^T\right)H$. The gain will be positive if the three-month spot rate at expiration turns out to be below the futures rate when the position was taken. Conversely for a short position. The gain/loss on a Eurodollar futures contract is closely related to the gain/loss on a forward rate agreement, as can be seen from substituting $S = T + 0.25$ into (12.6). Recall that the rates φ_T^T and $l_T^{T+0.25}$ are identical. However, it should be emphasized that in general the futures rate φ_t^T and the forward rate $L_t^{T,T+0.25}$ will be different due to the marking-to-market of the futures contract.

The final settlement is based on the terminal actual futures price

$$\begin{aligned}\mathcal{E}_T^T &\equiv 100 - 0.25\left(100 - \tilde{\mathcal{E}}_T^T\right)\\ &= 100 - 0.25\left(400\left[(B_T^{T+0.25})^{-1} - 1\right]\right)\\ &= 100\left[2 - (B_T^{T+0.25})^{-1}\right].\end{aligned}$$

It follows from Theorem 12.2 that the actual futures price at any earlier point in time t can be computed as

$$\mathcal{E}_t^T = \mathrm{E}_t^{\mathbb{Q}}\left[\mathcal{E}_T^T\right] = 100\left(2 - \mathrm{E}_t^{\mathbb{Q}}\left[(B_T^{T+0.25})^{-1}\right]\right).$$

The quoted futures price is therefore

$$\tilde{\mathcal{E}}_t^T = 4\mathcal{E}_t^T - 300 = 500 - 400\,\mathrm{E}_t^{\mathbb{Q}}\left[(B_T^{T+0.25})^{-1}\right]. \tag{12.8}$$

In several models of interest rate dynamics and bond prices, including the Vasicek and Cox-Ingersoll-Ross models introduced in Chapter 10, the expectation in (12.8) can be computed in closed form; see, e.g., Munk (2005b).

12.3 Options

In this section, we focus on European options. Some aspects of American options are discussed in Section 12.5.

12.3.1 General pricing results for European options

A European call option with an exercise price of K and expiration at time T gives a dividend at T of

$$C_T = \max(P_T - K, 0),$$

where P_T is the value at time T of the underlying variable of the option. For an option on a traded asset, P_T is the price of the underlying asset at the expiry date. For an option on a given interest rate, P_T denotes the value of this interest rate at the expiry date. With a call option you can speculate in high values of P_T. A call option on an asset offers protection to an investor who wants to purchase the underlying asset at time T. The call option ensures that the investor effectively pays at most K for the underlying asset. The call option price is the price of that protection.

Similarly, a European put option with an exercise price of K and expiration at time T gives a dividend at T of

$$\Pi_T = \max(K - P_T, 0).$$

With a put option you can speculate in low values of P_T. A put option offers protection to an investor who wants to sell the underlying asset at time T. The put option ensures that the effective selling price is at least K.

Prices of European call and put options on the same underlying variable are closely related. Since $C_T + K = \Pi_t + P_T$, it is clear that

$$C_t + KB_t^T = \Pi_t + B_t^T\,\mathrm{E}_t^{\mathbb{Q}^T}\left[P_T\right],$$

where $\mathbb{Q}^T$ is the T-forward martingale measure. In particular, if the underlying variable is the price of a non-dividend paying asset, we have $P_t = B_t^T\,\mathrm{E}_t^{\mathbb{Q}^T}[P_T]$ and thus the following result:

Theorem 12.3 (Put-call parity) *The prices of a European call option and a European put option on a non-dividend paying asset are related through the equation*

$$C_t + KB_t^T = \Pi_t + P_t. \tag{12.9}$$

A portfolio of a call option and K zero-coupon bonds maturing at time T gives exactly the same dividend as a portfolio of a put option and the underlying asset. The put-call parity (12.9) follows

by absence of arbitrage. A consequence of the put-call parity is that we can focus on the pricing of European call options. The prices of European put options will then follow immediately.

Now let us focus on the call option. In terms of the forward measure $\mathbb{Q}^T$ for maturity T, the time t price of the option is

$$C_t = B_t^T \, \mathrm{E}_t^{\mathbb{Q}^T} \left[\max(P_T - K, 0) \right]. \tag{12.10}$$

We can rewrite the payoff as

$$C_T = (P_T - K) \, \mathbf{1}_{\{P_T > K\}},$$

where $\mathbf{1}_{\{P_T>K\}}$ is the indicator for the event $P_T > K$. This indicator is a random variable whose value will be 1 if the realized value of P_T turns out to be larger than K and the value is 0 otherwise. Hence, the option price can be rewritten as[1]

$$\begin{aligned} C_t &= B_t^T \, \mathrm{E}_t^{\mathbb{Q}^T} \left[(P_T - K) \mathbf{1}_{\{P_T>K\}} \right] \\ &= B_t^T \left(\mathrm{E}_t^{\mathbb{Q}^T} \left[P_T \mathbf{1}_{\{P_T>K\}} \right] - K \, \mathrm{E}_t^{\mathbb{Q}^T} \left[\mathbf{1}_{\{P_T>K\}} \right] \right) \\ &= B_t^T \left(\mathrm{E}_t^{\mathbb{Q}^T} \left[P_T \mathbf{1}_{\{P_T>K\}} \right] - K \mathbb{Q}_t^T (P_T > K) \right) \\ &= B_t^T \, \mathrm{E}_t^{\mathbb{Q}^T} \left[P_T \mathbf{1}_{\{P_T>K\}} \right] - K B_t^T \mathbb{Q}_t^T (P_T > K). \end{aligned} \tag{12.11}$$

Here $\mathbb{Q}_t^T(P_T > K)$ denotes the probability (using the probability measure $\mathbb{Q}^T$) of $P_T > K$ given the information known at time t, i.e. the forward risk-adjusted probability of the option finishing in-the-money.

The term $B_t^T \, \mathrm{E}_t^{\mathbb{Q}^T}[P_T \mathbf{1}_{\{P_T>K\}}]$ is the value at time t of a dividend of $P_T \mathbf{1}_{\{P_T>K\}}$ at time T. For an option on a traded asset with a strictly positive price we can value the same payment using that underlying asset as the numeraire. In terms of the associated risk-adjusted measure $\mathbb{Q}^P$, the time t value of getting a dividend of D_T at time T is $P_t \, \mathrm{E}_t^{\mathbb{Q}^P}[D_T/P_T]$. Using this with $D_T = P_T \mathbf{1}_{\{P_T>K\}}$, we conclude that

$$B_t^T \, \mathrm{E}_t^{\mathbb{Q}^T} \left[P_T \mathbf{1}_{\{P_T>K\}} \right] = P_t \, \mathrm{E}_t^{\mathbb{Q}^P} \left[\mathbf{1}_{\{P_T>K\}} \right] = P_t \mathbb{Q}_t^P (P_T > K).$$

This assumes that the underlying asset pays no dividends in the interval $[t, T]$. Now the call price formula in the following theorem is clear. The put price can be derived analogously or from the put-call parity.

Theorem 12.4 *The price of a European call option on a non-dividend paying asset is given by*

$$C_t = P_t \mathbb{Q}_t^P (P_T > K) - K B_t^T \mathbb{Q}_t^T (P_T > K). \tag{12.12}$$

[1] In the computation we use the fact that the expected value of the indicator of an event is equal to the probability of that event. This follows from the general definition of an expected value, $\mathrm{E}[g(\omega)] = \int_{\omega \in \Omega} g(\omega) f(\omega) \, d\omega$, where $f(\omega)$ is the probability density function of the state ω and the integration is over all possible states. The set of possible states can be divided into two sets, namely the set of states ω for which $P_T > K$ and the set of ω for which $P_T \leq K$. Consequently,

$$\begin{aligned} \mathrm{E}[\mathbf{1}_{\{P_T>K\}}] &= \int_{\omega \in \Omega} \mathbf{1}_{\{P_T>K\}} f(\omega) \, d\omega \\ &= \int_{\omega : P_T > K} 1^\top f(\omega) \, d\omega + \int_{\omega : P_T \leq K} 0^\top f(\omega) \, d\omega \\ &= \int_{\omega : P_T > K} f(\omega) \, d\omega, \end{aligned}$$

which is exactly the probability of the event $P_T > K$.

The price Π_t *of a put option is given as*

$$\Pi_t = KB_t^T \mathbb{Q}_t^T(P_T \leq K) - P_t \mathbb{Q}_t^P(P_T \leq K). \tag{12.13}$$

Both probabilities in (12.12) show the probability of the option finishing in-the-money, but under two different probability measures. To compute the price of the European call option in a concrete model we "just" have to compute these probabilities. In some cases, however, it is easier to work directly on (12.10) or (12.11).

12.3.2 European option prices when the underlying is lognormal

If we assume that the value of the underlying variable at the maturity of the option is lognormally distributed under the forward measure for maturity T, a more explicit option pricing formula can be derived without too much work. For this reason many specific option pricing models build on assumptions leading to P_T being lognormal under the measure $\mathbb{Q}^T$.

If $\ln P_T \sim N(m, v^2)$ under the $\mathbb{Q}^T$-measure conditional on the information at time $t < T$, it follows that

$$\begin{aligned}\mathbb{Q}_t^T(P_T > K) &= \mathbb{Q}_t^T(\ln P_T > \ln K) = \mathbb{Q}_t^T\left(\frac{\ln P_T - m}{v} > \frac{\ln K - m}{v}\right) \\ &= \mathbb{Q}_t^T\left(\frac{\ln P_T - m}{v} < -\frac{\ln K - m}{v}\right) = N\left(\frac{m - \ln K}{v}\right),\end{aligned}$$

where $N(\cdot)$ is the cumulative probability distribution function of a normally distributed random variable with mean zero and variance one. The last equality follows since $(\ln P_T - m)/v \sim N(0,1)$. Moreover, it follows from Theorem B.3 in Appendix B that

$$\mathrm{E}_t^{\mathbb{Q}^T}\left[P_T \mathbf{1}_{\{P_T > K\}}\right] = \mathrm{E}_t^{\mathbb{Q}^T}[P_T] N\left(\frac{m - \ln K}{v} + v\right) = \mathrm{E}_t^{\mathbb{Q}^T}[P_T] N\left(\frac{\ln\left(\mathrm{E}_t^{\mathbb{Q}^T}[P_T]/K\right) + \frac{1}{2}v^2}{v}\right).$$

Substituting these results into (12.11), we get

$$C_t = B_t^T \mathrm{E}_t^{\mathbb{Q}^T}[P_T] N\left(\frac{\ln\left(\mathrm{E}_t^{\mathbb{Q}^T}[P_T]/K\right) + \frac{1}{2}v^2}{v}\right) - KB_t^T N\left(\frac{\ln\left(\mathrm{E}_t^{\mathbb{Q}^T}[P_T]/K\right) - \frac{1}{2}v^2}{v}\right)$$

In the typical case where P is the price of a non-dividend paying asset, we know that $P_t = B_t^T \mathrm{E}_t^{\mathbb{Q}^T}[P_T]$. Let us identify the relevant v^2. By convention the forward price of the underlying at time T for immediate delivery is $F_T^T = P_T$ so we can focus on the dynamics of the forward price $F_t^T = P_t/B_t^T$. This is easier since we know that the forward price is a martingale under the measure $\mathbb{Q}^T$ so that the drift is zero. By Itô's Lemma the sensitivity of the forward price is given by the sensitivity of the underlying and the sensitivity of the zero-coupon bond price so for this purpose we can ignore the drift terms in P_t and B_t^T. If we write their $\mathbb{Q}^T$-dynamics as

$$\begin{aligned}dP_t &= P_t\left[\ldots\, dt + \sigma(t)\, dz_{1t}^T\right], \\ dB_t^T &= B_t^T\left[\ldots\, dt + \sigma_B(T-t)\rho\, dz_{1t}^T + \sigma_B(T-t)\sqrt{1-\rho^2}\, dz_{2t}^T\right],\end{aligned}$$

where (z_1^T, z_2^T) is a two-dimensional standard Brownian motion under $\mathbb{Q}^T$, then $\sigma(t)$ is the volatility of the underlying asset, $\sigma_B(T-t)$ is the volatility of the zero-coupon bond price, and ρ is the

correlation between shocks to those two prices. We have assumed that ρ is a constant, that the volatility of the underlying asset $\sigma(t)$ is a deterministic function of time, and that $\sigma_B(\cdot)$ is a deterministic function of the time-to-maturity of the bond since these are the only reasonable assumptions that will lead to P_T being lognormal under $\mathbb{Q}^T$. Now the forward price dynamics will be

$$\begin{aligned} dF_t^T &= \frac{1}{B_t^T}\sigma(t)S_t\,dz_{1t}^T - \frac{S_t}{(B_t^T)^2}B_t^T\left(\sigma_B(T-t)\rho\,dz_{1t}^T + \sigma_B(T-t)\sqrt{1-\rho^2}\,dz_{2t}^T\right) \\ &= F_t^T\left[(\sigma(t)-\rho\sigma_B(T-t))\;dz_{1t}^T - \sqrt{1-\rho^2}\sigma_B(T-t)\,dz_{2t}^T\right], \end{aligned} \tag{12.14}$$

which implies that

$$\ln P_T = \ln F_T^T = \ln F_t^T - \frac{1}{2}v_F(T,t)^2 + \int_t^T (\sigma(u)-\rho\sigma_B(T-u))\;dz_{1u}^T - \int_t^T \sqrt{1-\rho^2}\sigma_B(T-u)\,dz_{2u}^T,$$

where

$$v_F(t,T) = \left(\int_t^T \left(\sigma(u)^2 + \sigma_B(T-u)^2 - 2\rho\sigma(u)\sigma_B(T-u)\right)\,du\right)^{1/2} \tag{12.15}$$

is the volatility of the forward price. Now we see that $\ln P_T \sim N(\ln F_t^T - \frac{1}{2}v_F(t,T)^2, v_F(t,T)^2)$ under $\mathbb{Q}^T$.

We summarize the above results in the following theorem:

Theorem 12.5 *If $\ln P_T$ conditional on time t information is normally distributed with variance v^2 under the forward measure $\mathbb{Q}^T$, the price of a European option maturing at time T is given by*

$$C_t = B_t^T\,\mathrm{E}_t^{\mathbb{Q}^T}[P_T]N(d) - KB_t^T N(d-v), \tag{12.16}$$

where

$$d = \frac{\ln\left(\mathrm{E}_t^{\mathbb{Q}^T}[P_T]/K\right) + \frac{1}{2}v^2}{v}.$$

In particular, if P is the price of a non-dividend paying asset and $\ln P_T$ is normally distributed under $\mathbb{Q}^T$, the call price is

$$C_t = P_t N\left(d(F_t^T,t)\right) - KB_t^T N\left(d(F_t^T,t) - v_F(t,T)\right), \tag{12.17}$$

where $F_t^T = P_t/B_t^T$,

$$d(F_t^T,t) = \frac{\ln\left(F_t^T/K\right) + \frac{1}{2}v_F(t,T)^2}{v_F(t,T)}$$

and $v_F(t,T)$ is given by (12.15).

12.3.3 The Black-Scholes-Merton model for stock option pricing

While option pricing models date back at least to Bachelier (1900), the most famous model is the Black-Scholes-Merton model developed by Black and Scholes (1973) and Merton (1973c) for the pricing of a European option on a stock. The model is formulated in continuous time and assumes that the risk-free interest rate r (continuously compounded) is constant over time and that the price S_t of the underlying stock follows a continuous stochastic process with a constant relative volatility, i.e.

$$dS_t = \mu_t S_t\,dt + \sigma S_t\,dz_t, \tag{12.18}$$

where σ is a constant and μ is a "nice" process. Furthermore, we assume that the underlying stock pays no dividends in the life of the option.

With constant interest rates, $B_t^T = e^{-r(T-t)}$ and the risk-neutral measure is identical to the forward measure, $\mathbb{Q} = \mathbb{Q}^T$. Since the risk-neutral expected rate of return of any asset is equal to the risk-free rate of return, the risk-neutral dynamics of the stock price is

$$dS_t = S_t \left[r\,dt + \sigma\,dz_t^{\mathbb{Q}} \right], \tag{12.19}$$

where $z^{\mathbb{Q}} = (z_t^{\mathbb{Q}})$ is a standard Brownian motion under $\mathbb{Q}$. It follows that the stock price is a geometric Brownian motion under $\mathbb{Q} = \mathbb{Q}^T$ and, in particular, we know from Section 2.6.7 that

$$\ln S_T = \ln S_t + \left(r - \frac{1}{2}\sigma^2 \right)(T-t) + \sigma(z_T^{\mathbb{Q}} - z_t^{\mathbb{Q}}).$$

Hence S_T is lognormal and $\mathrm{Var}_t^{\mathbb{Q}}[\ln S_T] = \sigma^2(T-t)$. We can apply Theorem 12.5 and since $\sigma_B(T-t) = 0$ with constant interest rates, $v_F = \sigma\sqrt{T-t}$. The forward price of the stock is $F_t^T = S_t e^{r(T-t)}$. We summarize in the following theorem:

Theorem 12.6 (Black-Scholes-Merton) *Assume that the stock pays no dividend, the stock price dynamics is of the form* (12.18) *and the short-term risk-free rate is constant. Then the price of a European call option on the stock is given by*

$$C_t = S_t N\left(d(S_t,t)\right) - K e^{-r(T-t)} N\left(d(S_t,t) - \sigma\sqrt{T-t} \right), \tag{12.20}$$

where

$$d(S_t,t) = \frac{\ln(S_t/K) + \left(r + \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}}.$$

Equation (12.20) is the famous Black-Scholes-Merton equation.

Alternatively, we can derive the above result using (12.12) which implies that the price of a European call option on a stock is given by

$$C_t = S_t \mathbb{Q}_t^S(S_T > K) - K B_t^T \mathbb{Q}_t(S_T > K),$$

where $\mathbb{Q}^S$ is the risk-adjusted measure associated with the underlying stock. With the risk-neutral dynamics (12.19), we have

$$\begin{aligned}
\mathbb{Q}_t(S_T > K) &= \mathbb{Q}_t(\ln S_T > \ln K) \\
&= \mathbb{Q}_t\left(\ln S_t + \left(r - \frac{1}{2}\sigma^2 \right)(T-t) + \sigma(z_T^{\mathbb{Q}} - z_t^{\mathbb{Q}}) > \ln K \right) \\
&= \mathbb{Q}_t\left(\frac{z_T^{\mathbb{Q}} - z_t^{\mathbb{Q}}}{\sqrt{T-t}} > -\frac{\ln(S_t/K) + \left(r - \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \right) \\
&= \mathbb{Q}_t\left(\frac{z_T^{\mathbb{Q}} - z_t^{\mathbb{Q}}}{\sqrt{T-t}} < \frac{\ln(S_t/K) + \left(r - \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \right) \\
&= N\left(\frac{\ln(S_t/K) + \left(r - \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}} \right).
\end{aligned}$$

According to (11.32), the dynamics of the stock price under the measure $\mathbb{Q}^S$ is

$$dS_t = S_t \left[\left(r + \sigma^2 \right) dt + \sigma\,dz_t^S \right] \tag{12.21}$$

so that S is also a geometric Brownian motion under the measure $\mathbb{Q}^S$. Analogously to the above equations it can be shown that

$$\mathbb{Q}_t^S(S_T > K) = N\left(\frac{\ln(S_t/K) + \left(r + \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}}\right).$$

Now the option price in (12.20) follows.

The Black-Scholes-Merton equation states the call option price in terms of five quantities:

(1) the price of the underlying stock,

(2) the price of the zero-coupon bond maturing at expiry of the option (or, equivalently, the risk-free interest rate),

(3) the time-to-expiration of the option,

(4) the exercise price (or, equivalently, the *moneyness* S_t/K of the option),

(5) the volatility of the underlying stock.

It can be shown (you are asked to do that in Exercise 12.2) by straightforward differentiation that

$$\frac{\partial C_t}{\partial S_t} = N\left(d(S_t,t)\right), \quad \frac{\partial^2 C_t}{\partial S_t^2} = \frac{n\left(d(S_t,t)\right)}{S_t\sigma\sqrt{T-t}}, \tag{12.22}$$

where $n(\cdot) = N'(\cdot)$ is the probability density function of a $N(0,1)$ random variable, and

$$\frac{\partial C_t}{\partial t} = -\frac{-S_t\sigma n\left(d(S_t,t)\right)}{2\sqrt{T-t}} - rKB_t^T N\left(d(S_t,t) - \sigma\sqrt{T-t}\right), \tag{12.23}$$

using that $B_t^T = \exp\{-r(T-t)\}$. In particular, the call option price is an increasing, convex function of the price of the underlying stock. The call price is increasing in the volatility σ, (obviously) decreasing in the exercise price K, and increasing in the zero-coupon bond price (and, hence, decreasing in the risk-free rate r).

Note that the Black-Scholes-Merton price of the call option does not involve any preference parameters or a market price of risk associated with the shock z to the underlying stock price. On the other hand, it involves the price of the underlying stock. In the Black-Scholes-Merton model the option is a redundant asset. There is only one source of risk and with the stock and the risk-free asset, the market is already complete. Any additional asset affected only by the same shock will be redundant.

The option can be perfectly replicated by a trading strategy in the stock and the risk-free asset. At any time $t < T$, the portfolio consists of $\theta_t^S = N\left(d(S_t,t)\right) \in (0,1)$ units of the stock and $\theta_t^B = -KN\left(d(S_t,t) - \sigma\sqrt{T-t}\right) \in (-K,0)$ units of the zero-coupon bond maturing at time T. Clearly, the value of this portfolio is identical to the value of the call option, $\theta_t^S S_t + \theta_t^B B_t^T = C_t$. Since $C_t = C(S_t,t)$, it follows from Itô's Lemma and the derivatives of C computed above that the dynamics of the call price is

$$\begin{aligned} dC_t &= \frac{\partial C_t}{\partial t}\,dt + \frac{\partial C_t}{\partial S_t}\,dS_t + \frac{1}{2}\frac{\partial^2 C_t}{\partial S_t^2}(dS_t)^2 \\ &= \left(N\left(d(S_t,t)\right)\mu(S_t,t) - rKB_t^T N\left(d(S_t,t) - \sigma\sqrt{T-t}\right)\right)dt + N\left(d(S_t,t)\right)\sigma S_t\,dz_t. \end{aligned} \tag{12.24}$$

The dynamics of the trading strategy is

$$\begin{aligned}\theta_t^S\, dS_t + \theta_t^B\, dB_t^T &= N\left(d(S_t,t)\right)\, dS_t - KN\left(d(S_t,t) - \sigma\sqrt{T-t}\right)\, dB_t^T \\ &= \left(N\left(d(S_t,t)\right)\mu(S_t,t) - rKB_t^T N\left(d(S_t,t) - \sigma\sqrt{T-t}\right)\right)\, dt + N\left(d(S_t,t)\right)\sigma S_t\, dz_t,\end{aligned} \tag{12.25}$$

which is identical to dC_t. The trading strategy is therefore replicating the option.

Applying the put-call parity (12.9), we obtain a European put option price of

$$\Pi(S_t,t) = KB_t^T N\left(-[d(S_t,t) - \sigma\sqrt{T-t}]\right) - S_t N\left(-d(S_t,t)\right). \tag{12.26}$$

In the above model, interest rates and hence bond price were assumed to be constant. However, we can easily generalize to the case where bond prices vary stochastically with a deterministic volatility and a constant correlation with the stock price. Assuming that the dynamics of the stock price and the price of the zero-coupon bond maturing at time T are given by

$$\begin{aligned}dS_t &= S_t\left[\ldots\, dt + \sigma\, dz_{1t}\right], \\ dB_t^T &= B_t^T\left[\ldots\, dt + \sigma_B(T-t)\rho\, dz_{1t} + \sigma_B(T-t)\sqrt{1-\rho^2}\, dz_{2t}\right],\end{aligned}$$

we are still in the setting of Section 12.3.2 and can apply Theorem 12.5.

Theorem 12.7 *Suppose the stock pays no dividend, has a constant volatility σ, and a constant correlation ρ with the price of the zero-coupon bond maturing at time T, and suppose that this bond has a deterministic volatility $\sigma_B(T-t)$. Then the price of a European call option on the stock is given by*

$$C_t = S_t N\left(d(F_t^T,t)\right) - KB_t^T N\left(d(F_t^T,t) - v_F(t,T)\right), \tag{12.27}$$

where $F_t^T = S_t/B_t^T$,

$$\begin{aligned}d(F_t^T,t) &= \frac{\ln(F_t^T/K) + \frac{1}{2}v_F(t,T)^2}{v_F(t,T)}, \\ v_F(t,T) &= \left(\int_t^T \left(\sigma^2 + \sigma_B(T-u)^2 - 2\rho\sigma\sigma_B(T-u)\right)\, du\right)^{1/2}.\end{aligned}$$

In practice $\sigma_B(T-t)$ is typically much smaller than σ and the approximation

$$v_F(t,T) \approx \sqrt{\int_t^T \sigma^2\, du} = \sigma\sqrt{T-t}$$

is not too bad. With that approximation you will get the same call price using (12.27) as by using the Black-Scholes-Merton equation (12.20) with the zero-coupon yield $y_t^T = -(\ln B_t^T)/(T-t)$ replacing r. In this sense the above theorem supports the use of the Black-Scholes-Merton equation even when interest rates are stochastic. Note, however, that the above theorem requires the bond price volatility to be a deterministic function of the time-to-maturity of the bond. This will only be satisfied if the short-term risk-free interest rate r_t follows a Gaussian process, e.g. an Ornstein-Uhlenbeck process as assumed in the Vasicek model introduced in Section 10.5.1. While such models are very nice to work with, they are not terribly realistic. On the other hand, for short maturities and relatively stable interest rates, it is probably reasonable to approximate the bond

price volatility with a deterministic function or even approximate it with zero as the Black-Scholes-Merton model implicitly does.

The assumption of a constant stock price volatility is important for the derivations of the Black-Scholes-Merton option pricing formula. Alas, it is not realistic. The volatility of a stock can be estimated from historical variations in the stock price and the estimate varies with the time period used in the estimation—both over short periods and long periods. Another measure of the volatility of a stock is its **implied volatility**. Given the current stock price S_t and interest rate r, we can define an implied volatility of the stock for any option traded upon that stock (i.e. for any exercise price K and any maturity T) as the value of σ you need to plug into the Black-Scholes-Merton formula to get a match with the observed market price of the option. Since the Black-Scholes-Merton option price is an increasing function of the volatility, there will be a unique value of σ that does the job. Looking at simultaneous prices of different options on the same stock, the implied volatility is found to vary with the exercise price and the maturity of the option. If the Black-Scholes-Merton assumptions were correct, you would find the same implied volatility for all options on the same underlying.

Various alternatives to the constant volatility assumption have been proposed. Black and Cox (1976) replace the constant σ by σS_t^α for some power α. Here the volatility is a function of the stock price and therefore perfectly correlated with the stock price. This extension does not seem sufficient. The stochastic volatility models of Hull and White (1987) and Heston (1993) allow the volatility to be affected by another exogenous shock than the shock to the stock price itself. With these extensions it is possible to match the option prices in the model with observed option prices. In the stochastic volatility models, the market is no longer complete, and the option prices will depend on the market price of volatility risk, which then has to be specified and estimated.

Occasionally stock prices change a lot over a very short period of time, for example in the so-called stock market crashes. Such dramatic variations are probably better modeled by jump processes than by pure diffusion models like those discussed in this book. Several papers study the pricing of options on stocks, when the stock price can jump. Some prominent examples are Merton (1976) and Madan, Carr, and Chang (1998).

12.3.4 Options on bonds

Consider a European call option on a zero-coupon bond. Let T be the maturity of the option and $T^* > T$ the maturity of the bond. K is the exercise price. Let C_t^{K,T,T^*} denote the price at time t of a European call option on this zero-coupon bond. The dividend of the option at time T is

$$C_T^{K,T,T^*} = \max\left(B_T^{T^*} - K, 0\right).$$

The option price is generally characterized by

$$C_t^{K,T,T^*} = B_t^{T^*}\mathbb{Q}_t^{T^*}\left(B_T^{T^*} > K\right) - KB_t^T\mathbb{Q}_t^T\left(B_T^{T^*} > K\right), \tag{12.28}$$

where $\mathbb{Q}^{T^*}$ and $\mathbb{Q}^T$ are the forward measures for maturities T^* and T, respectively.

If $B_T^{T^*}$ is lognormally distributed under the forward measure for maturity T, we know from Theorem 12.5 that we can find a nice closed-form solution. This is for example the case in the

Vasicek model introduced in Section 10.5.1. Given the Ornstein-Uhlenbeck process for the short-term risk-free rate,

$$dr_t = \kappa\,(\bar{r} - r_t)\;dt + \sigma_r\,dz_t, \tag{10.28}$$

and a constant market price of interest rate risk λ, the price of a zero-coupon bond price maturing at time s is

$$B_t^s = e^{-a(s-t)-b(s-t)r_t}, \tag{10.34}$$

cf. (10.34), where $a(\cdot)$ and $b(\cdot)$ are defined in (10.35) and (10.33), respectively. The change to the forward measure requires identification of the bond price sensitivity. An application of Itô's Lemma gives the bond price dynamics

$$dB_t^s = B_t^s\,[\ldots\,dt - \sigma_r b(s-t)\,dz_t]$$

so that the bond price sensitivity is $\sigma_B(s-t) = -\sigma_r b(s-t)$. (Since this is negative, the bond price volatility is $-\sigma_B(s-t) = \sigma_r b(s-t)$.) It now follows from (11.28) that the $\mathbb{Q}^T$-dynamics of the short-term interest rate is

$$\begin{aligned} dr_t &= (\kappa[\bar{r} - r_t] - \sigma_r[\lambda - \sigma_B(T-t)])\;dt + \sigma_r\,dz_t^T \\ &= \kappa\,(\tilde{r}(T-t) - r_t)\;dt + \sigma_r\,dz_t^T, \end{aligned}$$

where $\tilde{r}(\tau) = \bar{r} - \sigma_r\lambda/\kappa - \sigma_r^2 b(\tau)/\kappa$ and $z^T = (z_t^T)$ is a standard Brownian motion under $\mathbb{Q}^T$. Under the $\mathbb{Q}^T$-measure, the short rate behaves as an Ornstein-Uhlenbeck process but with a deterministically changing mean-reversion level $\tilde{r}(T-t)$. Hence, r_T will be normally distributed under $\mathbb{Q}^T$ and, consequently, the price of the underlying zero-coupon bond at the maturity of the option, $B_T^{T^*} = \exp\{-a(T^*-T) - b(T^*-T)r_T\}$, will be lognormally distributed under $\mathbb{Q}^T$. We can therefore apply Theorem 12.5 and conclude that the price of the option is

$$C_t^{K,T,T^*} = B_t^{T^*} N(d) - K B_t^T N\left(d - v_F(t,T,T^*)\right), \tag{12.29}$$

where

$$d = \frac{\ln\left(\frac{B_t^{T^*}}{KB_t^T}\right) + \frac{1}{2}v_F(t,T,T^*)^2}{v_F(t,T,T^*)}$$

and, using the fact that the underlying zero-coupon bond price is perfectly correlated with the price of the zero-coupon bond maturing at T,

$$\begin{aligned} v_F(t,T,T^*)^2 &= \int_t^T \left(\sigma_B(T^*-u) - \sigma_B(T-u)\right)^2\,du \\ &= \sigma_r^2 \int_t^T \left(b(T^*-u) - b(T-u)\right)^2\,du \\ &= \frac{\sigma_r^2}{\kappa^2}\int_t^T \left(e^{-\kappa(T-u)} - e^{-\kappa(T^*-u)}\right)^2\,du \\ &= \frac{\sigma_r^2}{\kappa^3}\left(1 - e^{-\kappa(T^*-T)}\right)^2\left(1 - e^{-2\kappa(T-t)}\right). \end{aligned}$$

In many other models of interest rates and bond prices, an option pricing formula very similar to (12.29) can be derived. For example, in the Cox-Ingersoll-Ross model introduced in Section 10.5.2, the price of a European call option on a zero-coupon bond is of the form

$$C_t^{K,T,T^*} = B_t^{T^*}\chi^2(h_1; f, g_1) - K B_t^T \chi^2(h_2; f, g_2),$$

where $\chi^2(\cdot; f, g)$ is the cumulative probability distribution function of a non-centrally χ^2-distributed random variable with f degrees of freedom and non-centrally parameter g. For details see, e.g., Munk (2005b, Ch. 7).

The options considered above are options on zero-coupon bonds. Traded bond options usually have a coupon bond as the underlying. Fortunately, the pricing formulas for options on zero-coupon bonds can, under some assumptions, be used in the pricing of options on coupon bonds. First some notation. The underlying coupon bond is assumed to pay Y_i at time T_i $(i = 1, 2, \ldots, n)$, where $T_1 < T_2 < \cdots < T_n$, so that the price of the bond is

$$B_t = \sum_{T_i > t} Y_i B_t^{T_i},$$

where we sum over all the future payment dates. Let $C_t^{K,T,\text{cpn}}$ denote the price at time t of a European call option on the coupon bond, where K is the exercise price and T is the expiration date of the option. In reasonable one-factor models, the price of a given zero-coupon bond will be a decreasing function of the short-term interest rate. In both the Vasicek model and the Cox-Ingersoll-Ross model the zero-coupon bond price is of the form $B_t^T = \exp\{-a(T-t) - b(T-t)r_t\}$ and since the b-function is positive in both models, the bond price is indeed decreasing in maturity. Then the following result, first derived by Jamshidian (1989), applies:

Theorem 12.8 *Suppose that the zero-coupon bond prices are of the form $B_t^T = B^T(r_t, t)$ and B^T is decreasing in r_t. Then the price of a European call on a coupon bond is*

$$C_t^{K,T,cpn} = \sum_{T_i > T} Y_i C_t^{K_i,T,T_i}, \tag{12.30}$$

where $K_i = B^{T_i}(r^*, T)$, and r^* is defined as the solution to the equation

$$B(r^*, T) \equiv \sum_{T_i > T} Y_i B^{T_i}(r^*, T) = K. \tag{12.31}$$

Proof: The payoff of the option on the coupon bond is

$$\max(B(r_T, T) - K, 0) = \max\left(\sum_{T_i > T} Y_i B^{T_i}(r_T, T) - K, 0 \right).$$

Since the zero-coupon bond price $B^{T_i}(r_T, T)$ is a monotonically decreasing function of the interest rate r_T, the whole sum $\sum_{T_i>T} Y_i B^{T_i}(r_T, T)$ is monotonically decreasing in r_T. Therefore, exactly one value r^* of r_T will make the option finish *at the money* so that (12.31) holds. Letting $K_i = B^{T_i}(r^*, T)$, we have that $\sum_{T_i>T} Y_i K_i = K$.

For $r_T < r^*$,

$$\sum_{T_i > T} Y_i B^{T_i}(r_T, T) > \sum_{T_i > T} Y_i B^{T_i}(r^*, T) = K, \quad B^{T_i}(r_T, T) > B^{T_i}(r^*, T) = K_i,$$

so that

$$\begin{aligned} \max\left(\sum_{T_i > T} Y_i B^{T_i}(r_T, T) - K, 0 \right) &= \sum_{T_i > T} Y_i B^{T_i}(r_T, T) - K \\ &= \sum_{T_i > T} Y_i \left(B^{T_i}(r_T, T) - K_i \right) \\ &= \sum_{T_i > T} Y_i \max\left(B^{T_i}(r_T, T) - K_i, 0 \right). \end{aligned}$$

For $r_T \geq r^*$,

$$\sum_{T_i>T} Y_i B^{T_i}(r_T, T) \leq \sum_{T_i>T} Y_i B^{T_i}(r^*, T) = K, \quad B^{T_i}(r_T, T) \leq B^{T_i}(r^*, T) = K_i,$$

so that

$$\max\left(\sum_{T_i>T} Y_i B^{T_i}(r_T, T) - K, 0\right) = 0 = \sum_{T_i>T} Y_i \max\left(B^{T_i}(r_T, T) - K_i, 0\right).$$

Hence, for *all* possible values of r_T we may conclude that

$$\max\left(\sum_{T_i>T} Y_i B^{T_i}(r_T, T) - K, 0\right) = \sum_{T_i>T} Y_i \max\left(B^{T_i}(r_T, T) - K_i, 0\right).$$

The payoff of the option on the coupon bond is thus identical to the payoff of a portfolio of options on zero-coupon bonds, namely a portfolio consisting (for each i with $T_i > T$) of Y_i options on a zero-coupon bond maturing at time T_i and an exercise price of K_i. Consequently, the value of the option on the coupon bond at time $t \leq T$ equals the value of that portfolio of options on zero-coupon bonds. A formal derivation goes as follows:

$$\begin{aligned}
C_t^{K,T,\text{cpn}} &= \mathrm{E}_{r,t}^{\mathbb{Q}}\left[e^{-\int_t^T r_u\,du} \max\left(B(r_T, T) - K, 0\right)\right] \\
&= \mathrm{E}_{r,t}^{\mathbb{Q}}\left[e^{-\int_t^T r_u\,du} \sum_{T_i>T} Y_i \max\left(B^{T_i}(r_T, T) - K_i, 0\right)\right] \\
&= \sum_{T_i>T} Y_i\, \mathrm{E}_{r,t}^{\mathbb{Q}}\left[e^{-\int_t^T r_u\,du} \max\left(B^{T_i}(r_T, T) - K_i, 0\right)\right] \\
&= \sum_{T_i>T} Y_i C_t^{K_i,T,T_i},
\end{aligned}$$

which completes the proof. □

To compute the price of a European call option on a coupon bond we must numerically solve one equation in one unknown (to find r^*) and calculate n' prices of European call options on zero-coupon bonds, where n' is the number of payment dates of the coupon bond after expiration of the option. For example, in the Vasicek model we can use (12.29).

Practitioners often use Black-Scholes-Merton type formulas for pretty much all types of options, including options on bonds. The formulas are based on the Black (1976) variant of the Black-Scholes-Merton model developed for stock option pricing, originally developed for options on futures on an asset with a lognormally distributed value. Black's formula for a European call option on a bond is

$$\begin{aligned}
C_t^{K,T,\text{cpn}} &= B_t^T\left[F_t^{T,\text{cpn}} N\left(d(F_t^{T,\text{cpn}}, t)\right) - K N\left(d(F_t^{T,\text{cpn}}, t) - \sigma_B\sqrt{T-t}\right)\right], \\
&= \tilde{B}_t N\left(d(F_t^{T,\text{cpn}}, t)\right) - K B_t^T N\left(d(F_t^{T,\text{cpn}}, t) - \sigma_B\sqrt{T-t}\right),
\end{aligned} \tag{12.32}$$

where σ_B is the volatility of the bond, $F_t^{T,\text{cpn}} = \tilde{B}_t / B_t^T$ is the forward price of the bond, $\tilde{B}_t = B_t - \sum_{t<T_i<T} Y_i B_t^{T_i}$ is the present value of the bond payments after maturity of the option, and

$$d(F_t^{T,\text{cpn}}, t) = \frac{\ln(F_t^{T,\text{cpn}}/K)}{\sigma\sqrt{T-t}} + \frac{1}{2}\sigma_B\sqrt{T-t}.$$

The use of Black's formula for bond options is not theoretically supported and may lead to prices allowing arbitrage. At best, it is a reasonable approximation to the correct price.

12.3.5 Interest rate options: caps and floors

An **(interest rate) cap** is designed to protect an investor who has borrowed funds on a floating interest rate basis against the risk of paying very high interest rates. Suppose the loan has a face value of H and payment dates $T_1 < T_2 < \cdots < T_n$, where $T_{i+1} - T_i = \delta$ for all i.[2] The interest rate to be paid at time T_i is determined by the δ-period money market interest rate prevailing at time $T_{i-1} = T_i - \delta$, i.e. the payment at time T_i is equal to $H\delta l^{T_i}_{T_i-\delta}$, cf. the notation for interest rates introduced in Section 10.2. Note that the interest rate is set at the beginning of the period, but paid at the end. Define $T_0 = T_1 - \delta$. The dates $T_0, T_1, \ldots, T_{n-1}$ where the rate for the coming period is determined are called the **reset dates** of the loan.

A cap with a face value of H, payment dates T_i $(i = 1, \ldots, n)$ as above, and a so-called cap rate K yields a time T_i payoff of $H\delta \max(l^{T_i}_{T_i-\delta} - K, 0)$, for $i = 1, 2, \ldots, n$. If a borrower buys such a cap, the net payment at time T_i cannot exceed $H\delta K$. The period length δ is often referred to as the **frequency** or the **tenor** of the cap.[3] In practice, the frequency is typically either 3, 6, or 12 months. Note that the time distance between payment dates coincides with the "maturity" of the floating interest rate. Also note that while a cap is tailored for interest rate hedging, it can also be used for interest rate speculation.

A cap can be seen as a portfolio of n **caplets**, namely one for each payment date of the cap. The i'th caplet yields a payoff at time T_i of

$$\mathcal{C}^i_{T_i} = H\delta \max\left(l^{T_i}_{T_i-\delta} - K, 0\right) \tag{12.33}$$

and no other payments. A caplet is a call option on the zero-coupon yield prevailing at time $T_i - \delta$ for a period of length δ, but where the payment takes place at time T_i although it is already fixed at time $T_i - \delta$.

In the following we will find the value of the i'th caplet before time T_i. Since the payoff becomes known at time $T_i - \delta$, we can obtain its value in the interval between $T_i - \delta$ and T_i by a simple discounting of the payoff, i.e.

$$\mathcal{C}^i_t = B^{T_i}_t H\delta \max\left(l^{T_i}_{T_i-\delta} - K, 0\right), \quad T_i - \delta \le t \le T_i.$$

In particular,

$$\mathcal{C}^i_{T_i-\delta} = B^{T_i}_{T_i-\delta} H\delta \max\left(l^{T_i}_{T_i-\delta} - K, 0\right). \tag{12.34}$$

To find the value before the fixing of the payoff, i.e. for $t < T_i - \delta$, we shall use two strategies. The first is simply to take relevant expectations of the payoff. Since the payoff comes at T_i, we know from Section 11.4 that the value of the payoff can be found as the product of the expected payoff computed under the T_i-forward martingale measure and the current discount factor for time T_i payments, i.e.

$$\mathcal{C}^i_t = H\delta B^{T_i}_t \, \mathrm{E}^{\mathbb{Q}^{T_i}}_t \left[\max\left(l^{T_i}_{T_i-\delta} - K, 0\right)\right], \quad t < T_i - \delta. \tag{12.35}$$

The price of a cap can therefore be determined as

$$\mathcal{C}_t = H\delta \sum_{i=1}^{n} B^{T_i}_t \, \mathrm{E}^{\mathbb{Q}^{T_i}}_t \left[\max\left(l^{T_i}_{T_i-\delta} - K, 0\right)\right], \quad t < T_0. \tag{12.36}$$

[2] In practice, there will not be exactly the same number of days between successive reset dates, and the calculations below must be slightly adjusted by using the relevant *day count convention*.

[3] The word tenor is sometimes used for the set of payment dates $T_1, \ldots, T_n$.

If each LIBOR rate $l_{T_i-\delta}^{T-i}$ is lognormally distributed under the $\mathbb{Q}^{T_i}$-forward measure, we can obtain a nice closed-form pricing formula. This is satisfied in the so-called LIBOR market model introduced by Miltersen, Sandmann, and Sondermann (1997) and Brace, Gatarek, and Musiela (1997). See Munk (2005b, Ch. 11) for a review. In fact, the resulting pricing formula is the Black formula often applied in practice:

$$\mathcal{C}_t^i = H\delta B_t^{T_i}\left[L_t^{T_i-\delta,T_i} N\left(d^i(L_t^{T_i-\delta,T_i},t)\right) - KN\left(d^i(L_t^{T_i-\delta,T_i},t) - \sigma_i\sqrt{T_i-\delta-t}\right)\right], \quad t < T_i-\delta, \tag{12.37}$$

where σ_i is the (relative) volatility of the forward LIBOR rate $L_t^{T_i-\delta,T_i}$, and d^i is given by

$$d^i(L_t^{T_i-\delta,T_i},t) = \frac{\ln(L_t^{T_i-\delta,T_i}/K)}{\sigma_i\sqrt{T_i-\delta-t}} + \frac{1}{2}\sigma_i\sqrt{T_i-\delta-t}.$$

The second pricing strategy links caps to bond options. Applying (10.5) on page 222, we can rewrite (12.34) as

$$\begin{aligned}
\mathcal{C}_{T_i-\delta}^i &= B_{T_i-\delta}^{T_i} H\max\left(1+\delta l_{T_i-\delta}^{T_i} - [1+\delta K], 0\right) \\
&= B_{T_i-\delta}^{T_i} H\max\left(\frac{1}{B_{T_i-\delta}^{T_i}} - [1+\delta K], 0\right) \\
&= H(1+\delta K)\max\left(\frac{1}{1+\delta K} - B_{T_i-\delta}^{T_i}, 0\right).
\end{aligned}$$

We can now see that the value at time $T_i-\delta$ is identical to the payoff of a European put option expiring at time $T_i-\delta$ that has an exercise price of $1/(1+\delta K)$ and is written on a zero-coupon bond maturing at time T_i. Accordingly, the value of the i'th caplet at an earlier point in time $t \le T_i-\delta$ must equal the value of that put option, i.e.

$$\mathcal{C}_t^i = H(1+\delta K)\Pi_t^{(1+\delta K)^{-1},T_i-\delta,T_i}. \tag{12.38}$$

To find the value of the entire cap contract we simply have to add up the values of all the caplets corresponding to the remaining payment dates of the cap. Before the first reset date, T_0, none of the cap payments are known, so the value of the cap is given by

$$\mathcal{C}_t = \sum_{i=1}^n \mathcal{C}_t^i = H(1+\delta K)\sum_{i=1}^n \Pi_t^{(1+\delta K)^{-1},T_i-\delta,T_i}, \quad t < T_0. \tag{12.39}$$

At all dates after the first reset date, the next payment of the cap will already be known. If we use the notation $T_{i(t)}$ for the nearest following payment date after time t, the value of the cap at any time t in $[T_0,T_n]$ (exclusive of any payment received exactly at time t) can be written as

$$\begin{aligned}
\mathcal{C}_t = {} & HB_t^{T_{i(t)}}\delta\max\left(l_{T_{i(t)}-\delta}^{T_{i(t)}} - K, 0\right) \\
& + (1+\delta K)H\sum_{i=i(t)+1}^n \Pi_t^{(1+\delta K)^{-1},T_i-\delta,T_i}, \quad T_0 \le t \le T_n.
\end{aligned} \tag{12.40}$$

If $T_{n-1} < t < T_n$, we have $i(t) = n$, and there will be no terms in the sum, which is then considered to be equal to zero. In various models of interest rate dynamics, nice pricing formulas for European options on zero-coupon bonds can be derived. This is for example the case in the Vasicek model studied above. Cap prices will then follow from prices of European puts on zero-coupon bonds.

Note that the interest rates and the discount factors appearing in the expressions above are taken from the money market, not from the government bond market. Also note that since caps and most other contracts related to money market rates trade OTC, one should take the default risk of the two parties into account when valuing the cap. Here, default simply means that the party cannot pay the amounts promised in the contract. Official money market rates and the associated discount function apply to loan and deposit arrangements between large financial institutions, and thus they reflect the default risk of these corporations. If the parties in an OTC transaction have a default risk significantly different from that, the discount rates in the formulas should be adjusted accordingly. However, it is quite complicated to do that in a theoretically correct manner, so we will not discuss this issue any further at this point.

An **(interest rate) floor** is designed to protect an investor who has lent funds on a floating rate basis against receiving very low interest rates. The contract is constructed just as a cap except that the payoff at time T_i $(i = 1, \ldots, n)$ is given by

$$\mathcal{F}^i_{T_i} = H\delta \max\left(K - l^{T_i}_{T_i-\delta}, 0\right), \tag{12.41}$$

where K is called the floor rate. Buying an appropriate floor, an investor who has provided another investor with a floating rate loan will in total at least receive the floor rate. Of course, an investor can also speculate in low future interest rates by buying a floor. The (hypothetical) contracts that only yield one of the payments in (12.41) are called **floorlets**. Obviously, we can think of a floorlet as a European put on the floating interest rate with delayed payment of the payoff.

Analogously to the analysis for caps, we can price the floor directly as

$$\mathcal{F}_t = H\delta \sum_{i=1}^{n} B^{T_i}_t \, \mathrm{E}^{\mathbb{Q}^{T_i}}_t \left[\max\left(K - L^{T_i-\delta,T_i}_{T_i-\delta}, 0\right)\right], \quad t < T_0. \tag{12.42}$$

Again a pricing formula consistent with the Black formula is obtained assuming lognormally distributed forward LIBOR rates. Alternatively, we can express the floorlet as a European call on a zero-coupon bond, and hence a floor is equivalent to a portfolio of European calls on zero-coupon bonds. More precisely, the value of the i'th floorlet at time $T_i - \delta$ is

$$\mathcal{F}^i_{T_i-\delta} = H(1 + \delta K) \max\left(B^{T_i}_{T_i-\delta} - \frac{1}{1+\delta K}, 0\right). \tag{12.43}$$

The total value of the floor contract at any time $t < T_0$ is therefore given by

$$\mathcal{F}_t = H(1 + \delta K) \sum_{i=1}^{n} C_t^{(1+\delta K)^{-1}, T_i-\delta, T_i}, \quad t < T_0, \tag{12.44}$$

and later the value is

$$\begin{aligned} \mathcal{F}_t = {} & H B^{T_{i(t)}}_t \delta \max\left(K - l^{T_{i(t)}}_{T_{i(t)}-\delta}, 0\right) \\ & + (1 + \delta K) H \sum_{i=i(t)+1}^{n} C_t^{(1+\delta K)^{-1}, T_i-\delta, T_i}, \quad T_0 \le t \le T_n. \end{aligned} \tag{12.45}$$

12.4 Interest rate swaps and swaptions

12.4.1 Interest rate swaps

Many different types of swaps are traded on the OTC markets, e.g. currency swaps, credit swaps, asset swaps, but we focus here on interest rate swaps. An **(interest rate) swap** is an exchange

of two cash flow streams that are determined by certain interest rates. In the simplest and most common interest rate swap, a *plain vanilla* swap, two parties exchange a stream of fixed interest rate payments and a stream of floating interest rate payments. The payments are in the same currency and are computed from the same (hypothetical) face value or notional principal. The floating rate is usually a money market rate, e.g. a LIBOR rate, possibly augmented or reduced by a fixed margin. The fixed interest rate is usually set so that the swap has zero net present value when the parties agree on the contract. While the two parties can agree upon any maturity, most interest rate swaps have a maturity between 2 and 10 years.

Let us briefly look at the uses of interest rate swaps. An investor can transform a floating rate loan into a fixed rate loan by entering into an appropriate swap, where the investor receives floating rate payments (netting out the payments on the original loan) and pays fixed rate payments. This is called a **liability transformation**. Conversely, an investor who has lent money at a floating rate, i.e. owns a floating rate bond, can transform this to a fixed rate bond by entering into a swap, where he pays floating rate payments and receives fixed rate payments. This is an **asset transformation**. Hence, interest rate swaps can be used for hedging interest rate risk on both (certain) assets and liabilities. On the other hand, interest rate swaps can also be used for taking advantage of specific expectations of future interest rates, i.e. for speculation.

Swaps are often said to allow the two parties to exploit their **comparative advantages** in different markets. Concerning interest rate swaps, this argument presumes that one party has a comparative advantage (relative to the other party) in the market for fixed rate loans, while the other party has a comparative advantage (relative to the first party) in the market for floating rate loans. However, these markets are integrated, and the existence of comparative advantages conflicts with modern financial theory and the efficiency of the money markets. Apparent comparative advantages can be due to differences in default risk premia. For details we refer the reader to the discussion in Hull (2006, Ch. 7).

Next, we will discuss the valuation of swaps. As for caps and floors, we assume that both parties in the swap have a default risk corresponding to the "average default risk" of major financial institutions reflected by the money market interest rates. For a description of the impact on the payments and the valuation of swaps between parties with different default risk, see Duffie and Huang (1996) and Huge and Lando (1999). Furthermore, we assume that the fixed rate payments and the floating rate payments occur at exactly the same dates throughout the life of the swap. This is true for most, but not all, traded swaps. For some swaps, the fixed rate payments only occur once a year, whereas the floating rate payments are quarterly or semi-annual. The analysis below can easily be adapted to such swaps.

In a plain vanilla interest rate swap, one party pays a stream of fixed rate payments and receives a stream of floating rate payments. This party is said to have a pay fixed, receive floating swap or a fixed-for-floating swap or simply a **payer swap**. The counterpart receives a stream of fixed rate payments and pays a stream of floating rate payments. This party is said to have a pay floating, receive fixed swap or a floating-for-fixed swap or simply a **receiver swap**. Note that the names payer swap and receiver swap refer to the fixed rate payments.

We consider a swap with payment dates $T_1, \ldots, T_n$, where $T_{i+1} - T_i = \delta$. The floating interest rate determining the payment at time T_i is the money market (LIBOR) rate $l_{T_i-\delta}^{T_i}$. In the following we assume that there is no fixed extra margin on this floating rate. If there were such an extra

charge, the value of the part of the flexible payments that is due to the extra margin could be computed in the same manner as the value of the fixed rate payments of the swap, see below. We refer to $T_0 = T_1 - \delta$ as the starting date of the swap. As for caps and floors, we call $T_0, T_1, \ldots, T_{n-1}$ the reset dates, and δ the frequency or the tenor. Typical swaps have δ equal to 0.25, 0.5, or 1 corresponding to quarterly, semi-annual, or annual payments and interest rates.

We will find the value of an interest rate swap by separately computing the value of the fixed rate payments (V^{fix}) and the value of the floating rate payments (V^{fl}). The fixed rate is denoted by K. This is a nominal, annual interest rate, so that the fixed rate payments equal $HK\delta$, where H is the notional principal or face value (which is not swapped). The value of the remaining fixed payments is simply

$$V_t^{\text{fix}} = \sum_{i=i(t)}^{n} HK\delta B_t^{T_i} = HK\delta \sum_{i=i(t)}^{n} B_t^{T_i}. \tag{12.46}$$

The floating rate payments are exactly the same as the coupon payments on a floating rate bond with annualized coupon rate $l_{T_i-\delta}^{T_i}$. Immediately after each reset date, the value of such a bond will equal its face value. To see this, first note that immediately after the last reset date $T_{n-1} = T_n - \delta$, the bond is equivalent to a zero-coupon bond with a coupon rate equal to the market interest rate for the last coupon period. By definition of that market interest rate, the time T_{n-1} value of the bond will be exactly equal to the face value H. In mathematical terms, the market discount factor to apply for the discounting of time T_n payments back to time T_{n-1} is $(1+\delta l_{T_{n-1}}^{T_n})^{-1}$, so the time T_{n-1} value of a payment of $H(1+\delta l_{T_{n-1}}^{T_n})$ at time T_n is precisely H. Immediately after the next-to-last reset date T_{n-2}, we know that we will receive a payment of $H\delta l_{T_{n-2}}^{T_{n-1}}$ at time T_{n-1} and that the time T_{n-1} value of the following payment (received at T_n) equals H. We therefore have to discount the sum $H\delta l_{T_{n-2}}^{T_{n-1}} + H = H(1+\delta l_{T_{n-2}}^{T_{n-1}})$ from T_{n-1} back to T_{n-2}. The discounted value is exactly H. Continuing this procedure, we get that immediately after a reset of the coupon rate, the floating rate bond is valued at par. Note that it is crucial for this result that the coupon rate is adjusted to the interest rate considered by the market to be "fair." Suppose we are interested in the value at some time t between T_0 and T_n. Let $T_{i(t)}$ be the nearest following payment date after time t. We know that the following payment at time $T_{i(t)}$ equals $H\delta l_{T_{i(t)-1}}^{T_{i(t)}}$ and that the value at time $T_{i(t)}$ of all the remaining payments will equal H. The value of the bond at time t will then be

$$B_t^{\text{fl}} = H(1+\delta l_{T_{i(t)}-\delta}^{T_{i(t)}})B_t^{T_{i(t)}}, \quad T_0 \le t < T_n. \tag{12.47}$$

This expression also holds at payment dates $t = T_i$, where it results in H, which is the value excluding the payment at that date.

The value of the floating rate bond is the value of both the coupon payments and the final repayment of face value so the value of the coupon payments only must be

$$\begin{aligned} V_t^{\text{fl}} &= H(1+\delta l_{T_{i(t)}-\delta}^{T_{i(t)}})B_t^{T_{i(t)}} - HB_t^{T_n} \\ &= H\delta l_{T_{i(t)}-\delta}^{T_{i(t)}} B_t^{T_{i(t)}} + H\left[B_t^{T_{i(t)}} - B_t^{T_n}\right], \quad T_0 \le t < T_n. \end{aligned}$$

At and before time T_0, the first term is not present, so the value of the floating rate payments is simply

$$V_t^{\text{fl}} = H\left[B_t^{T_0} - B_t^{T_n}\right], \quad t \le T_0. \tag{12.48}$$

We will also develop an alternative expression for the value of the floating rate payments of the swap. The time $T_i - \delta$ value of the coupon payment at time T_i is

$$H\delta l_{T_i-\delta}^{T_i} B_{T_i-\delta}^{T_i} = H\delta \frac{l_{T_i-\delta}^{T_i}}{1+\delta l_{T_i-\delta}^{T_i}},$$

where we have applied (10.5) on page 222. Consider a strategy of buying a zero-coupon bond with face value H maturing at $T_i - \delta$ and selling a zero-coupon bond with the same face value H but maturing at T_i. The time $T_i - \delta$ value of this position is

$$HB_{T_i-\delta}^{T_i-\delta} - HB_{T_i-\delta}^{T_i} = H - \frac{H}{1+\delta l_{T_i-\delta}^{T_i}} = H\delta \frac{l_{T_i-\delta}^{T_i}}{1+\delta l_{T_i-\delta}^{T_i}},$$

which is identical to the value of the floating rate payment of the swap. Therefore, the value of this floating rate payment at any time $t \leq T_i - \delta$ must be

$$H\left(B_t^{T_i-\delta} - B_t^{T_i}\right) = H\delta B_t^{T_i} \frac{\frac{B_t^{T_i-\delta}}{B_t^{T_i}} - 1}{\delta} = H\delta B_t^{T_i} L_t^{T_i-\delta,T_i}, \tag{12.49}$$

where we have applied (10.6) on page 222. Thus, the value at time $t \leq T_i - \delta$ of getting $H\delta l_{T_i-\delta}^{T_i}$ at time T_i is equal to $H\delta B_t^{T_i} L_t^{T_i-\delta,T_i}$, i.e. the unknown future spot rate $l_{T_i-\delta}^{T_i}$ in the payoff is replaced by the current forward rate for $L_t^{T_i-\delta,T_i}$ and then discounted by the current riskfree discount factor $B_t^{T_i}$. The value at time $t > T_0$ of all the remaining floating coupon payments can therefore be written as

$$V_t^{\text{fl}} = H\delta B_t^{T_{i(t)}} l_{T_{i(t)}-\delta}^{T_{i(t)}} + H\delta \sum_{i=i(t)+1}^{n} B_t^{T_i} L_t^{T_i-\delta,T_i}, \quad T_0 \leq t < T_n.$$

At or before time T_0, the first term is not present, so we get

$$V_t^{\text{fl}} = H\delta \sum_{i=1}^{n} B_t^{T_i} L_t^{T_i-\delta,T_i}, \quad t \leq T_0. \tag{12.50}$$

The value of a payer swap is

$$\mathbf{P}_t = V_t^{\text{fl}} - V_t^{\text{fix}},$$

while the value of a receiver swap is

$$\mathbf{R}_t = V_t^{\text{fix}} - V_t^{\text{fl}}.$$

In particular, the value of a payer swap at or before its starting date T_0 can be written as

$$\mathbf{P}_t = H\delta \sum_{i=1}^{n} B_t^{T_i} \left(L_t^{T_i-\delta,T_i} - K\right), \quad t \leq T_0, \tag{12.51}$$

using (12.46) and (12.50), or as

$$\mathbf{P}_t = H\left(\left[B_t^{T_0} - B_t^{T_n}\right] - \sum_{i=1}^{n} K\delta B_t^{T_i}\right), \quad t \leq T_0, \tag{12.52}$$

using (12.46) and (12.48). If we let $Y_i = K\delta$ for $i = 1, \ldots, n-1$ and $Y_n = 1 + K\delta$, we can rewrite (12.52) as

$$\mathbf{P}_t = H\left(B_t^{T_0} - \sum_{i=1}^{n} Y_i B_t^{T_i}\right), \quad t \leq T_0. \tag{12.53}$$

Also note the following relation between a cap, a floor, and a payer swap having the same payment dates and where the cap rate, the floor rate, and the fixed rate in the swap are all identical:

$$\mathcal{C}_t = \mathcal{F}_t + \mathbf{P}_t. \tag{12.54}$$

This follows from the fact that the payments from a portfolio of a floor and a payer swap exactly match the payments of a cap.

The **swap rate** $\tilde{l}^{\delta}_{T_0}$ prevailing at time T_0 for a swap with frequency δ and payments dates $T_i = T_0 + i\delta$, $i = 1, 2, \ldots, n$, is defined as the unique value of the fixed rate that makes the present value of a swap starting at T_0 equal to zero, i.e. $\mathbf{P}_{T_0} = \mathbf{R}_{T_0} = 0$. The swap rate is sometimes called the equilibrium swap rate or the par swap rate. Applying (12.51), we can write the swap rate as

$$\tilde{l}^{\delta}_{T_0} = \frac{\sum_{i=1}^{n} L^{T_i-\delta,T_i}_{T_0} B^{T_i}_{T_0}}{\sum_{i=1}^{n} B^{T_i}_{T_0}},$$

which can also be written as a weighted average of the relevant forward rates:

$$\tilde{l}^{\delta}_{T_0} = \sum_{i=1}^{n} w_i L^{T_i-\delta,T_i}_{T_0}, \tag{12.55}$$

where $w_i = B^{T_i}_{T_0} / \sum_{i=1}^{n} B^{T_i}_{T_0}$. Alternatively, we can let $t = T_0$ in (12.52) yielding

$$\mathbf{P}_{T_0} = H\left(1 - B^{T_n}_{T_0} - K\delta \sum_{i=1}^{n} B^{T_i}_{T_0}\right),$$

so that the swap rate can be expressed as

$$\tilde{l}^{\delta}_{T_0} = \frac{1 - B^{T_n}_{T_0}}{\delta \sum_{i=1}^{n} B^{T_i}_{T_0}}. \tag{12.56}$$

Substituting (12.56) into the expression just above it, the time T_0 value of an agreement to pay a fixed rate K and receive the prevailing market rate at each of the dates $T_1, \ldots, T_n$, can be written in terms of the current swap rate as

$$\begin{aligned} \mathbf{P}_{T_0} &= H\left(\tilde{l}^{\delta}_{T_0}\delta\left(\sum_{i=1}^{n} B^{T_i}_{T_0}\right) - K\delta\left(\sum_{i=1}^{n} B^{T_i}_{T_0}\right)\right) \\ &= \left(\sum_{i=1}^{n} B^{T_i}_{T_0}\right) H\delta\left(\tilde{l}^{\delta}_{T_0} - K\right). \end{aligned} \tag{12.57}$$

A **forward swap** (or deferred swap) is an agreement to enter into a swap with a future starting date T_0 and a fixed rate which is already set. Of course, the contract also fixes the frequency, the maturity, and the notional principal of the swap. The value at time $t \leq T_0$ of a forward payer swap with fixed rate K is given by the equivalent expressions (12.51)–(12.53). The **forward swap rate** $\tilde{L}^{\delta,T_0}_t$ is defined as the value of the fixed rate that makes the forward swap have zero value at time t. The forward swap rate can be written as

$$\tilde{L}^{\delta,T_0}_t = \frac{B^{T_0}_t - B^{T_n}_t}{\delta \sum_{i=1}^{n} B^{T_i}_t} = \frac{\sum_{i=1}^{n} L^{T_i-\delta,T_i}_t B^{T_i}_t}{\sum_{i=1}^{n} B^{T_i}_t}. \tag{12.58}$$

Note that both the swap rate and the forward swap rate depend on the frequency and the maturity of the underlying swap. To indicate this dependence, let $\tilde{l}^{\delta}_t(n)$ denote the time t swap

rate for a swap with payment dates $T_i = t + i\delta$, $i = 1, 2, \ldots, n$. If we depict the swap rate as a function of the maturity, i.e. the function $n \mapsto \tilde{l}_t^\delta(n)$ (only defined for $n = 1, 2, \ldots$), we get a **term structure of swap rates** for the given frequency. Many financial institutions participating in the swap market will offer swaps of varying maturities under conditions reflected by their posted term structure of swap rates. In Exercise 12.7, you are asked to show how the discount factors $B_{T_0}^{T_i}$ can be derived from a term structure of swap rates.

12.4.2 Swaptions

A swaption is an option on a swap. A European **swaption** gives its holder the right, but not the obligation, at the expiry date T_0 to enter into a specific interest rate swap that starts at T_0 and has a given fixed rate K. No exercise price is to be paid if the right is utilized. The rate K is sometimes referred to as the exercise rate of the swaption. We distinguish between a **payer swaption**, which gives the right to enter into a payer swap, and a **receiver swaption**, which gives the right to enter into a receiver swap. As for caps and floors, two different pricing strategies can be taken. One strategy is to link the swaption payoff to the payoff of another well-known derivative. The other strategy is to directly take relevant expectations of the swaption payoff.

Let us first see how we can link swaptions to options on bonds. Let us focus on a European receiver swaption. At time T_0, the value of a receiver swap with payment dates $T_i = T_0 + i\delta$, $i = 1, 2, \ldots, n$, and a fixed rate K is given by

$$\mathbf{R}_{T_0} = H\left(\sum_{i=1}^{n} Y_i B_{T_0}^{T_i} - 1\right),$$

where $Y_i = K\delta$ for $i = 1, \ldots, n-1$ and $Y_n = 1 + K\delta$; cf. (12.53). Hence, the time T_0 payoff of a receiver swaption is

$$\mathcal{R}_{T_0} = \max\left(\mathbf{R}_{T_0} - 0, 0\right) = H \max\left(\sum_{i=1}^{n} Y_i B_{T_0}^{T_i} - 1, 0\right), \tag{12.59}$$

which is equivalent to the payoff of H European call options on a bullet bond with face value 1, n payment dates, a period of δ between successive payments, and an annualized coupon rate K. The exercise price of each option equals the face value 1. The price of a European receiver swaption must therefore be equal to the price of these call options. In many models of interest rate dynamics, we can compute such prices quite easily. For the Vasicek model, the swaption prices follow from Equation 12.29 and Theorem 12.8.

Similarly, a European payer swaption yields a payoff of

$$\mathcal{P}_{T_0} = \max\left(\mathbf{P}_{T_0} - 0, 0\right) = \max\left(-\mathbf{R}_{T_0}, 0\right) = H \max\left(1 - \sum_{i=1}^{n} Y_i B_{T_0}^{T_i}, 0\right). \tag{12.60}$$

This is identical to the payoff from H European put options expiring at T_0 and having an exercise price of 1 with a bond paying Y_i at time T_i, $i = 1, 2, \ldots, n$, as its underlying asset.

Alternatively, we can apply (12.57) to express the payoff of a European payer swaption as

$$\mathcal{P}_{T_0} = \left(\sum_{i=1}^{n} B_{T_0}^{T_i}\right) H\delta \max\left(\tilde{l}_{T_0}^\delta - K, 0\right), \tag{12.61}$$

where $\tilde{l}^{\delta}_{T_0}$ is the (equilibrium) swap rate prevailing at time T_0. What is an appropriate numeraire for pricing this swaption? If we were to use the zero-coupon bond maturing at T_0 as the numeraire, we would have to find the expectation of the payoff $\mathcal{P}_{T_0}$ under the T_0-forward martingale measure $\mathbb{Q}^{T_0}$. But since the payoff depends on several different bond prices, the distribution of $\mathcal{P}_{T_0}$ under $\mathbb{Q}^{T_0}$ is rather complicated. It is more convenient to use another numeraire, namely the annuity bond, which at each of the dates $T_1, \ldots, T_n$ provides a payment of 1 dollar. The value of this annuity at time $t \leq T_0$ equals $G_t = \sum_{i=1}^{n} B_t^{T_i}$. In particular, the payoff of the swaption can be restated as

$$\mathcal{P}_{T_0} = G_{T_0} H \delta \max\left(\tilde{l}^{\delta}_{T_0} - K, 0\right),$$

and the payoff expressed in units of the annuity bond is simply $H\delta \max\left(\tilde{l}^{\delta}_{T_0} - K, 0\right)$. The risk-adjusted probability measure corresponding to the annuity being the numeraire is sometimes called the **swap martingale measure** and will be denoted by $\mathbb{Q}^G$ in the following. The price of the European payer swaption can now be written as

$$\mathcal{P}_t = G_t \, \mathrm{E}_t^{\mathbb{Q}^G}\left[\frac{\mathcal{P}_{T_0}}{G_{T_0}}\right] = G_t H \delta \, \mathrm{E}_t^{\mathbb{Q}^G}\left[\max\left(\tilde{l}^{\delta}_{T_0} - K, 0\right)\right], \tag{12.62}$$

so we only need to know the distribution of the swap rate $\tilde{l}^{\delta}_{T_0}$ under the swap martingale measure. In the so-called lognormal swap rate model introduced by Jamshidian (1997), the swap rate $\tilde{l}^{\delta}_{T_0}$ is assumed to be lognormally distributed under the $\mathbb{Q}^G$-measure and the resulting swaption pricing formula is identical to the Black formula for swaptions often applied by practitioners:

$$\mathcal{P}_t = H\delta\left(\sum_{i=1}^{n} B_t^{T_i}\right)\left[\tilde{L}_t^{\delta,T_0} N\left(d(\tilde{L}_t^{\delta,T_0}, t)\right) - K N\left(d(\tilde{L}_t^{\delta,T_0}, t) - \tilde{\sigma}\right)\right], \quad t < T_0, \tag{12.63}$$

where $\tilde{\sigma}$ is the (relative) volatility of the foward swap rate $\tilde{L}_t^{\delta,T_0}$ and

$$d(\tilde{L}_t^{\delta,T_0}, t) = \frac{\ln(\tilde{L}_t^{\delta,T_0}/K)}{\tilde{\sigma}\sqrt{T_0 - t}} + \frac{1}{2}\tilde{\sigma}\sqrt{T_0 - t}.$$

See Munk (2005b, Ch. 11) for a presentation and discussion of the lognormal swap rate model.

Similar to the put-call parity for options we have the following **payer-receiver parity** for European swaptions having the same underlying swap and the same exercise rate:

$$\mathcal{P}_t - \mathcal{R}_t = \mathbf{P}_t, \quad t \leq T_0, \tag{12.64}$$

cf. Exercise 12.8. In words, a payer swaption minus a receiver swaption is indistinguishable from a forward payer swap.

12.5 American-style derivatives

Consider an American-style derivative where the holder has the right to choose when to exercise the derivative, at least within some limits. Typically exercise can take place at the expiration date T or at any time before T. Let P_τ denote the payoff if the derivative is exercised at time $\tau \leq T$. In general, P_τ may depend on the evolution of the economy up to and including time τ, but it is usually a simple function of the time τ price of an underlying security or the time τ value of a particular interest rate. At each point in time the holder of the derivative must decide whether

or not he will exercise. Of course, this decision must be based on the available information, so we are seeking an entire exercise strategy that tell us exactly in what states of the world we should exercise the derivative. We can represent an exercise strategy by an indicator function $I(\omega, t)$, which for any given state of the economy ω at time t either has the value 1 or 0, where the value 1 indicates exercise and 0 indicates non-exercise. For a given exercise strategy I, the derivative will be exercised the first time $I(\omega, t)$ takes on the value 1. We can write this point in time as

$$\tau(I) = \min\{s \in [t, T] \mid I(\omega, s) = 1\}.$$

This is called a stopping time in the literature on stochastic processes. By our earlier analysis, the value of getting the payoff $V_{\tau(I)}$ at time $\tau(I)$ is given by $\mathrm{E}_t^{\mathbb{Q}}\left[e^{-\int_t^{\tau(I)} r_u\, du} P_{\tau(I)}\right]$. If we let $\mathcal{I}[t, T]$ denote the set of all possible exercise strategies over the time period $[t, T]$, the time t value of the American-style derivative must therefore be

$$V_t = \sup_{I \in \mathcal{I}[t,T]} \mathrm{E}_t^{\mathbb{Q}}\left[e^{-\int_t^{\tau(I)} r_u\, du} P_{\tau(I)}\right]. \tag{12.65}$$

An optimal exercise strategy I^* is such that

$$V_t = \mathrm{E}_t^{\mathbb{Q}}\left[e^{-\int_t^{\tau(I^*)} r_u\, du} P_{\tau(I^*)}\right].$$

Note that the optimal exercise strategy and the price of the derivative must be solved for simultaneously. This complicates the pricing of American-style derivatives considerably. In fact, in all situations where early exercise may be relevant, we will not be able to compute closed-form pricing formulas for American-style derivatives. We have to resort to numerical techniques. See Hull (2006) for an introduction to the standard techniques of binomial or trinomial trees, finite difference approximation of the partial differential equation that the pricing function must satisfy, and Monte Carlo simulation.

It is well-known that it is never strictly advantageous to exercise an American call option on a non-dividend paying asset before the final maturity date T, cf. Merton (1973c) and Hull (2006, Ch. 9). In contrast, premature exercise of an American put option on a non-dividend paying asset will be advantageous for sufficiently low prices of the underlying asset. If the underlying asset pays dividends at discrete points in time, it can be optimal to exercise an American call option prematurely but only immediately before each dividend payment date. Regarding early exercise of put options, it can never be optimal to exercise an American put on a dividend-paying asset just before a dividend payment, but at all other points in time early exercise will be optimal for sufficiently low prices of the underlying asset.

12.6 Concluding remarks

This chapter has given an introduction to standard derivatives and the pricing of such derivatives. Numerous continuous-time models have been proposed in the literature for the pricing of derivatives on stocks, interest rates, bonds, commodities, foreign exchange, and other variables. Also many "exotic" variations of the basic derivatives are traded and studied in the literature. In many cases the prices of some relevant derivatives cannot be computed explicitly given the modeling assumptions found to be reasonable so the prices have to be computed by approximations or

numerical solution techniques. The design of efficient computational techniques for derivatives pricing is also an active research area.

The interested reader can find much more information on derivatives in specialized textbooks such as Björk (2004), Brigo and Mercurio (2001), Hull (2006), James and Webber (2000), Munk (2005b), Musiela and Rutkowski (1997). The market for derivatives with payoffs depending on credit events, such as the default of a given corporation, has been rapidly growing recently. Such derivatives and their pricing are studied in textbooks such as Bielecki and Rutkowski (2002), Duffie and Singleton (2003), Lando (2004), and Schönbucher (2003).

12.7 Exercises

EXERCISE 12.1 Consider a coupon bond with payment dates $T_1 < T_2 < \cdots < T_n$. For each $i = 1, 2, \ldots, n$, let Y_i be the sure payment at time T_i. For some $t < T < T_i$, let Φ_t^{T,T_i} denote the futures price at time t for delivery at time T of the zero-coupon bond maturing at time T_i with a unit payment. Show that futures price at time t for delivery at time T of the coupon bond satisfies

$$\Phi_t^{T,\text{cpn}} = \sum_{T_i > T} Y_i \Phi_t^{T,T_i}.$$

EXERCISE 12.2 Show by differentiation that the Black-Scholes-Merton call option price satisfies (12.22) and (12.23).

EXERCISE 12.3 Show that the no-arbitrage price of a European call option on a non-dividend paying stock must satisfy

$$\max\left(0, S_t - KB_t^T\right) \leq C_t \leq S_t.$$

Show that the no-arbitrage price of a European call on a zero-coupon bond will satisfy

$$\max\left(0, B_t^S - KB_t^T\right) \leq C_t^{K,T,S} \leq B_t^S(1 - K)$$

provided that all interest rates are non-negative.

EXERCISE 12.4 We will adapt the Black-Scholes-Merton model and option pricing formula to three cases in which the underlying asset provides dividend payments before the expiration of the option at time T.

I. Discrete dividends known in absolute terms. Suppose that the underlying asset pays dividends at n points in time before time T, namely $t_1 < t_2 < \cdots < t_n$. All the dividends are known already. Let D_j denote the dividend at time t_j. The time t value of all the remaining dividends is then

$$D_t^* = \sum_{t_j > t} D_j e^{-r(t_j - t)},$$

where r is the constant interest rate. Define $S_t^* = S_t - D_t^*$. Note that $S_T^* = S_T$.

(a) Show that S_t^* is the necessary investment at time t to end up with one unit of the underlying asset at time T.

(b) Assuming that S_t^* has constant volatility σ, so that

$$dS_t^* = S_t^* \left(\mu(\cdot)\, dt + \sigma\, dz_t\right)$$

for some drift term $\mu(\cdot)$, derive a Black-Scholes-Merton-type equation for a European call option on this asset. State the option price in terms of S_t (and the remaining dividends). Compare with the standard Black-Scholes-Merton formula—in particular, check whether σ is equal to the volatility of S_t under the assumptions on S^*.

II. Discrete dividends known as a percentage of the price of the underlying asset. Again assume that dividends are paid at $t_1 < t_2 < \cdots < t_n$, but now assume that the dividend at time t_j is known to be $D_j = \delta_j S_{t_j-}$, where δ_j is a known constant and S_{t_j-} is the price just before the dividend is paid out. The ex-dividend price is then $S_{t_j} = (1-\delta_j)S_{t_j-}$. Define the process S^* by

$$S_t^* = S_t \prod_{t_j > t} (1-\delta_j), \quad t < t_n,$$

and $S_t^* = S_t$ for $t \geq t_n$. Answer the questions (a) and (b) above using this definition of S^*.

III. Continuous dividend payments at a known rate. Now suppose that the underlying asset pays dividends continuously at a constant and known relative rate δ. This means that over any very short time interval $[t, t+\Delta t]$, the total dollar dividends is $\int_t^{t+\Delta t} \delta S_u\, du$ or approximately $\delta S_{t+\Delta t}\Delta t$. Define

$$S_t^* = S_t e^{-\delta(T-t)}.$$

Again, answer the questions (a) and (b) using this new definition of S^*. *Hint: For part (a), you may want to divide the interval $[t,T]$ into N equally long subintervals and assume that dividends are paid only at the end of each subinterval. Use the result* $\lim_{N\to\infty}(1+\delta\frac{T-t}{N})^N = e^{\delta(T-t)}$ *to go to the continuous-time limit.*

EXERCISE 12.5 Let $S_1 = (S_{1t})$ and $S_2 = (S_{2t})$ be the price processes of two assets. Consider the option to exchange (at zero cost) one unit of asset 2 for one unit of asset 1 at some prespecified date T. The payoff is thus $\max(S_{1T} - S_{2T}, 0)$. The assets have no dividends before time T.

(a) Argue that the time t value of this option can be written as

$$V_t = S_{2t}\, \mathrm{E}_t^{\mathbb{Q}_2}\left[\max\left(\frac{S_{1T}}{S_{2T}} - 1, 0\right)\right],$$

where $\mathbb{Q}_2$ is the risk-adjusted probability measure associated with asset 2.

Suppose that S_1 and S_2 are both geometric Brownian motions so that we may write their joint dynamics as

$$\begin{aligned} dS_{1t} &= S_{1t}\left[\mu_1\, dt + \sigma_1\, dz_{1t}\right], \\ dS_{2t} &= S_{2t}\left[\mu_2\, dt + \rho\sigma_2\, dz_{1t} + \sqrt{1-\rho^2}\sigma_2\, dz_{2t}\right]. \end{aligned}$$

(b) Find the dynamics of S_{1t}/S_{2t} under the probability measure $\mathbb{Q}_2$.

(c) Use the two previous questions and your knowledge of lognormal random variables to show that

$$V_t = S_{1t}N(d_1) - S_{2t}N(d_2), \tag{12.66}$$

where

$$d_1 = \frac{\ln(S_{1t}/S_{2t})}{v} + \frac{1}{2}v, \quad d_2 = d_1 - v, \quad v = \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)(T-t)}.$$

This formula was first given by Margrabe (1978).

(d) Give pricing formulas (in terms of S_{1t} and S_{2t}) for an option with payoff $\max(S_{1T}, S_{2T})$ and an option with payoff $\min(S_{1T}, S_{2T})$.

(e) What happens to the pricing formula (12.66) if asset 2 is a zero-coupon bond maturing at time T with a payment of K? And if, furthermore, interest rates are constant, what then?

EXERCISE 12.6 Let $F_t^{T,S}$ and $\Phi_t^{T,S}$ denote the forward price and futures price at time t, respectively, for delivery at time $T > t$ of a zero-coupon bond maturing at time $S > T$. Under the assumptions of the Vasicek model introduced in Section 10.5.1, show that

$$F_t^{T,S} = \exp\{-[a(S-t) - a(T-t)] - [b(S-t) - b(T-t)]r_t\},$$
$$\Phi_t^{T,S} = \exp\{-\tilde{a}(T-t, S-T) - [b(S-t) - b(T-t)]r_t\},$$

where $a(\cdot)$ and $b(\cdot)$ are given by (10.35) and (10.33),

$$\tilde{a}(T-t, S-T) = a(S-T) + \kappa\hat{r}b(S-T)b(T-t) - \frac{\sigma_r^2}{2}b(S-T)^2\left(b(T-t) - \frac{\kappa}{2}b(T-t)^2\right)$$

and $\hat{r} = \bar{r} - \sigma_r\lambda/\kappa$.

EXERCISE 12.7 Let $\tilde{l}_{T_0}^{\delta}(k)$ be the equilibrium swap rate for a swap with payment dates $T_1, T_2, \ldots, T_k$, where $T_i = T_0 + i\delta$ as usual. Suppose that $\tilde{l}_{T_0}^{\delta}(1), \ldots, \tilde{l}_{T_0}^{\delta}(n)$ are known. Find a recursive procedure for deriving the associated discount factors $B_{T_0}^{T_1}, B_{T_0}^{T_2}, \ldots, B_{T_0}^{T_n}$.

EXERCISE 12.8 Show the parity (12.64). Show that a payer swaption and a receiver swaption (with identical terms) will have identical prices, if the exercise rate of the contracts is equal to the forward swap rate $\tilde{L}_t^{\delta,T_0}$.

EXERCISE 12.9 Consider a swap with starting date T_0 and a fixed rate K. For $t \le T_0$, show that $V_t^{\text{fl}}/V_t^{\text{fix}} = \tilde{L}_t^{\delta,T_0}/K$, where $\tilde{L}_t^{\delta,T_0}$ is the forward swap rate.

Appendix A

A review of basic probability concepts

Any asset pricing model must handle uncertainty. Therefore we need to apply some concepts and results from probability theory. We will be a bit more formal than many textbooks on statistics for business and economics. This section is meant to give a short introduction. We will discuss further issues in later chapters when we need them in our asset pricing models.

The basic mathematical object for studies of uncertain events is a **probability space**, which is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of a state space Ω, a sigma-algebra $\mathcal{F}$, and a probability measure $\mathbb{P}$. Any study of uncertain events must explicitly or implicitly specify the probability space. Let us discuss each of the three elements of a probability space in turn.

The **state space** Ω is the set of possible states or outcomes of the uncertain object. Only one of these states will be realized. For example, if one studies the outcome of a throw of a die (the number of "eyes" on the upside), the state space is $\Omega = \{1, 2, 3, 4, 5, 6\}$. An event is a set of possible outcomes, i.e. a subset of the state space. In the example with the die, some events are $\{1, 2, 3\}$, $\{4, 5\}$, $\{1, 3, 5\}$, $\{6\}$, and $\{1, 2, 3, 4, 5, 6\}$. This is an example where a finite state space is natural. For other uncertain objects it is natural to take an infinite state space. If we only want to study the dividend of a given stock at a given point in time, an appropriate state space is $\mathbb{R}_+ \equiv [0, \infty)$ since the dividend may in principle be any non-negative real number. In our asset pricing models we want to study the entire economy over a certain time span so the state space has to list all the possible realizations of dividends of all assets and incomes of all individuals. Of course, this requires a large state space. Note that some authors use the term sample space instead of state space.

The second component of a probability space, $\mathcal{F}$, is the set of events to which a probability can be assigned, i.e. the **set of "probabilizable" or "measurable" events**. Hence, $\mathcal{F}$ is a set of subsets of the state space! It is required that

(i) the entire state space can be assigned a probability, i.e. $\Omega \in \mathcal{F}$;

(ii) if some event $F \subseteq \Omega$ can be assigned a probability, so can its complement $F^c \equiv \Omega \setminus F$, i.e. $F \in \mathcal{F} \Rightarrow F^c \in \mathcal{F}$; and

(iii) given a sequence of probabilizable events, the union is also probabilizable, i.e. $F_1, F_2, \cdots \in \mathcal{F} \Rightarrow \cup_{i=1}^{\infty} F_i \in \mathcal{F}$.

A set $\mathcal{F}$ with these properties is called a **sigma-algebra**, a sigma-field, or a tribe. We will stick to the term sigma-algebra.

An alternative way to represent the probabilizable events is by a partition $\mathbf{F}$ of Ω. By a partition $\mathbf{F}$ of Ω we mean a collection $A_1, \ldots, A_k$ of disjoint subsets of Ω, i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$, so that the union of these subsets equals the entire set Ω, i.e. $\Omega = A_1 \cup \cdots \cup A_k$. With a finite state space $\Omega = \{\omega_1, \omega_2, \ldots, \omega_S\}$ the natural partition is

$$\mathbf{F} = \Big\{\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_S\}\Big\},$$

which intuitively means that we will learn exactly which state is realized. Given a partition $\mathbf{F}$ we can define an associated sigma-algebra $\mathcal{F}$ as the set of all unions of (countably many) sets in $\mathbf{F}$ including the "empty union", i.e. the empty set $\emptyset$. Again, if $\Omega = \{\omega_1, \omega_2, \ldots, \omega_S\}$ and $\mathbf{F} = \Big\{\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_S\}\Big\}$, the corresponding sigma-algebra is the set of all subsets of Ω. On the other hand we can also go from a sigma-algebra $\mathcal{F}$ to a partition $\mathbf{F}$. Just remove all sets in $\mathcal{F}$ that are unions of the sets in $\mathcal{F}$. Again this includes the empty set $\emptyset$ since that is is an "empty union" of the other sets in $\mathcal{F}$. If the state space is infinite, the equivalence between a partition and a sigma-algebra may break down, and the sigma-algebra formulation is the preferred one; see for example the discussion in Björk (2004, App. B).

We can think of the sigma-algebra $\mathcal{F}$ or the associated partition $\mathbf{F}$ as representing full information about the realization of the state. In some cases it can be relevant also to model some limited information about the realized state. Many models in financial economics are designed to capture uncertainty about many different variables or objects, for example the dividends on a large number of stocks. It may be relevant to formalize what can be learned about the true state by just observing the dividends of one particular stock. In other models some individuals are assumed to know more about some uncertain objects than other individuals. Less-than-full information can be represented formally by a sigma-algebra $\mathcal{G}$ on Ω, which is coarser than $\mathcal{F}$ in the sense that any set in $\mathcal{G}$ is also in $\mathcal{F}$. In terms of partitions, a partition $\mathbf{G}$ of Ω represent less information than $\mathbf{F}$ if any set in $\mathbf{G}$ is the union of sets in $\mathbf{F}$. In the example with the throw of a die, full information is represented by the partition

$$\mathbf{F} = \Big\{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\Big\}$$

or the associated sigma-algebra. An example of less-than-perfect information is represented by the partition

$$\mathbf{G} = \Big\{\{1, 3, 5\}, \{2, 4, 6\}\Big\}$$

or the associated sigma-algebra

$$\mathcal{G} = \Big\{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\Big\}.$$

With $\mathbf{G}$, you will only know whether the die will show an odd or an even number of eyes on the upside. As mentioned above the link between partitions and sigma-algebras is more delicate in infinite state spaces and so is the notion of information. Dubra and Echenique (2004) gives an example in an economic setting where one partition represents more information than another

partition but the sigma-algebra associated with the second partition seems to represent more information than the sigma-algebra associated with the first!

The final component of a probability space is a **probability measure** $\mathbb{P}$, which formally is a function from the sigma-algebra $\mathcal{F}$ into the interval $[0,1]$. To each event $F \in \mathcal{F}$, the probability measure assigns a number $\mathbb{P}(F)$ in the interval $[0,1]$. This number is called the $\mathbb{P}$-probability (or simply the probability) of F. A probability measure must satisfy the following conditions:

(i) $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}(\emptyset) = 0$;

(ii) the probability of the state being in the union of disjoint sets is equal to the sum of the probabilities for each of the sets, i.e. given $F_1, F_2, \cdots \in \mathcal{F}$ with $F_i \cap F_j = \emptyset$ for all $i \neq j$, we have $\mathbb{P}(\cup_{i=1}^{\infty} F_i) = \sum_{i=1}^{\infty} \mathbb{P}(F_i)$.

If the state space Ω is finite, say $\Omega = \{\omega_1, \omega_2, \ldots, \omega_S\}$, and each $\{\omega_i\}$ is probabilizable, a probability measure $\mathbb{P}$ is fully specified by the individual state probabilities $\mathbb{P}(\omega_i)$, $i = 1, 2, \ldots, S$.

Many different probability measures can be defined on the same sigma-algebra, $\mathcal{F}$, of events. In the example of the die, a probability measure $\mathbb{P}$ corresponding to the idea that the die is "fair" is defined by

$$\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \cdots = \mathbb{P}(\{6\}) = 1/6.$$

Another probability measure, $\mathbb{Q}$, can be defined by

$$\mathbb{Q}(\{1\}) = 1/12, \quad \mathbb{Q}(\{2\}) = \cdots = \mathbb{Q}(\{5\}) = 1/6, \quad \mathbb{Q}(\{6\}) = 3/12,$$

which may be appropriate if the die is believed to be "unfair."

Two probability measures $\mathbb{P}$ and $\mathbb{Q}$ defined on the same state space and sigma-algebra $(\Omega, \mathcal{F})$ are called equivalent if the two measures assign probability zero to exactly the same events, i.e. if $\mathbb{P}(A) = 0 \Leftrightarrow \mathbb{Q}(A) = 0$. The two probability measures in the die example are equivalent. In the stochastic models of financial markets switching between equivalent probability measures turns out to be useful.

A **random variable** is a function X from the state space Ω into the real numbers $\mathbb{R}$. To each possible outcome $\omega \in \Omega$ the function assigns a real number $X(\omega)$. A random variable is thus the formal way to represent a state-dependent quantity. To be meaningful, the function X must be $\mathcal{F}$-measurable. This means that for any interval $I \in \mathbb{R}$, the set $\{\omega \in \Omega | X(\omega) \in I\}$ belongs to $\mathcal{F}$, i.e. we can assign a probability to the event that the random variable takes on a value in the interval I. A random variable is thus defined relative to a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Any random variable is associated with a probability distribution. We can represent the distribution by the **cumulative distribution function** $F_X : \mathbb{R} \to \mathbb{R}$ defined for any $x \in \mathbb{R}$ by

$$F_X(x) = \mathbb{P}(X \leq x) \equiv \mathbb{P}(\{\omega \in \Omega | X(\omega) \leq x\}).$$

If the random variable can only take on finitely many different values $x_1, x_2, \ldots, x_m \in \mathbb{R}$, it is said to be a discrete-valued or simply a **discrete random variable**, and we can represent the probability distribution by the numbers $f_X(x_i) \equiv \mathbb{P}(X = x_i) \equiv \mathbb{P}(\{\omega \in \Omega | X(\omega) = x_i\})$. Note that this is surely the case if the state space Ω itself is finite. A random variable X is said to be a continuous-valued or simply **continuous random variable** if it can take on a continuum of

possible values and a function $f_X : \mathbb{R} \to \mathbb{R}$ exists such that

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy.$$

The function f_X is then called the **probability density function** of X. It is also possible to construct random variables that are neither discrete or continuous but they will not be important for our purposes. In any case we can represent the probability distribution more abstractly by a **distribution measure** μ_X, which is a probability measure on the real numbers $\mathbb{R}$ equipped with the so-called Borel-algebra $\mathcal{B}$. The Borel-algebra can be defined as the smallest sigma-algebra that includes all intervals. The Borel-algebra includes all subsets of $\mathbb{R}$ "you can think of" but there are in fact some very obscure subsets of $\mathbb{R}$ which are not in the Borel-algebra. Fortunately, this will be unimportant for our purposes. The distribution measure is defined for any $B \in \mathcal{B}$ by

$$\mu_X(B) = \mathbb{P}(X \in B) \equiv \mathbb{P}(\{\omega \in \Omega | X(\omega) \in B\}).$$

It is often useful to summarize the probability distribution of a random variable in a few informative numbers. The most frequently used are the expected value (or mean) and the variance. For a discrete random variable X that can take on the values $x_1, \ldots, x_m \in \mathbb{R}$ the **expected value** $\mathrm{E}[X]$ is defined by

$$\mathrm{E}[X] = \sum_{i=1}^{m} x_i \mathbb{P}(X = x_i).$$

For a continuous random variable X with probability density function f_X, the expected value is defined as

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

if this integral is finite; otherwise the random variable does not have an expected value. Similarly we can define the expected value of a function $g(X)$ as $\mathrm{E}[g(X)] = \sum_{i=1}^{m} g(x_i)\mathbb{P}(X = x_i)$ or $\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$, respectively.

For a general random variable X we can define the expected value of $g(X)$ as

$$\mathrm{E}[g(X)] = \int_{\Omega} g(X(\omega))\, d\mathbb{P}(\omega)$$

which is an integral with respect to the probability measure $\mathbb{P}$. For functions that are just modestly nice (so-called Borel functions) one can rewrite the expected value as

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, d\mu_X(x)$$

which is an integral with respect to the distribution measure μ_X of the random variable. We do not want to go into the theory of integration with respect to various measures so let us just note that for discrete and continuous random variables the general definition simplifies to the definitions given in the paragraph above.

The **variance** of a random variable X is generally defined as

$$\mathrm{Var}[X] = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2.$$

The **standard deviation** of X is $\sigma[X] = \sqrt{\mathrm{Var}[X]}$. The n'th moment of the random variable X is $\mathrm{E}[X^n]$, while the n'th central moment is $\mathrm{E}[(X - \mathrm{E}[X])^n]$. In particular, the variance is the second central moment.

It can be shown that, for any constants a and b,

$$\mathrm{E}[aX + b] = a\,\mathrm{E}[X] + b, \quad \mathrm{Var}[aX + b] = a^2\,\mathrm{Var}[X].$$

In the example with the throw of a die, the random variable X defined by $X(\omega) = \omega$ for all $\omega \in \Omega$ simply represents the uncertain number of eyes on the upside of the die after the throw. Other random variables may be relevant also. Suppose that you bet 10 dollars with a friend that the number of eyes on the upside will be even or odd. If even, you will win 10 dollars, if odd, you will loose 10 dollars. A random variable Y capturing your uncertain gain on the bet can be defined as follows:

$$Y(\omega) = \begin{cases} 10, & \text{if } \omega \in \{2,4,6\}, \\ -10, & \text{if } \omega \in \{1,3,5\}. \end{cases}$$

If the die is believed to be fair, corresponding to the probability measure $\mathbb{P}$, the distribution associated with the random variable Y is given by

$$\mathbb{P}\left(Y = -10\right) = \mathbb{P}\left(\omega \in \{1,3,5\}\right) = \frac{1}{2}, \quad \mathbb{P}\left(Y = 10\right) = \mathbb{P}\left(\omega \in \{2,4,6\}\right) = \frac{1}{2}$$

or by the cumulative distribution function

$$F_Y(x) \equiv \mathbb{P}\left(Y(\omega) \leq x\right) = \begin{cases} 0, & \text{for } x < -10, \\ \frac{1}{2}, & \text{for } -10 \leq x < 10, \\ 1, & \text{for } x \geq 10. \end{cases}$$

Observing the realization of a random variable can give you some information about which state ω was realized. If the random variable X takes on different values in all states, i.e. you cannot find $\omega_1, \omega_2 \in \Omega$ with $\omega_1 \neq \omega_2$ and $X(\omega_1) \neq X(\omega_2)$, observing the realized value $X(\omega)$ will tell you exactly which state was realized. On the other extreme, if X takes on the same value in all states, you cannot infer anything from observing $X(\omega)$. Other random variables will tell you something but not all. In the example above, observing the realization of the random variable Y will tell you either that the realized state is in $\{1,3,5\}$ or in $\{2,4,6\}$. We can represent this by the partition

$$\mathbf{F}_Y = \Big\{\{1,3,5\}, \{2,4,6\}\Big\}$$

or the associated sigma-algebra

$$\mathcal{F}_Y = \Big\{\emptyset, \{1,3,5\}, \{2,4,6\}, \Omega\Big\}.$$

More generally, we can define the sigma-algebra associated with a random variable $X : \Omega \to \mathbb{R}$ to be the smallest sigma-algebra on Ω with respect to which X is a measurable function. This sigma-algebra will be denoted $\mathcal{F}_X$. Just think of this as the information generated by X.

We have defined a random variable to be a function from Ω to $\mathbb{R}$. Given two random variables X_1 and X_2 on the same probability space, we can form the vector $(X_1, X_2)^\top$, which is then a (measurable) function from Ω to $\mathbb{R}^2$ said to be a two-dimensional random variable. For example, X_1 could represent the uncertain dividend of one asset and X_2 the uncertain dividend of another asset. Similarly we can define random variables of any other (integer) dimension. This will often be notationally convenient.

For a two-dimensional random variable $(X_1, X_2)^\top$ the joint or simultaneous cumulative distribution function is the function $F_{(X_1,X_2)} : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$F_{(X_1,X_2)}(x_1, x_2) = \mathbb{P}\left(X_1 \leq x_1, X_2 \leq x_2\right) \equiv \mathbb{P}\left(\{\omega \in \Omega | X_1(\omega) \leq x_1 \text{ and } X_2(\omega) \leq x_2\}\right).$$

If both X_1 and X_2 are discrete random variables, the vector random variable $(X_1, X_2)^\top$ is also discrete, and the joint probability distribution is characterized by probabilities $\mathbb{P}\left(X_1 = x_1, X_2 = x_2\right)$. The two-dimensional random variable (X_1, X_2) is said to be continuous if a function $f_{(X_1,X_2)} : \mathbb{R}^2 \to \mathbb{R}$ exists such that

$$F_{(X_1,X_2)}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} f_{(X_1,X_2)}(y_1, y_2)\, dy_1\, dy_2$$

and $f_{(X_1,X_2)}$ is then called the joint or simultaneous probability density function of (X_1, X_2).

Given the joint distribution of $(X_1, X_2)^\top$, we can find the distributions of X_1 and X_2, the so-called marginal distributions. For example, if $(X_1, X_2)^\top$ is continuous with joint probability density function $f_{(X_1,X_2)}$, we can find the marginal probability density function of X_1 by integrating over all possible values of X_2, i.e.

$$f_{X_1}(x_1) = \int_{-\infty}^{+\infty} f_{(X_1,X_2)}(x_1, x_2)\, dx_2.$$

Two random variables X_1 and X_2 are said to be **independent** if

$$\mathbb{P}\left(X_1 \in B_1, X_2 \in B_2\right) = \mathbb{P}\left(X_1 \in B_1\right)\, \mathbb{P}\left(X_2 \in B_2\right)$$

for all Borel sets $B_1, B_2 \subseteq \mathbb{R}$ or, equivalently, if

$$F_{(X_1,X_2)}(x_1, x_2) = F_{X_1}(x_1)\, F_{X_2}(x_2)$$

for all $(x_1, x_2) \in \mathbb{R}^2$.

We can easily extend the definition of expected value to functions of multi-dimensional random variables. If $(X_1, X_2)^\top$ is a two-dimensional continuous random variable and $g : \mathbb{R}^2 \to \mathbb{R}$, the expected value of $g(X_1, X_2)$ is defined as

$$\mathrm{E}[g(X_1, X_2)] = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x_1, x_2) f_{(X_1,X_2)}(x_1, x_2)\, dx_1\, dx_2.$$

We define the **covariance** between X_1 and X_2 by

$$\mathrm{Cov}[X_1, X_2] = \mathrm{E}\left[(X_1 - \mathrm{E}[X_1])\,(X_2 - \mathrm{E}[X_2])\right] = \mathrm{E}[X_1 X_2] - \mathrm{E}[X_1]\,\mathrm{E}[X_2].$$

In particular, $\mathrm{Cov}[X_1, X_1] = \mathrm{Var}[X_1]$. The **correlation** between X_1 and X_2 is

$$\rho[X_1, X_2] = \frac{\mathrm{Cov}[X_1, X_2]}{\sigma[X_1]\sigma[X_2]},$$

which is a number in the interval $[-1, 1]$. Some useful properties of covariances are

$$\mathrm{Cov}[X_1, X_2] = \mathrm{Cov}[X_2, X_1], \quad \mathrm{Cov}[aX_1 + bX_2, X_3] = a\,\mathrm{Cov}[X_1, X_3] + b\,\mathrm{Cov}[X_2, X_3],$$

where X_1, X_2, X_3 are three random variables and $a, b \in \mathbb{R}$. An often used result is

$$\mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2\,\mathrm{Cov}[X_1, X_2].$$

If X_1 and X_2 are independent, one can show that $\mathrm{Cov}[X_1, X_2] = 0$ and, consequently, $\rho[X_1, X_2] = 0$.

If $\boldsymbol{X} = (X_1, \ldots, X_K)^\top$ is a K-dimensional random variable, its variance-covariance matrix is the $K \times K$ matrix $\mathrm{Var}[\boldsymbol{X}]$ with (i, j)'th entry given by $\mathrm{Cov}[X_i, X_j]$. If $\boldsymbol{X}$ is a K-dimensional random variable and $\underline{\underline{A}}$ is an $M \times K$ matrix, then

$$\mathrm{Var}[\underline{\underline{A}}\boldsymbol{X}] = \underline{\underline{A}}\,\mathrm{Var}[\boldsymbol{X}]\underline{\underline{A}}^\top. \tag{A.1}$$

If $\boldsymbol{X} = (X_1, \ldots, X_K)^\top$ is a K-dimensional random variable and $\boldsymbol{Y} = (Y_1, \ldots, Y_L)^\top$ is an L-dimensional random variable, their covariance matrix $\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}]$ is the $K \times L$ matrix whose (k, l)'th entry is $\mathrm{Cov}[X_k, Y_l]$. If $\boldsymbol{X}$ is a K-dimensional random variable, $\boldsymbol{Y}$ is an L-dimensional random variable, $\underline{\underline{A}}$ is an $M \times K$ matrix, and $\boldsymbol{a}$ is an M-dimensional vector, then

$$\mathrm{Cov}[\boldsymbol{a} + \underline{\underline{A}}\boldsymbol{X}, \boldsymbol{Y}] = \underline{\underline{A}}\,\mathrm{Cov}[\boldsymbol{X}, \boldsymbol{Y}]. \tag{A.2}$$

Conditional expectations....

Appendix B

Results on the lognormal distribution

A random variable Y is said to be lognormally distributed if the random variable $X = \ln Y$ is normally distributed. In the following we let m be the mean of X and s^2 be the variance of X, so that

$$X = \ln Y \sim N(m, s^2).$$

The probability density function for X is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi s^2}} \exp\left\{-\frac{(x-m)^2}{2s^2}\right\}, \qquad x \in \mathbb{R}.$$

Theorem B.1 *The probability density function for Y is given by*

$$f_Y(y) = \frac{1}{\sqrt{2\pi s^2} y} \exp\left\{-\frac{(\ln y - m)^2}{2s^2}\right\}, \qquad y > 0,$$

and $f_Y(y) = 0$ for $y \leq 0$.

This result follows from the general result on the distribution of a random variable which is given as a function of another random variable; see any introductory text book on probability theory and distributions.

Theorem B.2 *For $X \sim N(m, s^2)$ and $\gamma \in \mathbb{R}$ we have*

$$\mathrm{E}\left[e^{-\gamma X}\right] = \exp\left\{-\gamma m + \frac{1}{2}\gamma^2 s^2\right\}.$$

Proof: Per definition we have

$$\mathrm{E}\left[e^{-\gamma X}\right] = \int_{-\infty}^{+\infty} e^{-\gamma x} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-m)^2}{2s^2}}\, dx.$$

Manipulating the exponent we get

$$\begin{aligned}
\mathrm{E}\left[e^{-\gamma X}\right] &= e^{-\gamma m + \frac{1}{2}\gamma^2 s^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{1}{2s^2}\left[(x-m)^2 + 2\gamma(x-m)s^2 + \gamma^2 s^4\right]}\, dx \\
&= e^{-\gamma m + \frac{1}{2}\gamma^2 s^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-[m-\gamma s^2])^2}{2s^2}}\, dx \\
&= e^{-\gamma m + \frac{1}{2}\gamma^2 s^2},
\end{aligned}$$

where the last equality is due to the fact that the function

$$x \mapsto \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-[m-\gamma s^2])^2}{2s^2}}$$

is a probability density function, namely the density function for an $N(m-\gamma s^2, s^2)$ distributed random variable. □

Using this theorem, we can easily compute the mean and the variance of the lognormally distributed random variable $Y = e^X$. The mean is (let $\gamma = -1$)

$$\mathrm{E}[Y] = \mathrm{E}\left[e^X\right] = \exp\left\{m + \frac{1}{2}s^2\right\}. \tag{B.1}$$

With $\gamma = -2$ we get

$$\mathrm{E}\left[Y^2\right] = \mathrm{E}\left[e^{2X}\right] = e^{2(m+s^2)},$$

so that the variance of Y is

$$\begin{aligned} \mathrm{Var}[Y] &= \mathrm{E}\left[Y^2\right] - (\mathrm{E}[Y])^2 \\ &= e^{2(m+s^2)} - e^{2m+s^2} \\ &= e^{2m+s^2}\left(e^{s^2} - 1\right). \end{aligned} \tag{B.2}$$

The next theorem provides an expression of the truncated mean of a lognormally distributed random variable, i.e. the mean of the part of the distribution that lies above some level. We define the indicator variable $\mathbf{1}_{\{Y>K\}}$ to be equal to 1 if the outcome of the random variable Y is greater than the constant K and equal to 0 otherwise.

Theorem B.3 *If $X = \ln Y \sim N(m, s^2)$ and $K > 0$, then we have*

$$\begin{aligned} \mathrm{E}\left[Y\mathbf{1}_{\{Y>K\}}\right] &= e^{m+\frac{1}{2}s^2} N\left(\frac{m - \ln K}{s} + s\right) \\ &= \mathrm{E}[Y]\, N\left(\frac{m - \ln K}{s} + s\right). \end{aligned}$$

Proof: Because $Y > K \Leftrightarrow X > \ln K$, it follows from the definition of the expectation of a random variable that

$$\begin{aligned} \mathrm{E}\left[Y\mathbf{1}_{\{Y>K\}}\right] &= \mathrm{E}\left[e^X \mathbf{1}_{\{X>\ln K\}}\right] \\ &= \int_{\ln K}^{+\infty} e^x \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-m)^2}{2s^2}}\, dx \\ &= \int_{\ln K}^{+\infty} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-[m+s^2])^2}{2s^2}} e^{\frac{2ms^2+s^4}{2s^2}}\, dx \\ &= e^{m+\frac{1}{2}s^2} \int_{\ln K}^{+\infty} f_{\bar{X}}(x)\, dx, \end{aligned}$$

where

$$f_{\bar{X}}(x) = \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{(x-[m+s^2])^2}{2s^2}}$$

is the probability density function for an $N(m+s^2, s^2)$ distributed random variable. The calculations

$$\begin{aligned}\int_{\ln K}^{+\infty} f_{\bar{X}}(x)\,dx &= \text{Prob}(\bar{X} > \ln K)\\ &= \text{Prob}\left(\frac{\bar{X}-[m+s^2]}{s} > \frac{\ln K-[m+s^2]}{s}\right)\\ &= \text{Prob}\left(\frac{\bar{X}-[m+s^2]}{s} < -\frac{\ln K-[m+s^2]}{s}\right)\\ &= N\left(-\frac{\ln K-[m+s^2]}{s}\right)\\ &= N\left(\frac{m-\ln K}{s}+s\right)\end{aligned}$$

complete the proof. □

Theorem B.4 *If $X = \ln Y \sim N(m, s^2)$ and $K > 0$, we have*

$$\begin{aligned}\text{E}\left[\max\left(0, Y-K\right)\right] &= e^{m+\frac{1}{2}s^2} N\left(\frac{m-\ln K}{s}+s\right) - KN\left(\frac{m-\ln K}{s}\right)\\ &= \text{E}\left[Y\right] N\left(\frac{m-\ln K}{s}+s\right) - KN\left(\frac{m-\ln K}{s}\right).\end{aligned}$$

Proof: Note that

$$\begin{aligned}\text{E}\left[\max\left(0, Y-K\right)\right] &= \text{E}\left[(Y-K)\mathbf{1}_{\{Y>K\}}\right]\\ &= \text{E}\left[Y\mathbf{1}_{\{Y>K\}}\right] - K\text{Prob}\left(Y>K\right).\end{aligned}$$

The first term is known from Theorem B.3. The second term can be rewritten as

$$\begin{aligned}\text{Prob}\left(Y>K\right) &= \text{Prob}\left(X>\ln K\right)\\ &= \text{Prob}\left(\frac{X-m}{s} > \frac{\ln K-m}{s}\right)\\ &= \text{Prob}\left(\frac{X-m}{s} < -\frac{\ln K-m}{s}\right)\\ &= N\left(-\frac{\ln K-m}{s}\right)\\ &= N\left(\frac{m-\ln K}{s}\right).\end{aligned}$$

The claim now follows immediately. □

Bibliography

Allais, M. (1953). Le Comportement de l'Homme Rationnel devant le Risque – Critique des Postulats et Axiomes de l'École Américaine. *Econometrica 21*, 503–546.

Anscombe, F. and R. Aumann (1963). A Definition of Subjective Probability. *Annals of Mathematical Statistics 34*, 199–205.

Arrow, K. (1951). An Extension of the Basic Theorems of Classical Welfare Economics. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 507–532. University of California Press.

Arrow, K. (1953). Le Rôle des Valeurs Boursières pour la Repartition la Meillure des Risques. *Econometrie 40*, 41–47. English translation: Arrow (1964).

Arrow, K. (1964). The Role of Securities in the Optimal Allocation of Risk-Bearing. *Review of Economic Studies 31*, 91–96.

Arrow, K. (1970). *Essays in the Theory of Risk Bearing*. North-Holland.

Babbs, S. H. and N. J. Webber (1994, February). A Theory of the Term Structure with an Official Short Rate. Working paper, Warwick Business School, University of Warwick, Coventry CV4 7AL, UK.

Bachelier, L. (1900). *Théorie de la Spéculation*, Volume 3 of *Annales de l'Ecole Normale Supérieure*. Gauthier-Villars. English translation in Cootner (1964).

Bakshi, G. S. and Z. Chen (1996). Inflation, Asset Prices and the Term Structure of Interest Rates in Monetary Economies. *The Review of Financial Studies 9*(1), 241–275.

Bakshi, G. S. and Z. Chen (1997). An Alternative Valuation Model for Contingent Claims. *Journal of Financial Economics 44*, 123–165.

Balduzzi, P., G. Bertola, and S. Foresi (1997). A Model of Target Changes and the Term Structure of Interest Rates. *Journal of Mathematical Economics 39*, 223–249.

Başak, S. and D. Cuoco (1998). An Equilibrium Model with Restricted Stock Market Participation. *The Review of Financial Studies 11*(2), 309–341.

Berk, J., R. Green, and V. Naik (1999). Optimal Investment, Growth Options, and Security Returns. *The Journal of Finance 54*, 1553–1608.

Bernoulli, D. (1954). Exposition of a New Theory on the Measurement of Risk. *Econometrica 22*(1), 23–36. Translation of the 1738 version.

Bielecki, T. R. and M. Rutkowski (2002). *Credit Risk: Modeling, Valuation and Hedging*. Springer.

Björk, T. (2004). *Arbitrage Theory in Continuous Time* (Second ed.). Oxford University Press.

Black, F. (1976). The Pricing of Commodity Contracts. *Journal of Financial Economics 3*, 167–179.

Black, F. and J. Cox (1976). Valuing Corporate Securities: Some Effects of Bond Indenture Provisions. *The Journal of Finance 31*, 351–367.

Black, F. and M. Scholes (1973). The Pricing of Options and Corporate Liabilities. *Journal of Political Economy 81*(3), 637–654.

Bossaerts, P. (2002). *The Paradox of Asset Pricing*. Princeton University Press.

Brace, A., D. Gatarek, and M. Musiela (1997). The Market Model of Interest Rate Dynamics. *Mathematical Finance 7*(2), 127–155.

Brav, A., G. M. Constantinides, and G. C. Geczy (2002). Asset Pricing with Heterogenous Consumers and Limited Participation: Empirical Evidence. *Journal of Political Economy 110*, 793–824.

Breeden, D. T. (1979). An Intertemporal Asset Pricing Model with Stochastic Consumption and Investment Opportunities. *Journal of Financial Economics 7*, 265–296.

Breeden, D. T. (1986). Consumption, Production, Inflation and Interest Rates. *Journal of Financial Economics 16*, 3–39.

Breeden, D. T., M. R. Gibbons, and R. H. Litzenberger (1989). Empirical Tests of the Consumption-Oriented CAPM. *The Journal of Finance 44*, 231–262.

Brennan, M. J., A. W. Wang, and Y. Xia (2004). Estimation and Test of a Simple Model of Intertemporal Capital Asset Pricing. *The Journal of Finance 59*, 1743–1775.

Brigo, D. and F. Mercurio (2001). *Interest Rate Models – Theory and Practice*. Springer-Verlag.

Brown, S., W. Goetzmann, and S. A. Ross (1995). Survival. *The Journal of Finance 50*, 853–873.

Browning, M. (1991). A Simple Nonadditive Preference Structure for Models of Household Behavior over Time. *Journal of Political Economy 99*(3), 607–637.

Campbell, J. Y. (1986). A Defense of Traditional Hypotheses About the Term Structure of Interest Rates. *The Journal of Finance 41*, 617–630.

Campbell, J. Y. and J. H. Cochrane (1999). By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior. *Journal of Political Economy 107*, 205–251.

Campbell, J. Y., A. W. Lo, and A. C. MacKinlay (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Chan, Y. L. and L. Kogan (2002). Catching Up with the Joneses: Heterogeneous Preferences and the Dynamics of Asset Prices. *Journal of Political Economy 110*(6), 1255–1285.

Chen, N.-f., R. Roll, and S. A. Ross (1986). Economic Forces and the Stock Market. *Journal of Business 59*, 383–403.

Cicchetti, C. J. and J. A. Dubin (1994). A Microeconomic Analysis of Risk Aversion and the Decision to Self-Insure. *Journal of Political Economy 102*(1), 169–186.

Cochrane, J. H. (2001). *Asset Pricing*. Princeton University Press.

Constantinides, G. M. (1990). Habit Formation: A Resolution of the Equity Premium Puzzle. *Journal of Political Economy 98*, 519–543.

Constantinides, G. M. (1992). A Theory of the Nominal Term Structure of Interest Rates. *The Review of Financial Studies 5*(4), 531–552.

Constantinides, G. M., J. B. Donaldson, and R. Mehra (2002). Junior Can't Borrow: A New Perspective on the Equity Premium Puzzle. *Quarterly Journal of Economics 117*, 269–296.

Constantinides, G. M. and D. Duffie (1996). Asset Pricing with Heterogeneous Consumers. *Journal of Political Economy 104*(2), 219–240.

Cont, R. and P. Tankov (2004). *Financial Modeling with Jump Processes*. Chapman & Hall.

Cootner, P. H. (Ed.) (1964). *The Random Character of Stock Market Prices*. MIT Press.

Cox, J. C., J. E. Ingersoll, Jr., and S. A. Ross (1981a). A Re-examination of Traditional Hypotheses about the Term Structure of Interest Rates. *The Journal of Finance 36*(4), 769–799.

Cox, J. C., J. E. Ingersoll, Jr., and S. A. Ross (1981b). The Relation between Forward Prices and Futures Prices. *Journal of Financial Economics 9*, 321–346.

Cox, J. C., J. E. Ingersoll, Jr., and S. A. Ross (1985a). An Intertemporal General Equilibrium Model of Asset Prices. *Econometrica 53*(2), 363–384.

Cox, J. C., J. E. Ingersoll, Jr., and S. A. Ross (1985b). A Theory of the Term Structure of Interest Rates. *Econometrica 53*(2), 385–407.

Cox, J. C., S. A. Ross, and M. Rubinstein (1979). Option Pricing: A Simplified Approach. *Journal of Financial Economics 7*, 229–263.

Culbertson, J. M. (1957, November). The Term Structure of Interest Rates. *Quarterly Journal of Economics*, 489–504.

Danthine, J.-P. and J. B. Donaldson (2002). *Intermediate Financial Theory*.

Debreu, G. (1954). Valuation Equilibrium and Pareto Optimum. *Proceedings of the National Academy of Sciences 40*, 588–592.

Detemple, J. B. and F. Zapatero (1991). Asset Prices in An Exchange Economy with Habit Formation. *Econometrica 59*, 1633–1658.

Dothan, M. U. (1990). *Prices in Financial Markets*. Oxford University Press.

Drèze, J. (1971). Market Allocation under Uncertainty. *European Economic Review 15*, 133–165.

Dubra, J. and F. Echenique (2004). Information is not about Measurability. *Mathematical Social Sciences 47*, 177–185.

Duffie, D. (2001). *Dynamic Asset Pricing Theory* (Third ed.). Princeton University Press.

Duffie, D. and L. G. Epstein (1992a). Asset Pricing with Stochastic Differential Utility. *The Review of Financial Studies 5*(3), 411–436.

Duffie, D. and L. G. Epstein (1992b). Stochastic Differential Utility. *Econometrica 60*(2), 353–394.

Duffie, D. and M. Huang (1996). Swap Rates and Credit Quality. *The Journal of Finance 51*(3), 921–949.

Duffie, D. and K. Singleton (2003). *Credit Risk: Pricing, Measurement, and Management*. Princeton University Press.

Duffie, D. and R. Stanton (1992). Pricing Continuously Resettled Contingent Claims. *Journal of Economic Dynamics and Control 16*, 561–574.

Dybvig, P. H. and C.-f. Huang (1988). Nonnegative Wealth, Absence of Arbitrage, and Feasible Consumption Plans. *The Review of Financial Studies 1*(4), 377–401.

Epstein, L. G. and S. E. Zin (1989). Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework. *Econometrica 57*(4), 937–969.

Epstein, L. G. and S. E. Zin (1991). Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: An Empirical Analysis. *Journal of Political Economy 99*, 263–286.

Fama, E. F. (1970). Multiperiod Consumption-Investment Decisions. *American Economic Review 60*(1), 163–174. Correction: Fama (1976).

Fama, E. F. (1976). Multiperiod Consumption-Investment Decisions: A Correction. *American Economic Review 66*(4), 723–724.

Fama, E. F. (1981). Stock Returns, Real Activity, Inflation, and Money. *American Economic Review 71*, 545–565.

Fama, E. F. and K. R. French (1996). The CAPM is Wanted, Dead or Alive. *The Journal of Finance 51*(5), 1947–1958.

Fama, E. F. and M. Gibbons (1982). Inflation, Real Returns and Capital Investment. *Journal of Mathematical Economics 9*, 297–323.

Fishburn, P. (1970). *Utility Theory for Decision Making*. John Wiley and Sons.

Fisher, I. (1896). Appreciation and Interest. *Publications of the American Economic Association*, 23–29 and 88–92.

Fisher, M. and C. Gilles (1998). Around and Around: The Expectations Hypothesis. *The Journal of Finance 52*(1), 365–383.

Fleming, W. H. and H. M. Soner (1993). *Controlled Markov Processes and Viscosity Solutions*, Volume 25 of *Applications of Mathematics*. New York: Springer-Verlag.

Friend, I. and M. E. Blume (1975). The Demand for Risky Assets. *American Economic Review 65*(5), 900–922.

Geman, H. (1989). The Importance of the Forward Neutral Probability in a Stochastic Approach of Interest Rates. Working paper, ESSEC.

Goldstein, R. and F. Zapatero (1996). General Equilibrium with Constant Relative Risk Aversion and Vasicek Interest Rates. *Mathematical Finance 6*, 331–340.

Gollier, C. (2001). *The Economics of Risk and Time*. MIT Press.

Gomes, J., L. Kogan, and L. Zhang (2003). Equilibrium Cross-Section of Reutrns. *Journal of Political Economy 111*, 693–732.

Grether, D. M. and C. R. Plott (1979). Economic Theory of Choice and the Preference Reversal Phenomenon. *American Economic Review 69*(4), 623–638.

Grossman, S. J., A. Melino, and R. J. Shiller (1987). Estimating the Continuous-Time Consumption-Based Asset-Pricing Model. *Journal of Business & Economic Statistics 5*(3), 315–327.

Hakansson, N. H. (1970). Optimal Investment and Consumption Strategies Under Risk for a Class of Utility Functions. *Econometrica 38*(5), 587–607.

Hansen, L. P. and R. Jagannathan (1991). Implications of Security Market Data for Models of Dynamic Economies. *Journal of Political Economy 99*, 225–262.

Hansen, L. P. and S. F. Richard (1987). The Role of Conditioning Information in Deducing Testable Restrictions Implied by Dynamic Asset Pricing Models. *Econometrica 55*, 587–614.

Harrison, J. M. and D. M. Kreps (1979). Martingales and Arbitrage in Multiperiod Securities Markets. *Journal of Economic Theory 20*, 381–408.

Harrison, J. M. and S. R. Pliska (1981). Martingales and Stochastic Integrals in the Theory of Continuous Trading. *Stochastic Processes and their Applications 11*, 215–260. Addendum: Harrison and Pliska (1983).

Harrison, J. M. and S. R. Pliska (1983). A Stochastic Calculus Model of Continuous Trading: Complete Markets. *Stochastic Processes and their Applications 15*, 313–316.

He, H. and D. M. Modest (1995). Market Frictions and Consumption-Based Asset Pricing. *Journal of Political Economy 103*(11), 94–117.

Heston, S. L. (1993). A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *The Review of Financial Studies 6*(2), 327–343.

Hicks, J. R. (1939). *Value and Capital*. Oxford: Clarendon Press.

Huang, C.-f. and R. H. Litzenberger (1988). *Foundations for Financial Economics*. Prentice-Hall.

Huge, B. and D. Lando (1999). Swap Pricing with Two-Sided Default Risk in a Rating-Based Model. *European Finance Review 3*(3), 239–268.

Hull, J. and A. White (1987). The Pricing of Options on Assets with Stochastic Volatility. *The Journal of Finance 42*(2), 281–300.

Hull, J. C. (2006). *Options, Futures, and Other Derivatives* (6th ed.). Prentice-Hall, Inc.

Ingersoll, Jr., J. E. (1987). *Theory of Financial Decision Making*. Rowman & Littlefield.

James, J. and N. Webber (2000). *Interest Rate Modelling*. Wiley.

Jamshidian, F. (1987). Pricing of Contingent Claims in the One Factor Term Structure Model. Working paper, Merrill Lynch Capital Markets.

Jamshidian, F. (1989). An Exact Bond Option Formula. *The Journal of Finance 44*(1), 205–209.

Jamshidian, F. (1997). LIBOR and Swap Market Models and Measures. *Finance and Stochastics 1*, 293–330.

Kan, R. (1992, June). Shape of the Yield Curve under CIR Single Factor Model: A Note. Working paper, University of Chicago.

Karlin, S. and H. M. Taylor (1981). *A Second Course in Stochastic Processes*. Academic Press, Inc.

Korn, R. and H. Kraft (2001). A Stochastic Control Approach to Portfolio Problems with Stochastic Interest Rates. *SIAM Journal on Control and Optimization 40*(4), 1250–1269.

Kothari, S. P., J. Shanken, and R. G. Sloan (1995). Another Look at the Cross-Section of Expected Stock Returns. *The Journal of Finance 50*, 185–224.

Kraft, H. (2004). Optimal Portfolios with Stochastic Short Rate: Pitfalls when the Short Rate is Non-Gaussian or the Market Price of Risk is Unbounded. Working paper, University of Kaiserslautern. Available at SSRN: `http://ssrn.com/abstract=666385`.

Kreps, D. M. (1990). *A Course in Microeconomic Theory*. Harvester Wheatsheaf.

Kreps, D. M. and E. Porteus (1978). Temporal Resolution of Uncertainty and Dynamic Choice Theory. *Econometrica 46*, 185–200.

Lando, D. (2004). *Credit Risk Modeling*. Princeton University Press.

Lettau, M. and S. Ludvigson (2001). Resurrecting the (C)CAPM: A Cross-Sectional Test when Risk Premia Are Time-Varying. *Journal of Political Economy 109*, 1238–1287.

Liew, J. and M. Vassalou (2000). Can Book-to-Market, Size, and Momentum Be Risk Factors that Predict Economic Growth? *Journal of Financial Economics 57*, 221–245.

Lintner, J. (1965). The Valuation of Risky Assets and the Selection of Risky Investment in Stock Portfolios and Capital Budgets. *Review of Economics and Statistics 47*, 13–37.

Lo, A. W. and A. C. MacKinlay (1990). Data-Snooping Biases in Tests of Financial Asset Pricing Models. *The Review of Financial Studies 3*, 431–467.

Longstaff, F. A. and E. S. Schwartz (1992). Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model. *The Journal of Finance 47*(4), 1259–1282.

Lustig, H. N. and S. G. van Nieuwerburgh (2005). Housing Collateral, Consumption Insurance, and Risk Premia: An Empirical Perspective. *The Journal of Finance 60*(3), 1167–1219.

Lutz, F. (1940, November). The Structure of Interest Rates. *Quarterly Journal of Economics*, 36–63.

Madan, D. B., P. P. Carr, and E. C. Chang (1998). The Variance-Gamma Process and Option Pricing. *European Finance Review 2*, 79–105.

Margrabe, W. (1978). The Value of an Option to Exchange One Asset for Another. *The Journal of Finance 33*(1), 177–198.

Markowitz, H. (1952). Portfolio Selection. *The Journal of Finance 7*, 77–91.

Markowitz, H. (1959). *Portfolio Selection: Efficient Diversification of Investment*. Wiley.

Marshall, D. (1992). Inflation and Asset Returns in a Monetary Economy. *The Journal of Finance 47*, 1315–1342.

McCulloch, J. H. (1993). A Reexamination of Traditional Hypotheses about the Term Structure of Interest Rates: A Comment. *The Journal of Finance 48*, 779–789.

Mehra, R. and E. C. Prescott (1985). The Equity Premium: A Puzzle. *Journal of Monetary Economics 15*, 145–162.

Mehra, R. and E. C. Prescott (2003). The Equity Premium in Retrospect. In G. M. Constantinides, M. Harris, and R. Stulz (Eds.), *Handbook of the Economics of Finance*, Volume 1B, Chapter 14. Elsevier.

Menzly, L., T. Santos, and P. Veronesi (2002). The Time Series of the Cross Section of Asset Prices. Working paper, University of Chicago.

Merton, R. C. (1969). Lifetime Portfolio Selection Under Uncertainty: The Continuous-Time Case. *Review of Economics and Statistics 51*, 247–257. Reprinted as Chapter 4 in Merton (1992).

Merton, R. C. (1971). Optimum Consumption and Portfolio Rules in a Continuous-Time Model. *Journal of Economic Theory 3*, 373–413. Erratum: Merton (1973a). Reprinted as Chapter 5 in Merton (1992).

Merton, R. C. (1973a). Erratum. *Journal of Economic Theory 6*, 213–214.

Merton, R. C. (1973b). An Intertemporal Capital Asset Pricing Model. *Econometrica 41*(5), 867–887. Reprinted in an extended form as Chapter 15 in Merton (1992).

Merton, R. C. (1973c). Theory of Rational Option Pricing. *Bell Journal of Economics and Management Science 4*(Spring), 141–183. Reprinted as Chapter 8 in Merton (1992).

Merton, R. C. (1976). Option Pricing When Underlying Stock Returns are Discontinuous. *Journal of Financial Economics 3*, 125–144. Reprinted as Chapter 9 in Merton (1992).

Merton, R. C. (1992). *Continuous-Time Finance*. Padstow, UK: Basil Blackwell Inc.

Miltersen, K. R., K. Sandmann, and D. Sondermann (1997). Closed Form Solutions for Term Structure Derivatives with Log-Normal Interest Rates. *The Journal of Finance 52*(1), 409–430.

Modigliani, F. and R. Sutch (1966, May). Innovations in Interest Rate Policy. *American Economic Review*, 178–197.

Mossin, J. (1966). Equilibrium in a Capital Asset Market. *Econometrica 35*, 768–783.

Munk, C. (2005a). Dynamic Asset Allocation. Lecture notes, University of Southern Denmark.

Munk, C. (2005b). Fixed-Income Analysis: Securities, Pricing, and Risk Management. Lecture notes, University of Southern Denmark.

Musiela, M. and M. Rutkowski (1997). *Martingale Methods in Financial Modelling*, Volume 36 of *Applications of Mathematics*. Springer-Verlag.

Negishi, T. (1960). Welfare Economics and Existence of an Equilibrium for a Competetive Economy. *Metroeconometrica 12*, 92–97.

Nielsen, L. T. and M. Vassalou (2006). The Instantaneous Capital Market Line. *Economic Theory 28*, 651–664.

Ogaki, M. and Q. Zhang (2001). Decreasing Relative Risk Aversion and Tests of Risk Sharing. *Econometrica 69*(2), 515–526.

Øksendal, B. (1998). *Stochastic Differential Equations* (Fifth ed.). Springer-Verlag.

Omberg, E. (1989). The Expected Utility of the Doubling Strategy. *The Journal of Finance 44*(2), 515–524.

Petkova, R. (2006). Do the Fama-French Factors Proxy for Innovations in Predictive Variables? *The Journal of Finance 61*(2), 581–612.

Piazzesi, M. (2001). An Econometric Model of the Yield Curve with Macroeconomic Jump Effects. Working paper, UCLA and NBER.

Piazzesi, M., M. Schneider, and S. Tuzel (2006). Housing, Consumption, and Asset Pricing. Forthcoming in *Journal of Financial Economics*.

Pindyck, R. S. (1988). Risk Aversion and the Determinants of Stock Market Behavior. *Review of Economic Studies 70*(2), 183–190.

Pratt, J. (1964). Risk Aversion in the Small and the Large. *Econometrica 32*, 122–136.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). *Numerical Recipes in C* (Second ed.). Cambridge University Press.

Riedel, F. (2000). Decreasing Yield Curves in a Model with an Unknown Constant Growth Rate. *European Finance Review 4*(1), 51–67.

Riedel, F. (2004). Heterogeneous Time Preferences and Humps in the Yield Curve – The Preferred Habitat Theory Revisited. *The European Journal of Finance 10*(1), 3–22.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton, New Jersey: Princeton University Press.

Roll, R. (1977). A Critique of the Asset Pricing Theory's Tests. *Journal of Financial Economics 4*, 129–176.

Ross, S. A. (1976). The Arbitrage Theory of Capital Asset Pricing. *Journal of Economic Theory 13*, 341–360.

Ross, S. A. (1978). A Simple Approach to the Valuation of Risky Streams. *Journal of Business 51*, 453–475.

Rubinstein, M. (1974). An Aggregation Theorem for Securities Markets. *Journal of Financial Economics 1*, 225–244.

Rubinstein, M. (1976). The Strong Case for the Generalized Logarithmic Utility Model as the Premier Model of Financial Markets. *The Journal of Finance 31*(2), 551–571.

Samuelson, P. A. (1969). Lifetime Portfolio Selection by Dynamic Stochastic Programming. *Review of Economics and Statistics 51*, 239–246.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley.

Schönbucher, P. (2003). *Credit Derivatives Pricing Models: Models, Pricing, and Implementation*. Wiley.

Sharpe, W. (1964). Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance 19*, 425–442.

Storesletten, K., C. I. Telmer, and A. Yaron (2004). Cyclical Dynamics in Idiosyncratic Labor Market Risk. *Journal of Political Economy 112*(3), 695–717.

Sydsaeter, K. and P. J. Hammond (2005). *Essential Mathematics for Economic Analysis* (2 ed.). Prentice-Hall, Pearson Education.

Sydsaeter, K., P. J. Hammond, A. Seierstad, and A. Strom (2005). *Further Mathematics for Economic Analysis* (1 ed.). Prentice-Hall, Pearson Education.

Sydsaeter, K., A. Strom, and P. Berck (2000). *Economists' Mathematical Manual* (3 ed.). Springer-Verlag.

Szpiro, G. G. (1986). Measuring Risk Aversion: An Alternative Approach. *Review of Economic Studies 68*(1), 156–159.

Telmer, C. I. (1993). Asset-Pricing Puzzles and Incomplete Markets. *The Journal of Finance 48*(5), 1803–1832.

Vasicek, O. (1977). An Equilibrium Characterization of the Term Structure. *Journal of Financial Economics 5*, 177–188.

Vassalou, M. (2003). News Related to Future GDP Growth as a Risk Factor in Equity Returns. *Journal of Financial Economics 68*, 47–73.

von Neumann, J. and O. Morgenstern (1944). *Theory of Games and Economic Behavior*. New Jersey: Princeton University Press.

Wachter, J. A. (2006). A Consumption-Based Model of the Term Structure of Interest Rates. *Journal of Financial Economics 79*, 365–399.

Wang, J. (1996). The Term Structure of Interest Rates in a Pure Exchange Economy with Heterogeneous Investors. *Journal of Financial Economics 41*, 75–110.

Weil, P. (1989). The Equity Premium Puzzle and the Risk-free Rate Puzzle. *Journal of Mathematical Economics 24*(3), 401–421.

Wilcox, D. W. (1992). The Construction of U.S. Consumption Data: Some Facts and Their Implications for Empirical Work. *American Economic Review 82*(4), 922–941.

Yogo, M. (2006). A Consumption-Based Explanation of Expected Stock Returns. *The Journal of Finance 61*(2), 539–580.