Essential Mathematics for Market Risk Management

For other titles in the Wiley Finance series
please see www.wiley.com/finance

Essential Mathematics for Market Risk Management

Simon Hubbert

WILEY

A John Wiley & Sons, Ltd., Publication

This edition first published 2012

Registered office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Hubbert, Simon.
Essential mathematics for market risk management / Simon Hubbert. – 2nd ed.
p. cm. – (The Wiley finance series)
Includes bibliographical references and index.
ISBN 978-1-119-97952-4 (hardback)
1. Risk management – Mathematical models. 2. Capital market – Mathematical models. I. Title.
HD61.H763 2012
658.15′50151 – dc23

2011039267

A catalogue record for this book is available from the British Library.

ISBN 978-1-119-97952-4 (hardback) ISBN 978-1-119-95301-2 (ebk)
ISBN 978-1-119-95302-9 (ebk) ISBN 978-1-119-95303-6 (ebk)

Set in 10/12pt Times by Laserwords Private Limited, Chennai, India
Printed in Great Britain by CPI Group (UK) Ltd, Croydon, CR0 4YY

For my parents, Michelle and Nancy.
And dedicated to the memory of my brother, Craig.

Contents

Preface

The aim of this book is to provide the reader with a clear exposition of some of the fundamental mathematical tools and techniques that are frequently used in financial risk management. The book has been written with a wide audience in mind. For instance, it should appeal to numerate graduates who seek an accessible and self-contained account of the science behind the evolving story of financial risk management. In addition, it should also be of interest to the market practitioner who is interested in gaining a deeper understanding of the mathematical theory which underpins some of the most commonly used quantitative (black-box) techniques.

Most of the existing books devoted to financial risk management tend to fall into two categories, those that tackle a large number of topics with only brief overviews of the mathematical ideas (e.g., Hull (2007), Dowd (2002) and Jorion (2006)) and, on the other hand, rigorous mathematical expositions that are too advanced for an introductory level (e.g., McNeil, Frey and Embrechts (2005) and Moix (2001)). In view of this I have designed this book to occupy the middle ground, namely one that delivers an accessible yet thorough mathematical account of a broad sweep of carefully selected topics that an experienced risk manager is likely to encounter on a regular basis. In order to maintain focus I have devoted the book entirely to the mathematics of market risk management; there are already a whole host of excellent texts that cover the science of credit risk management, Bielecki, and Rutkowski (2010) and Schönbucher (2003) being excellent examples. The book, as its title suggests, is focused firmly on the essential mathematics of the subject and so, by design, it should equip the reader with the required scientific background to either embark on a rewarding career in risk management or to study the subject at a more advanced level. In particular, it is hoped that this text will serve as a useful companion to Alexander (2008a), Alexander (2008b) and Christoffersen (2003); three excellent books which place the emphasis firmly on practical examples and implementation.

The book itself has evolved from two courses on risk management that I teach regularly at Birkbeck, University of London. Both courses form part of a wider qualification in financial engineering, one at graduate diploma level and the other at masters level. The graduate diploma courses at Birkbeck are aimed at students who are familiar with basic calculus, linear algebra and probability theory, and they are designed to serve as a stepping stone to the more technically demanding masters level courses. Students who take this route invariably perform extremely well and, in view of this, the book represents a blend of introductory material (from the graduate diploma) and advanced topics (from the masters course). The

field of market risk management is so vast that one could devote an entire textbook to several of its sub-branches (e.g., volatility modelling, simulation methods, extreme value theory) and thus I do not claim that this text represents an exhaustive account of state-of-the-art topics in this field. However, it is hoped that the book will inspire the reader to go on and investigate these topics in more depth.

It is a pleasure to thank the people who have helped make this book possible. I would like to acknowledge my colleagues Brad Baxter and Raymond Brummelhuis at Birkbeck for their support and encouragement. I also gratefully appreciate many of my past students for their valuable feedback on the structure and content of the book; special thanks go to Mafalda Alabort Jordan who provided many of the figures that appear in Chapter 19.

Essential Mathematics for Market Risk Management
by Simon Hubbert

1 Introduction

In life we simply cannot avoid the presence of risk. However, we tend to avoid its potential impact because, on the whole, we do a good job of risk management; we wear a bicycle helmet when cycling, we fasten our seat belts in a moving car, we use gloves when handling corrosive substances, etc. In the world of financial investments the universally held view is that the more risk we take the more we stand to gain but, just as importantly, the greater the chance we will lose. The task of the financial risk manager is to be aware of the presence of risk, to understand how it can damage a potential investment and, most of all, to be able to reduce the exposure to it in order to avert a potential disaster. It is the aim of this book to develop the mathematical tools which can be used to manage and control risks that are inherent in the financial market. We will be guided by two basic principles. Firstly, we shall endeavour to ensure that, on average, a financial investment provides a healthy return rate for a tolerable amount of risk. Secondly, we shall be prepared for rare market events whose impact could trigger a potentially catastrophic loss. The purpose of this chapter is to shed light on both the day-to-day issues and also the big challenges that a typical risk manager is likely to face, thus it serves as aperitif to stimulate the mathematical journey ahead.

1.1 BASIC CHALLENGES IN RISK MANAGEMENT

We open our discussion by considering a seemingly simple problem. Assume that we are armed with a wealth of $\$W$ and we decide to invest this today, at time t, in a single financial asset for a period of τ days into the future. The value of the asset today is known and denoted by $S(t)$ but its future value $S(t+\tau)$ is uncertain. We think of our asset being a simple market product such as a share in a stock, an amount of foreign currency or the ownership of a bond or some other commodity. In this situation there are two possible strategies:

- The holding strategy.
  If we believe the asset price will rise then we simply buy it today and sell it in the future at the (hopefully) higher price. In which case we make a profit from the purchase.
- The short-selling strategy.
  If we believe the asset price will fall then we can profit out of this situation by employing the strategy of short selling. This is summarized as follows.

t	$t+\tau$
Borrow the asset today and sell it immediately to receive $S(t)$	Buy the asset for $S(t+\tau)$ and return it to the lender

If, as we suspected, the value of the asset falls (i.e., if $S(t+\tau) < S(t)$) then we have made a profit.

The risk profiles of the two strategies are very different. The asset price can never drop below zero but, theoretically, it can grow without bound. For the holding strategy this means potentially unlimited profits and a bounded loss. However, for short selling the reverse is true and there is a potential for unlimited losses. In view of this one finds that, in practice, the process of pure short selling is supplemented by certain restrictions and safeguards.

We now suppose that we choose to invest our $\$W$ in a collection of n risky assets denoted by $\{S_1, \ldots, S_n\}$. Our strategy is simply to invest a fraction of our wealth, w_i say, in asset S_i for $i = 1, \ldots, n$. We shall assume that short selling is allowed and so some of the w_i may be negative. This scenario leads us to our next challenging problem:

The Portfolio Problem

How can we choose an optimal set of weights $\{w_1, \ldots, w_n\}$*, so that our overall investment is likely to yield a promising return with minimal risk?*

Occasionally in mathematics one finds that seemingly complex problems have the most elegant and rewarding solutions. The portfolio problem above is such an example and it is the perfect starting point for our mathematical journey through risk management. The problem itself was solved in the early 1950s by Harry Markowitz (1952) in his PhD studies. The route that Markowitz took to derive his famous solution is as follows:

- Establish a formula for the random return rate for the portfolio, denoted by r_p, as a function of its weights $w_1, \ldots, w_n$.
- Use basic probability theory to derive expressions for the expected return μ_p (a measure of potential reward) and the volatility σ_p (a measure of risk) of r_p.
- We now search for the weights $w_1^*, \ldots, w_n^*$ that provide a desired level of expected return while ensuring that the risk involved is as small as possible.

The mathematical tools needed to attack this problem are developed in Chapters 2–4 and its full solution is delivered in Chapter 5; this will represent our first major landmark result.

Before Markowitz's theory emerged most investment decisions were made on the basis of gut instinct or simple advice such as *don't put all of your eggs in one basket*, there was little in the way of quantitative analysis. Markowitz gave investment theory a scientific footing and, in Chapter 6, we will discover some intriguing consequences of his pioneering work. Indeed these discoveries subsequently inspired many other researchers to investigate more deeply the relationship between the value of an asset and its perceived riskiness. This is a tough problem and one that is made even more difficult by the fact that asset prices do not always move of their own accord. More often than not we find that asset prices are related to each other. Strong price fluctuations in one asset will influence the movements of another and vice versa, we say they possess a correlation structure. This leads us to address the following.

The Modelling Challenge

How can we accurately model the way the price of a risky asset evolves through time?

The Correlation Challenge

How can we accurately model the correlation structure of a collection of many risky assets?

In the early 1960s Markowitz encouraged a PhD student, William Sharpe, to investigate these problems. To do this Sharpe imagined a world where all investors build their portfolios with Markowitz weights and, in this setting, he developed the famous Capital Asset Pricing Model (CAPM) Sharpe (1964). Chapter 7 of this book is devoted to the mathematical derivation of this model. We shall demonstrate some of its practical uses and its consequences, including the follwing intriguing discovery:

Asset prices are related to each other through their responses to a ***single*** *risk factor.* (1.1)

This revelation tells us that, in the Markowitz world, a single known risk factor can be viewed as the main driving force behind the movements and co-movements of all our risky assets. This conclusion is a remarkable one and, not surprisingly, it fuelled much debate amongst financial economists. Indeed, a great deal of empirical work has been done over the years to test the validity of the CAPM and its underlying assumptions.

In the 1970s a more cavalier approach to the development of financial risk models was taken. Specifically, inspired by the CAPM, the following more general situation was considered:

Asset prices are related to each other through their responses to ***several*** *risk factors.* (1.2)

In response to the above hypothesis a more general class of risk model was proposed, the so-called linear factor model. We will examine this popular approach in greater detail in Chapter 8 of this book. The most appealing feature of the linear factor model is the fact that there is a great deal of flexibility in the choice and composition of the driving factors. This flexibility leads us to an important practical risk management challenge.

The Factor Selection Challenge

How do we choose the number and nature of the driving risk factors?

We shall conclude Chapter 8 by describing how principal components analysis, a famous dimension-reduction tool from multivariate statistics, can be used to deliver a useful scientific solution to this challenge.

1.2 VALUE AT RISK

In the late 1980s fund managers and traders with complicated risk positions looked increasingly to a new breed of so-called derivative products as a means of dampening their risk profiles. Derivatives are literally products that are derived from simpler assets like those we have already encountered (i.e., stocks and shares, commodities, foreign currencies and bonds). When used correctly derivatives are able to protect those who hold them against risk; they can be viewed as a kind of insurance policy. However, as their popularity began to rise it became clear that the misuse of these products can have devastating consequences. Indeed, throughout the mid-1990s a whole host of derivatives-related disasters finally led to a much needed shake-up in the way banks were regulated. New tighter controls were imposed on financial institutions and consequently the industry as a whole had to rethink its approach to

risk management. In the present day all financial institutions have dedicated research teams of applied scientists who employ sophisticated mathematical and statistical methods to quantify and control exposure to risk. The risk-management revolution was initiated in the early 1990s when the famous Basel committee (on banking supervision) began a consultation process which, essentially, set about addressing the following important questions.

Ensuring Against Large Losses

How can investment banks measure their exposure to unfavourable and unanticipated movements of the basic financial assets?

How can they use this measure to determine their capital adequacy requirements?

In order to attack this problem the committee proposed that each investment bank should divide its market positions into two books, the trading book and the bank book. The trading book, as its name suggests, contains all products that are used as part of an active day-to-day trading strategy (e.g., investment portfolios and derivatives would belong in the trading book). In contrast, the bank book consists of positions that are held over a much longer time horizon such as long-term loans.

The Basel committee directed its attention on the trading book and investigated how its riskiness could be quantified. The value of each product in the trading book has a price which can be discovered on the market (provided there is enough liquidity). The prices of these products in the future however are unknown, and thus, even though we may know the value of the trading book today, its value tomorrow or at any time in the future is unknown. When market conditions are calm one would hope that the trading portfolio would report a daily profit or at least only a mild, manageable loss. However, we cannot control market conditions and history dictates that, once in a while, we can expect a financial storm where an increase in market volatility can wipe away significant value from a financial product. In view of this a natural question to ask could be the following:

What is the largest loss the trading book is likely to suffer 99 out of every 100 days?

The answer to this question is known as the Value at Risk (VaR) for the trading book at the 99% confidence level; obviously the same question can be posed for other confidence levels, e.g., 95% represents the maximum likely loss in 95 out of 100 days. The idea of measuring the VaR of a portfolio is popular with practitioners; it represents a potential monetary loss and, in that respect, it is concise, practical and easy to understand. In 1996, the Basel committee added their own support to the VaR concept by proposing that banks could use VaR as a measure of its trading book's exposure to market risk. The final Basel committee report is viewed as pioneering for two reasons:

1. It endorses that investment banks can use their own internal models to calculate VaR estimates.
2. It provides all investment banks with a universal formula which they can use to calculate their own capital adequacy requirements; the formula is based upon the bank's own VaR estimates.

Value at Risk is widely regarded as one of the key milestones in the new risk-management revolution. However, the simplicity of the VaR concept disguises the complexity involved

in its measurement. For instance, before a single computation takes place we need to ensure that we have access to all relevant financial data, both historical and real time. Thus, a typical financial institution faces the following significant task:

The IT Challenge

Construct an IT system with the following functionality:

- *Real-time position data for all products in the trading portfolio are gathered and correctly mapped to the risk calculation engine.*
- *A database that is dynamically populated with historical prices at regular intervals (e.g., daily prices) is accessible.*

This IT challenge is enormous, especially for multinational investment banks whose trading portfolio consists of products that span the global markets. Not surprisingly most investment banks choose to hand these data management projects over to one of the many IT consultancy firms with specialized skills in database architecture.

The VaR concept can be viewed as the trigger for a new approach to risk management; indeed, it marks the starting point of an exciting area of science where academic progress and real-world applications are in constant exchange. We consider the VaR calculation challenge in two parts.

The VaR Calculation Challenge

For a given confidence level $\alpha \in (0, 1)$ *how can we measure the corresponding Value at Risk for a portfolio which consists entirely of:*

1. *Basic financial assets such as stocks and shares, commodities, foreign currencies and bonds.*
2. *More complex derivatives products.*

We take up the first part of the VaR challenge in Chapter 9 where we examine its mathematical properties. We shall discover some of VaR's enticing features, however we also reveal some unfortunate problems. We endeavour to correct these problems by investigating alternative risk measures, and ask whether such candidates can be viewed as serious competitors to VaR.

In Chapter 10 we turn to the practical calculation of VaR and its associated challenges. As a starting point, we propose a basic model which assumes that the random variable representing the daily portfolio loss is normally distributed. In particular, for this simplified framework, we will show how we can derive neat closed-form solutions to almost all of the crucial VaR-related challenges.

The second part of the VaR challenge involves an additional level of complexity as we now allow derivative products to be included in the portfolio. In order to attack this problem we need some advanced results from probability theory and statistics and these are developed in Chapters 11 and 12. At the simplest level we can invest in a single derivative whose value depends upon the price of its underlying asset. Mathematically we say that the derivative price is a function of the asset price and write

$$\text{derivative price at } t = f(S(t)),$$

where f is some non-linear function. In order to examine the potential profit/loss associated with the derivative we need to determine, as accurately as possible, the form of f. This leads us to our next challenge:

Derivative Pricing

For a given derivative how can we determine the relationship between its value and the level of the underlying asset?

Derivative pricing is a branch of mathematical finance in its own right and there are a whole host of excellent textbooks written on this subject (e.g., Higham (2004), Joshi (2005), Neftci (1996) and Wilmott, Howison and Dewynne (1995)). However, in Chapter 13 we provide a self-contained derivation of the celebrated Black–Scholes option pricing model for the simplest plain European options. This model dates back to the early 1970s and yet its impact on the development of modern mathematical finance cannot be overstated; a great deal of the pioneering work on derivative pricing can be viewed as an extension or an innovation of the original Black–Scholes model.

We will not pursue derivative pricing in any further depth, but will simply assume that a calculation engine exists and is able to deliver a price for any derivative we encounter. In this situation we are able to tackle the problem of computing VaR estimates for a portfolio of derivatives. In a deliberate effort to reduce the computational burden of this problem we shall investigate the possibility of providing a closed-form solution. We remark that this problem is difficult for at least two reasons:

1. The number of underlying assets (upon which the derivatives are written) can be very large, i.e., the problem is a high-dimensional one.
2. Even if we understand the probabilistic nature of a particular asset it is much harder to predict how a non-linear function (i.e., a derivative) of it will behave.

In the late 1990s Britten-Jones and Schaeffer (1999) tackled the above issues and proposed the following recipe:

- Step 1. Dimension reduction.
  A linear factor model is proposed as a model for the changes in the underlying asset returns. It is assumed that the number of factors is much smaller than the number of assets and thus the size (dimension) of the problem is greatly reduced.
- Step 2. Probabilistic assumptions.
  Some simplifying assumptions are proposed for the probabilistic laws that govern the random nature of the risk factors.
- Step 3. Function approximations.
  A local approximation of the non-linear derivative function is made.

In Chapter 14 we will develop the above steps in detail and show how the local approximations to the derivatives can be used to provide closed-form expressions for so-called non-linear Value at Risk.

1.3 FURTHER CHALLENGES IN RISK MANAGEMENT

The early attempts to calculate VaR were made in the mid-1990s and, during this time, the main priority for most practitioners was to establish a straightforward solution that could

be implemented with ease. As a result these early attempts were based upon rather simple assumptions regarding the random behaviour of the financial losses/returns. Towards the end of the 1990s almost all financial institutions took advantage of the rapid advances in information technology, where faster computing speed coupled with increased data storage enabled teams of quantitative analysts to perform deeper scientific investigations. A particularly important example is to use historical data to help gain an insight into the characteristic properties of the underlying financial variables; indeed, this becomes the focus of our next challenge:

Statistical Investigation

Using realized price data, perform a statistical investigation to determine the key empirical properties of asset losses/returns.

In Chapters 15–18 we develop the statistical tools and techniques needed to tackle this problem. Then, in Chapter 19, we put these tools into action and conduct a scientific investigation whose aim is to pin down the key statistical properties that characterize the true nature of financial losses/returns. These properties are commonly referred to as the stylized facts and they serve as a guide for the development of new and improved risk models; a successful mathematical model should capture as many (if not all) of these properties as possible.

One particular result of our investigation is the observation that extreme values tend to occur more often than some of the basic models would predict, with large losses occurring more often than large profits. In relation to this we also discover that the future volatility of a basic financial asset is closely related to its past. This is an important observation because it implies that when an asset experiences a period of high volatility the likelihood of an extreme swing is increased; unfortunately, the swing can be downward as well as upward. These observations lead us to one of the central questions that all risk managers must address:

The Volatility Challenge

How can we construct a time-dependent volatility model which accurately captures the stylized facts of financial losses?

The topic of volatility modelling is so large that it can also be regarded as a branch of mathematical finance in its own right, indeed there are several textbooks devoted to this topic (e.g., Gouriéroux (1997), Poon (2005) and Taylor (2007)). We take on the volatility challenge in Chapter 20 where, rather than provide a bite-size review of the many different approaches, we present the mathematical story of one of the most popular, the so-called GARCH family of models. GARCH models have found a wide range of applications in financial modelling because, as we shall discover, they have the ability to capture almost all of the stylized facts and, what's more, they are also fairly simple to implement. The GARCH modelling framework is also extremely flexible; once one understands the basic model, it is then possible to introduce extensions designed to enhance its performance. This is reflected in the vast range of innovative GARCH-type volatility models on the market.

In order to motivate the next important challenge we recall that our Value at Risk measure, as we know it, is designed to cope with those unanticipated events which typically occur two

or three times in a year. Unfortunately, however, experience has shown that financial markets can also be exposed to tornado-like events such as terrorist attacks, political instabilities and natural disasters. These events have the potential to wipe billions off the value of global stock markets. Thus, one of the new challenges of mathematical risk management is to develop a methodology to cater for such extreme events. In this respect we face a new challenge:

The Challenge of Quantifying Losses Due to Rare Events

How do we assign appropriate probabilities to potential extreme movements of a financial asset?

We tackle this problem in Chapter 21 where we appeal to extreme value theory (EVT), a branch of probability theory that is concerned with describing the statistical properties of extreme events. EVT has applications in many areas of science and engineering. In particular, hydrologists have successfully used EVT to help predict the likelihood and size of potentially damaging floods, the hydrologist then uses these findings to estimate the optimal height of a dam which is to be constructed to protect against such floods. In finance the application of EVT is much the same; the risk manager uses the theory to model the likelihood and size of a portfolio loss due to a financial storm, he can then use this data to determine the size of the buffer fund which is designed to absorb such losses.

A common situation in finance (and other branches of science and engineering) is that closed-form solutions to real-world problems only tend to be available in the simplest of cases. For instance, the fair price of a plain European option can be derived analytically, however numerical methods are needed for most non-standard options. We find this in risk management too, for instance if we (erroneously) assume that the portfolio loss random variable is normally distributed then we can derive expressions for almost any risk measure, however if a more sophisticated risk model is used then we must turn to numerical techniques. This leads us to our next challenge:

Numerical Methods for Risk Quantification

How can we develop numerical techniques to compute the Value at Risk for a financial portfolio?

In order to address this problem we must study the mathematical ideas behind one of the most crucial numerical tools in risk management – the ability to perform numerical simulations. We take on this challenge in Chapter 22 where we demonstrate how simulation techniques can be used to deliver estimates of financial risk measures such as Value at Risk. In particular we describe how to design a simulation algorithm whose purpose is to generate a range of potential future prices for each asset and/or derivative in the portfolio, these are then combined to produce a simulated value of the portfolio. Then, as more and more simulated values are generated, a clearer picture of the crucial statistical properties of the portfolio loss random variable emerges and, as a result, estimates for VaR (and other risk measures) can easily be derived. The success of the method lies in the specification of the algorithm. It may depend upon the past price history of the portfolio (historical simulation) or it may depend upon some mathematical model that is calibrated to real market prices (Monte Carlo simulation).

Obviously, from a practitioner's perspective, an accurate closed-form expression for VaR is highly desirable. Indeed, in the late 1990s several alternative VaR methodologies were proposed, each delivering closed-form solutions while attempting to simultaneously capture the true statistical properties of the loss random variable. In Chapter 23 we shall present two of the most commonly used methods and by doing so we bring the story of VaR calculation methods to a close.

At the end of the day the model that is finally selected to compute the VaR of the trading portfolio is of particular importance. The resulting VaR calculations will determine the size of the institution's buffer fund, the amount of regulatory capital it must set aside to help absorb unexpected losses. The buffer fund cannot be used for investment purposes, it is off-limits and its size must adhere to the regulator's formula. If we choose a model that consistently overestimates the true VaR then we will be overcommitting funds that could otherwise be used to generate profits. On the other hand, if we choose a model that consistently underestimates the true VaR then it will be punished; the regulator will revise its formula so that the size of the buffer fund is increased, i.e., the regulator penalizes a substandard VaR model. In view of these two influences we must select the most accurate calculation method to suit the characteristics of our trading portfolio, i.e., we must face the following challenge:

Verification of Risk Models

How can we scientifically test the performance of a particular VaR model?

We address this question in Chapter 24, the final chapter of the book. Specifically we develop Christoffersen's testing methodology Christoffersen (1998) that dates back to the late 1990s. The idea here is to appeal to our earlier review of statistical testing (Chapter 18) and use this to propose a certain test statistic whose value is dependent upon the past performance of the VaR model. The test statistic itself can be viewed as a random quantity which obeys certain known probability laws and we can use this fact to construct a decision rule that determines whether the model should be accepted or rejected.

Essential Mathematics for Market Risk Management
by Simon Hubbert

2 Applied Linear Algebra for Risk Managers

Many of the problems in risk management are said to be high dimensional because they involve a large number of underlying variables. For instance, problems involving financial portfolios are high dimensional because a portfolio is made up of many financial assets and its value is determined by the monetary amounts that are invested in these assets. Applied linear algebra is the branch of mathematics that provides the framework needed to set up and solve these problems and, in this chapter, we present the most crucial results.

2.1 VECTORS AND MATRICES

The fundamental objects of applied linear algebra are vectors and matrices. A vector, by definition, is a column of real numbers, i.e., elements of $\mathbb{R}$. The number of entries in the column is called the dimension of the vector and we write

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n \text{ is an } n\text{-dimensional vector.}$$

Suppose we have a collection of m vectors, each of the same dimension:

$$\left\{ \mathbf{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}, \mathbf{a}_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix}, \ldots, \mathbf{a}_m = \begin{pmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{pmatrix} \right\}.$$

A matrix is a means of displaying all of this information in one object. We write

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \in \mathbb{R}^{n \times m},$$

and say that $\mathbf{A}$ is an $n \times m$ matrix as it consists of n rows (the dimension of the columns) and m columns (the number of vectors on display).

Any matrix is completely defined by its elements and, in general, if $\mathbf{A} \in \mathbb{R}^{n \times m}$ then we let $\mathbf{A}_{kl}$ denote the entry appearing in the kth row and lth column; commonly called the

(k, l)th entry. Using this notation the matrix above can be defined in a much more compact way by writing

$$\mathbf{A} \in \mathbb{R}^{n\times m} \text{ such that } \mathbf{A}_{kl} := a_{kl} \text{ for } (1 \le k \le n),\ (1 \le l \le m).$$

One of the simplest operations one can perform on a matrix is to turn its rows into columns and vice versa; the result is called the transpose and we write

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}^T = \underbrace{\begin{pmatrix} a_{11} & a_{21} & \cdots & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & \cdots & a_{m2} \\ \vdots & \vdots & & & \vdots \\ a_{1m} & a_{2m} & \cdots & \cdots & a_{mn} \end{pmatrix}}_{=\mathbf{A}^T \in \mathbb{R}^{m\times n}},$$

or, in compact form, $\mathbf{A}^T \in \mathbb{R}^{m\times n}$ defined by

$$\left(\mathbf{A}^T\right)_{kl} = \mathbf{A}_{lk} = a_{lk} \text{ for } (1 \le l \le m), (1 \le k \le n).$$

We note that an n-dimensional column vector can be viewed as an $n \times 1$ matrix. It is often useful to refer to a column as the transpose of its row vector, i.e., we shall often see

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n (= \mathbb{R}^{n\times 1}).$$

We are now in a position to assemble our first working toolkit of important facts, definitions and handy results of applicable linear algebra.

A Basis of Spanning Vectors

In geometry, whenever we describe the location of a point in the plane or in three-dimensional space we generally do so with reference to a standard coordinate system, i.e., we write

$$\begin{pmatrix} x \\ y \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \end{pmatrix} \text{ for 2-dimensional space } \mathbb{R}^2$$

and

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = x \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} + z \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \text{ for 3-dimensional space } \mathbb{R}^3.$$

This idea extends to higher-dimensional spaces and, for a general $n > 1$, we define

$$\mathcal{U} = \left\{ \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \mathbf{e}_2 = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{pmatrix}, \ldots, \mathbf{e}_n = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix} \right\}. \tag{2.1}$$

The set $\mathcal{U}$ is said to form the standard basis for n-dimensional space because any n-dimensional vector $\mathbf{x}$ can be written uniquely as a linear combination of the elements of $\mathcal{U}$, i.e.,

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n.$$

There are other collections of vectors that can also serve as an n-dimensional basis, however the standard basis $\mathcal{U}$ is the simplest of its kind. In general we say that a collection of n vectors

$$\mathcal{B} = \left\{ \mathbf{a}_1 = \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix}, \mathbf{a}_2 = \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix}, \ldots, \mathbf{a}_n = \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{nn} \end{pmatrix} \right\}$$

forms a basis for $\mathbb{R}^n$ if, for each vector $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$, there exists a unique coordinate vector $\mathbf{c} = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$, such that

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = c_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} + c_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix} + \cdots + c_n \begin{pmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{nn} \end{pmatrix},$$

or, in more compact form, that

$$\mathbf{x} = \sum_{j=1}^{n} c_j \mathbf{a}_j \quad \text{for a unique } \mathbf{c} = (c_1, \ldots, c_n)^T \in \mathbb{R}^n.$$

In order to investigate whether the coordinate vector is unique we suppose the contrary, i.e., assume that, for a given $\mathbf{x} \in \mathbb{R}^n$ there are two distinct coordinate vectors $\mathbf{c}$ and $\mathbf{c}'$ such that

$$\mathbf{x} = \sum_{j=1}^{n} c_j \mathbf{a}_j = \sum_{j=1}^{n} c'_j \mathbf{a}_j.$$

If this is the case we can subtract the two representations to deduce that the zero vector $\mathbf{0} = (0, \ldots, 0)^T \in \mathbb{R}^n$ also has more than one coordinate representation, for instance

$$\mathbf{0} = \sum_{j=1}^{n} (c_j - c'_j)\mathbf{a}_j \quad \text{and} \quad \mathbf{0} = \sum_{j=1}^{n} 0 \cdot \mathbf{a}_j.$$

We can deduce from this that a vector $\mathbf{x} \in \mathbb{R}^n$ has a unique coordinate vector provided the equation

$$\sum_{j=1}^{n} c_j \cdot \mathbf{a}_j = 0 \quad \text{implies} \quad c_1 = c_2 = \cdots = c_n = 0. \tag{2.2}$$

A collection of column vectors $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ that satisfies (2.2) is said to form a **linearly independent** set. Thus, any set of n linearly independent vectors (of dimension n) serves as a basis for $\mathbb{R}^n$.

The Magnitude of a Vector

The magnitude of a vector $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ is denoted by $\|\mathbf{x}\|$ and is defined by

$$\|\mathbf{x}\| = \sqrt{x_1^2 + \cdots + x_n^2}. \tag{2.3}$$

We collect the following related facts:

- We say that $\mathbf{x}$ is a unit vector if $\|\mathbf{x}\| = 1$. We note that the standard basis vectors (2.1) are simple examples of unit vectors.
- The quantity $d(\mathbf{x}, \mathbf{y})$ denotes the distance between any two n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, and is defined by
$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|.$$

Inner and Outer Product of Two Vectors

The inner product between two n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ is defined by

$$\begin{aligned} < \mathbf{x}, \mathbf{y} > = \mathbf{x}^T \mathbf{y} &= (x_1, \ldots, x_n) \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &= x_1 y_1 + \cdots + x_n y_n = \sum_{j=1}^{n} x_j y_j \in \mathbb{R}. \end{aligned} \tag{2.4}$$

We collect the following properties:

- The inner product is symmetric, i.e., $< \mathbf{x}, \mathbf{y} > = < \mathbf{y}, \mathbf{x} >$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.
- The inner product is linear, i.e., if $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ are n-dimensional vectors then
$$< \alpha\mathbf{x} + \beta\mathbf{y}, \mathbf{z} > = \alpha < \mathbf{x}, \mathbf{z} > + \beta < \mathbf{y}, \mathbf{z} > \quad \alpha, \beta \in \mathbb{R}.$$
- The inner product of a vector $\mathbf{x} \in \mathbb{R}^n$ with itself is the square of its magnitude, i.e.,
$$< \mathbf{x}, \mathbf{x} > = \sum_{j=1}^{n} x_j^2 = \|x\|^2.$$

- If we take the inner product of a vector $\mathbf{x}$ with one of the standard basis vectors $\mathbf{e}_k \in \mathcal{U}$ we find

$$< \mathbf{x}, \mathbf{e}_k > = x_k, k = 1, \ldots, n, \tag{2.5}$$

  i.e., the result is the kth entry of $\mathbf{x}$.
- There exists an angle $\theta \in [-\pi, \pi)$ such that the following equation holds:

$$< \mathbf{x}, \mathbf{y} > = \|\mathbf{x}\| \|\mathbf{y}\| \cos\theta.$$

  We say that θ is the angle between $\mathbf{x}$ and $\mathbf{y}$.
- In view of the above, we say that two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are

$$\text{orthogonal if } \theta = \pm\pi/2, \text{ i.e., if } < \mathbf{x}, \mathbf{y} > = 0.$$

  Furthermore, we say that they are

$$\text{orthonormal if } < \mathbf{x}, \mathbf{y} > = 0 \text{ and } \|\mathbf{x}\| = \|\mathbf{y}\| = 1.$$

The outer product of two n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ is defined by

$$\begin{aligned} \mathbf{x} \otimes \mathbf{y} = \mathbf{x}\mathbf{y}^T &= \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} (y_1, \ldots, y_n) \\ &= \begin{pmatrix} x_1 y_1 & \cdots & x_1 y_n \\ \vdots & & \vdots \\ x_n y_1 & \cdots & x_n y_n \end{pmatrix} \in \mathbb{R}^{n \times n}. \end{aligned} \tag{2.6}$$

We note that while the inner product of two vectors is a scalar, the outer product is an $n \times n$ matrix whose entries are defined by

$$(\mathbf{x} \otimes \mathbf{y})_{kl} = x_k x_l \quad \text{for } 1 \leq k, l \leq n.$$

Multiplication with Matrices and Vectors

There are many ways of transforming an n-dimensional vector $\mathbf{x}$; we have seen, for example, that taking the inner product of $\mathbf{x}$ with another vector results in a real number, whereas the outer product results in a matrix.

If we multiply $\mathbf{x}$ by a $1 \times n$ matrix, i.e., a row vector, then the calculation we are faced with is the same as the inner product; we have

$$(a_1, \ldots, a_n) \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = (\text{row vector}) \cdot (\text{column vector}) = \sum_{i=1}^{n} a_i x_i.$$

The result of multiplying $\mathbf{x}$ by an $m \times n$ matrix $\mathbf{A}$ is the m-dimensional vector whose elements are computed by multiplying $\mathbf{x}$ by each row of $\mathbf{A}$ from top to bottom. To demonstrate this, we have

$$\begin{aligned}\mathbf{Ax} &= \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} \\ &= \begin{pmatrix} (\text{row } 1 \text{ of } \mathbf{A}) \cdot \mathbf{x} \\ (\text{row } 2 \text{ of } \mathbf{A}) \cdot \mathbf{x} \\ \vdots \\ (\text{row } m \text{ of } \mathbf{A}) \cdot \mathbf{x} \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j} x_j \\ \sum_{j=1}^{n} a_{2j} x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj} x_i \end{pmatrix}. \end{aligned} \tag{2.7}$$

It is often helpful to represent a matrix in a more compact form. In particular, we can write

$$\mathbf{A} = (\mathbf{a}_1 \mathbf{a}_2 \cdots \mathbf{a}_n)$$

to signify that $\mathbf{A}$ is the $m \times n$ matrix whose columns are $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^m$. With this notation we can develop (2.7) to show that

$$\mathbf{Ax} = \begin{pmatrix} \sum_{j=1}^{n} a_{1j} x_j \\ \sum_{j=1}^{n} a_{2j} x_j \\ \vdots \\ \sum_{j=1}^{n} a_{mj} x_j \end{pmatrix} = \sum_{j=1}^{n} x_j \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix} = \sum_{j=1}^{n} x_j \mathbf{a}_j. \tag{2.8}$$

We remark that if $\mathbf{x}$ corresponds to one of the standard basis vectors $\mathbf{e}_k \in \mathcal{U}$ $(1 \leq k \leq n)$ then we find that

$$\mathbf{Ae}_k = \begin{pmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{nk} \end{pmatrix} \qquad 1 \leq k \leq n, \tag{2.9}$$

i.e., matrix multiplication of the kth standard basis vector picks out the kth column of the matrix $\mathbf{A}$.

The above development establishes that a matrix $\mathbf{A}$, with n columns, transforms an n-dimensional vector $\mathbf{x}$ into a (potentially) new vector. Indeed, if the matrix has m rows then the vector $\mathbf{x}$ is transformed into the m-dimensional vector $\mathbf{Ax}$, defined by (2.7).

The column vector $\mathbf{x}$ is simply an $n \times 1$ matrix, thus the multiplication law follows

$$(m \times n) \text{ matrix } \mathbf{A} \cdot (n \times 1) \text{ matrix } \mathbf{x} = (m \times 1) \text{ matrix } \mathbf{Ax},$$

such that

$$j\text{th entry of } \mathbf{Ax} = (\text{row } j \text{ of } \mathbf{A}) \cdot \mathbf{x}, \quad j = 1, \ldots, m.$$

We can easily extend this definition to show how two matrices multiply together. Given that $\mathbf{A}$ is an $m \times n$ matrix, let $\mathbf{B}$ denote an $n \times p$ matrix, then the matrix product $\mathbf{AB}$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{pmatrix}$$

results in an $n \times p$ matrix $\mathbf{AB}$ according to the law

$$(m \times n) \text{ matrix } \mathbf{A} \cdot (n \times p) \text{ matrix } \mathbf{B} = (m \times p) \text{ matrix } \mathbf{AB},$$

where the (k, l)th element of $\mathbf{AB}$ is defined by

$$\begin{aligned} (\mathbf{AB})_{kl} &= (\text{row } k \text{ of } \mathbf{A}) \cdot (\text{column } l \text{ of } \mathbf{B}) \\ &= \sum_{i=1}^{n} \mathbf{A}_{ki}\mathbf{B}_{il} = \sum_{i=1}^{n} a_{ki}b_{il} \qquad 1 \le k \le m,\ 1 \le l \le p. \end{aligned}$$

We now investigate the transpose of the product $\mathbf{AB}$, where we find

$$\begin{aligned} (\mathbf{AB})_{kl}^{T} = (\mathbf{AB})_{lk} &= \sum_{i=1}^{n} \mathbf{A}_{li}\mathbf{B}_{ik} = \sum_{i=1}^{n} \mathbf{B}_{ik}\mathbf{A}_{li} \\ &= \sum_{i=1}^{n} \mathbf{B}_{ki}^{T}\mathbf{A}_{il}^{T} = \left(\mathbf{B}^{T}\mathbf{A}^{T}\right)_{kl}. \end{aligned}$$

Thus, we have discovered the following very useful result:

$$\text{if } \mathbf{A} \in \mathbb{R}^{m \times n} \text{ and } \mathbf{B} \in \mathbb{R}^{n \times p} \text{ then } (\mathbf{AB})^{T} = \mathbf{B}^{T}\mathbf{A}^{T} \in \mathbb{R}^{p \times n}. \tag{2.10}$$

2.2 MATRIX ALGEBRA IN PRACTICE

Our motivation for studying the properties of matrices arises from the fact that there are so many situations in finance where we are called upon to solve some system of m linear equations with n unknowns of the form

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & y_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & y_2 \\ \vdots & & \vdots & & & & \vdots & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & y_m. \end{array}$$

In matrix–vector form this system is equivalent to

$$\mathbf{Ax} = \begin{pmatrix} a_{11} & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & \cdots & a_{2n} \\ \vdots & \vdots & & & \vdots \\ a_{m1} & a_{m2} & \cdots & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \mathbf{y}.$$

The general problem we face is to find a vector $\mathbf{x}$ that solves the above equation. If we let $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ denote the m-dimensional column vectors of $\mathbf{A}$ then we can recast the problem as follows:

$$\text{express } \mathbf{y} \in \mathbb{R}^m \text{ as a linear combination of } \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \mathbb{R}^m,$$

i.e., our task amounts to finding a coordinate vector $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ such that

$$\mathbf{y} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \cdots + x_n\mathbf{a}_n.$$

In general, there is no guarantee that we are able to solve this problem. Indeed, the nature of the problem depends upon the size of the matrix $\mathbf{A}$ and on the properties of its column vectors. We consider three cases:

- $m = n$
  In this case the number of equations matches the number of variables. The existence of a unique solution depends crucially on the defining matrix $\mathbf{A}$; specifically, if its columns $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n$ form a basis for $\mathbb{R}^n$, i.e., if they are linearly independent, then a unique solution exists.
  If, on the other hand, it can be shown that one or more of the columns can be expressed as a linear combination of the others then there can be no unique solution.
- $m < n$
  In this case there are fewer equations than there are variables, we say the system is under-determined. There can be no unique solution to such a problem but there may be many different solutions; so here the task is to effectively filter through these solutions to find a suitable one.
- $m > n$
  In this case there are more equations than there are variables, we say the system is over-determined. There may be no exact solution to this problem and so we resort to finding a best approximation.

Square Matrices

If a matrix $\mathbf{A}$ has the same number of rows and columns we say that it is square and write $\mathbf{A} \in \mathbb{R}^{n \times n}$, where n is sometimes called the size of the matrix. A simple example of a square matrix is the diagonal matrix, so called because it is defined by the entries that populate its main diagonal; all other entries are zero. We write

$$\mathbf{D} = \text{diag } (\lambda_1, \lambda_2 \ldots, \lambda_n) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

An important diagonal matrix arises when all of the diagonal entries equal one, we call this the identity matrix and write

$$\mathbf{I}_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

A triangular matrix is another simple example of a square matrix. As its name suggests this matrix is defined by its entries that appear in the triangle above or below the main diagonal; all remaining entries are zero. There are two possible types.

- The lower triangular matrix:

$$\mathbf{L} = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix} \quad \text{where} \quad \mathbf{L}_{kl} = 0 \quad k < l.$$

- The upper triangular matrix:

$$\mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{pmatrix} \quad \text{where} \quad \mathbf{U}_{kl} = 0 \quad k > l.$$

We now introduce a few characteristics that a square matrix may or may not display.

- **Symmetry**
  A square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ is said to be symmetric if it is the transpose of itself, i.e., if

$$\mathbf{A} = \mathbf{A}^T \quad \Rightarrow \quad \mathbf{A}_{kl} = \mathbf{A}_{lk} \quad 1 \leq k, l \leq n.$$

- **Invertibility**
  If $\mathbf{A} \in \mathbb{R}^{n \times n}$ and there exists a matrix $\mathbf{A}^{-1} \in \mathbb{R}^{n \times n}$ that satisfies

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n = \mathbf{A}\mathbf{A}^{-1}$$

  then we say that $\mathbf{A}$ is invertible (or non-singular) and that the matrix $\mathbf{A}^{-1}$ is its inverse.
  We note that if $\mathbf{A}$ is invertible then we can display the solution of a linear system explicitly, that is,

$$\text{the vector } \mathbf{x} = \mathbf{A}^{-1}\mathbf{y} \text{ uniquely solves the linear system } \mathbf{A}\mathbf{x} = \mathbf{y}.$$

Thus, in view of our previous discussion on linear systems we can deduce the following result:

$$\begin{aligned} \mathbf{A} \in \mathbb{R}^{n \times n} \text{ is invertible} &\Leftrightarrow \mathbf{A}\mathbf{x} = \mathbf{0} \text{ implies } \mathbf{x} = \mathbf{0} \\ &\Leftrightarrow \text{its column vectors } \mathbf{a}_1, \ldots, \mathbf{a}_n \\ &\quad \text{are linearly independent.} \end{aligned} \tag{2.11}$$

- **Orthogonality**
  An invertible matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is said to be orthogonal if its inverse and its transpose coincide, i.e., if
  $$\mathbf{A}^{-1} = \mathbf{A}^T, \text{ i.e., } \mathbf{A}^T\mathbf{A} = \mathbf{I}_n = \mathbf{A}\mathbf{A}^T.$$
  Thus, we note that columns of an orthogonal matrix are not only linearly independent but also orthonormal.

Symmetric Bilinear Forms

Suppose we are given an $n \times n$ symmetric matrix $\mathbf{A}$. We can use this to construct a useful function $B_{\mathbf{A}} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ defined by the rule

$$B_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) = \mathbf{y}^T\mathbf{A}\mathbf{x} =< \mathbf{y}, \mathbf{A}\mathbf{x} > . \tag{2.12}$$

We say that B_A is the symmetric bilinear form based on $\mathbf{A}$. As its name suggests, B_A displays the following symmetry property:

$$< \mathbf{y}, \mathbf{A}\mathbf{x} > = B_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) = B_{\mathbf{A}}(\mathbf{y}, \mathbf{x}) =< \mathbf{A}\mathbf{y}, \mathbf{x} > . \tag{2.13}$$

Another interesting observation arises if we let $\mathbf{x} = \mathbf{e}_k$ and $\mathbf{y} = \mathbf{e}_l$ denote any two elements of the standard basis (2.1), then, using (2.5) and (2.9), the resulting bilinear form provides the (k, l)th element of $\mathbf{A}$, i.e.,

$$B_{\mathbf{A}}(\mathbf{e}_k, \mathbf{e}_l) = \mathbf{e}_l^T\mathbf{A}\mathbf{e}_k = a_{kl}. \tag{2.14}$$

We can use the definition of the inner product (2.4) and identity (2.7) to show that the bilinear form can be expressed as a double sum, as follows:

$$\begin{aligned} B_{\mathbf{A}}(\mathbf{x}, \mathbf{y}) &= (y_1, \ldots, y_n) \begin{pmatrix} \sum_{j=1}^n a_{1j}x_j \\ \sum_{i=1}^n a_{2j}x_j \\ \vdots \\ \sum_{j=1}^n a_{nj}x_j \end{pmatrix} \\ &= \sum_{i=1}^n y_i \sum_{j=1}^n a_{ij}x_j \\ &= \sum_{i=1}^n \sum_{j=1}^n a_{ij}y_i x_j. \end{aligned}$$

Finally, we can also define the quadratic form based on $\mathbf{A}$; this is a function $Q_{\mathbf{A}} : \mathbb{R}^n \to \mathbb{R}$ that arises as a special case of the bilinear form, and is given by

$$Q_{\mathbf{A}}(\mathbf{x}) = B_{\mathbf{A}}(\mathbf{x}, \mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} =< \mathbf{x}, \mathbf{A}\mathbf{x} >, \tag{2.15}$$

or, equivalently, as the double sum

$$Q_{\mathbf{A}}(x_1, \ldots, x_n) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j. \tag{2.16}$$

Bilinear and quadratic forms arise naturally in many areas of finance and engineering. The quadratic form is often used as a local approximation to a more complicated function; it can be viewed as the higher-dimensional analogue of the more familiar 1-dimensional quadratic function $q(x) = ax^2$.

2.3 EIGENVECTORS AND EIGENVALUES

We can transform any n-dimensional vector $\mathbf{x}$ into a new vector by multiplying it by a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$. Unfortunately, when the dimension n is greater than 3 it is impossible to visualize such a transformation. There are, however, some powerful ideas from linear algebra that enable us to gain some insight into this higher-dimensional world.

One of the easiest transformations we can make is to simply scale our n-dimensional vector by some number $\lambda \in \mathbb{R}$, i.e., $\mathbf{v} \mapsto \lambda \cdot \mathbf{v}$. A remarkable result tells us that, given any symmetric matrix $\mathbf{A}$, there exists a special collection of vectors for which the matrix transformation

$$\mathbf{v} \mapsto \mathbf{A}\mathbf{v} \text{ is equivalent to } \mathbf{v} \mapsto \lambda \mathbf{v} \text{ for some } \lambda \in \mathbb{R}.$$

These vectors in some sense belong to the matrix $\mathbf{A}$ and, together with their associated scale factors, they play an extremely important role in many branches of mathematics. The following definition establishes the notation and terminology.

Definition 2.1. *Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix. A unit vector $\boldsymbol{v} \in \mathbb{R}^n$ is said to be an eigenvector of $\boldsymbol{A}$ if there exists a scalar λ such that*

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}.$$

In addition, we say that λ is the eigenvalue of $\boldsymbol{A}$ associated with $\boldsymbol{v}$.

Suppose $\mathbf{v}$ is an eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ and consider the following space:

$$\mathcal{E}_{\mathbf{v}} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \alpha \mathbf{v} \quad \text{for some } \alpha \in \mathbb{R}\}.$$

For any vector $\mathbf{x}(= \alpha \mathbf{v}) \in \mathcal{E}_{\mathbf{v}}$ we have that

$$\mathbf{A}\mathbf{x} = \mathbf{A}\alpha\mathbf{v} = \alpha\mathbf{A}\mathbf{v} = \lambda\alpha\mathbf{v} = \lambda\mathbf{x}.$$

The space $\mathcal{E}_{\mathbf{v}}$ can be visualized as an infinite straight line in $\mathbb{R}^n$, it emanates from the origin out to infinity in the direction of $\mathbf{v}$ and $-\mathbf{v}$. Any vector $\mathbf{x}$ that lies on this line (i.e., any $\mathbf{x} \in \mathcal{E}_{\mathbf{v}}$) also satisfies $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. We call $\mathcal{E}_{\mathbf{v}}$ the eigenspace of $\mathbf{A}$ generated by the eigenvector $\mathbf{v}$.

The following result highlights a useful characteristic of the eigenvectors of a matrix.

Theorem 2.2. *Let $\mathbf{A}$ be an $n \times n$ symmetric matrix. If $\mathbf{v}_1$ and $\mathbf{v}_2$ are two eigenvectors of $\mathbf{A}$ that correspond to two different non-zero eigenvalues λ_1 and λ_2 then $\mathbf{v}_1$ and $\mathbf{v}_2$ must be orthonormal.*

Proof. The eigenvectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are unit vectors by definition, thus we need to demonstrate that $< \mathbf{v}_1, \mathbf{v}_2 > = 0$. Consider the following development:

$$\begin{aligned}
\lambda_1 < \mathbf{v}_1, \mathbf{v}_2 > &= < \lambda_1 \mathbf{v}_1, \mathbf{v}_2 > = < \mathbf{A}\mathbf{v}_1, \mathbf{v}_2 > \\
&= B_A(\mathbf{v}_1, \mathbf{v}_2) \quad \text{symmetric bilinear form on } \mathbf{A} \\
&= B_A(\mathbf{v}_2, \mathbf{v}_1) \quad \text{using symmetry, see (2.13)} \\
&= < \mathbf{v}_1, \mathbf{A}\mathbf{v}_2 > = < \mathbf{v}_1, \lambda_2 \mathbf{v}_2 > = \lambda_2 < \mathbf{v}_1, \mathbf{v}_2 > .
\end{aligned}$$

Since the eigenvalues are distinct, the above equation implies that $< \mathbf{v}_1, \mathbf{v}_2 > = 0$ as required. □

The mathematical methods required to compute the eigenvalues of a symmetric matrix are beyond the scope of this book, but we mention here some helpful background facts.

- For each symmetric matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ we can construct a polynomial of degree n,

$$p_n(\lambda) = (-1)^n \lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1 \lambda + \alpha_0, \tag{2.17}$$

  whose n roots $\lambda_1, \ldots, \lambda_n$ are real numbers and coincide with the eigenvalues of $\mathbf{A}$, i.e., we have

$$p_n(\lambda) = (-1)^n (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n). \tag{2.18}$$

  We call p_n the characteristic polynomial of $\mathbf{A}$.
- The expressions for the coefficients of $p_n(\lambda)$ are rather complicated with the exception of α_{n-1}, where it can be shown that

$$\begin{aligned}
\alpha_{n-1} &= (-1)^{n-1}(a_{11} + \cdots + a_{nn}) \\
&= (-1)^{n-1} \,.\, (\text{sum of diagonal elements of } \mathbf{A}).
\end{aligned} \tag{2.19}$$

  We remark that the sum of the diagonal elements of a square matrix is often referred to as the trace of the matrix; we write

$$\text{Trace}(\mathbf{A}) = a_{11} + \cdots + a_{nn}, \tag{2.20}$$

  and conclude from (2.19) that

$$\alpha_{n-1} = (-1)^{n-1}\text{Trace}(\mathbf{A}). \tag{2.21}$$

  We observe that if the above factorization (2.18) were expanded then the coefficient multiplying λ^{n-1} is equal to

$$(-1)^{n-1}(\lambda_1 + \cdots + \lambda_n).$$

Thus, in view of (2.21) we see that, by matching coefficients, we have

$$\text{Trace}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i. \tag{2.22}$$

The concept of the trace of a matrix arises frequently in statistical theory; we will encounter it again in later chapters of this book. In view of this we present the following helpful result, which can easily be established by direct calculation.

Lemma 2.3. *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times m}$ *and* $\boldsymbol{B} \in \mathbb{R}^{m\times n}$*, then*

$$\text{Trace}(\boldsymbol{AB}) = \text{Trace}(\boldsymbol{BA}). \tag{2.23}$$

Proof. By direct calculation we see that

$$\begin{aligned}\text{Trace}(\mathbf{AB}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ji} \\ &= \sum_{j=1}^{m}\sum_{i=1}^{n} b_{ji}a_{ij} = \text{Trace}(\mathbf{BA}).\end{aligned}$$

□

- The determinant of a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is defined as the product of its eigenvalues, and we write

$$\det(\mathbf{A}) = \lambda_1 \cdots \lambda_n.$$

We remark that there are several equivalent definitions for the determinant of a matrix. The reason for its name is an illustrative one, if the determinant of a matrix is zero, i.e., if any of its eigenvalues is zero, then its inverse does not exist; otherwise the matrix is invertible.

- The determinant of a 2-dimensional matrix is particularly easy to compute. Indeed, for a general 2×2 matrix, not necessarily symmetric, we have

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc,$$

and when $\det(\mathbf{A}) \neq 0$ the inverse is given by

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{2.24}$$

We can now use the facts that we have established above to deduce that for each symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ there are n eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ associated with real-valued eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$. There is no guarantee, in general, that these eigenvalues are distinct and non-zero; however, when they are we can deduce from Theorem 2.2 that the matrix

$$\boldsymbol{\Gamma} = (\mathbf{v}_1\mathbf{v}_2 \cdots \mathbf{v}_n) \in \mathbb{R}^{n\times n}$$

is orthonormal, that is

$$\mathbf{\Gamma}^T\mathbf{\Gamma} = \mathbf{\Gamma}^T(\mathbf{v}_1\mathbf{v}_2\cdots\mathbf{v}_n) = (\mathbf{\Gamma}^T\mathbf{v}_1\mathbf{\Gamma}^T\mathbf{v}_2\cdots\mathbf{\Gamma}^T\mathbf{v}_n) = \mathbf{I}_n,$$

or equivalently,

$$\mathbf{\Gamma}^T\mathbf{v}_k = \mathbf{e}_k \quad k = 1, \ldots, n. \tag{2.25}$$

Armed with this vision we can now prove one of the most famous matrix decomposition results.

Theorem 2.4. *Let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix and let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n\}$ denote its eigenvectors associated with the eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$. Let $\mathbf{\Gamma} = (\boldsymbol{v}_1\boldsymbol{v}_2\cdots\boldsymbol{v}_n)$ and $\boldsymbol{D} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$. If the eigenvalues of $\boldsymbol{A}$ are non-zero and distinct then*

$$\mathbf{\Gamma}^T\boldsymbol{A}\mathbf{\Gamma} = \boldsymbol{D}, \text{ or equivalently, } \boldsymbol{A} = \mathbf{\Gamma}\boldsymbol{D}\mathbf{\Gamma}^T.$$

Proof.

$$\begin{aligned}
\mathbf{\Gamma}^T\mathbf{A}\mathbf{\Gamma} &= \mathbf{\Gamma}^T\mathbf{A}(\mathbf{v}_1\ \mathbf{v}_2\cdots\mathbf{v}_n) \\
&= \mathbf{\Gamma}^T(\mathbf{A}\mathbf{v}_1\ \mathbf{A}\mathbf{v}_2\cdots\mathbf{A}\mathbf{v}_n) \\
&= \mathbf{\Gamma}^T(\lambda_1\mathbf{v}_1\ \lambda_2\mathbf{v}_2\cdots\lambda_n\mathbf{v}_n) \\
&= \left(\lambda_1\mathbf{\Gamma}^T\mathbf{v}_1\ \lambda_2\mathbf{\Gamma}^T\mathbf{v}_2\cdots\lambda_n\mathbf{\Gamma}^T\mathbf{v}_n\right) \\
&= (\lambda_1\mathbf{e}_1\ \lambda_2\mathbf{e}_2\cdots\lambda_n\mathbf{e}_n) \quad \text{using (2.25)} \\
&= \text{diag}(\lambda_1,\ \lambda_2, \ldots, \lambda_n) = \mathbf{D}.
\end{aligned}$$

□

2.4 POSITIVE DEFINITE MATRICES

Positive definite matrices arise in many areas of finance and risk management. They are theoretically interesting objects to study and they have important computational properties, especially in optimization and data-fitting algorithms. We will frequently encounter these matrices and, by way of an introduction, we provide the definition and briefly compose a few interesting properties.

Definition 2.5. *A symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is said to be non-negative definite if its quadratic form $Q_{\boldsymbol{A}}$ satisfies*

$$Q_{\boldsymbol{A}}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0 \quad \text{for all vectors } \boldsymbol{x} \in \mathbb{R}^n.$$

If the quadratic form satisfies the stronger condition

$$Q_{\boldsymbol{A}}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0 \quad \text{for all non-zero vectors } \boldsymbol{x} \in \mathbb{R}^n$$

then we say that $\boldsymbol{A}$ is a positive definite matrix.

- If **A** is positive definite then all of its eigenvalues are positive numbers and hence it is invertible.
- The inverse of a positive definite matrix is also a positive definite matrix.
- If **A** is positive definite then all of its diagonal entries are positive, this follows from (2.14) since

$$\mathbf{e}_k^T \mathbf{A} \mathbf{e}_k = a_{kk} > 0 \quad \text{for } k = 1, \ldots, n.$$

We close this section, and hence the chapter, by presenting a result that captures one of the most useful properties of a positive definite matrix.

Theorem 2.6. *If a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive definite then there exists a unique lower triangular matrix $\boldsymbol{R}$, with positive diagonal entries, such that*

$$\boldsymbol{A} = \boldsymbol{R}\boldsymbol{R}^T. \tag{2.26}$$

The decomposition (2.26) of a positive definite matrix is often called the square root or Choleski decomposition, and the lower triangular matrix **R** is called the Choleski factor which, in view of (2.26), can be interpreted as the positive square root of **A**. We shall encounter the Choleski decompostion on several occasions in this book; it serves as a useful theoretical tool and, as we will discover in Chapter 22, it also plays a crucial role in the numerical computation of Value at Risk.

Essential Mathematics for Market Risk Management
by Simon Hubbert

3 Probability Theory for Risk Managers

The future of most financial investments is uncertain. We are unable to make statements such as

. . . in T days' time our investment WILL be worth. . .

Instead, the uncertain world is full of open questions such as

. . . what is the LIKELIHOOD that, in T days' time,
our investment will be worth. . .?

Probability theory is the branch of mathematics that studies the likelihood of random events and so provides us with the scientific approach we need to describe and analyse the nature of financial investments. In this chapter we set up a toolbox of rudimentary probability theory that enables us to make a start. As the story of mathematical risk management evolves, then so too will this toolbox.

3.1 UNIVARIATE THEORY

In the certain world we talk of a function f as an object that acts upon a set A of underlying variables. The function itself specifies a rule which governs how each $x \in A$ transforms into a (potentially) new value $f(x)$. We write

$$f : A \to \mathbb{R} \quad \text{by} \quad f : x \mapsto f(x).$$

3.1.1 Random variables

In the uncertain world there are no such rules that fix the outcome of a variable, we deal with random quantities that can take on any one of a range of possible values; some of these outcomes will be more likely to occur than others. We consider two distinct cases.

Discrete Random Variables

Here the random quantity X can take on any one of a finite number of specific values $x_1 < x_2 < \cdots < x_n$ say. We represent these n potential events as $\{X = x_j\}$ $(1 \leq j \leq n)$, and we assume that for each event there is a probability, denoted by

$$p_j = \mathbb{P}\{X = x_j\} \in [0, 1] \quad \text{for} \quad j = 1, \ldots, n. \tag{3.1}$$

These probabilities sum up to one and each one can be regarded as the relative frequency with which the event under inspection would occur if we were able to observe the variable infinitely often. In summary, we have the following definition:

Definition 3.1. *If X is a random quantity that can take on any one of the values $x_1 < x_2 < \cdots < x_n$ and there are numbers $p_1, p_2, \ldots, p_n$ defined by* (3.1) *that satisfy*

$$\sum_{j=1}^{n} p_j = 1,$$

then we say X is a ***discrete*** *random variable.*

One useful way to view the properties of a discrete random variable X is to use the function

$$F : \mathbb{R} \to [0, 1] \quad \text{given by} \quad F : x \mapsto \mathbb{P}\{X \leq x\}.$$

We note, from the above, that probability does not begin to accumulate until x hits x_1 (i.e., $F(x) = 0$ for $x < x_1$), then as x increases from x_1 to x_n the distribution $F(x)$ begins to pick up probability mass along the way and $F(x)$ increases accordingly. Finally, once x moves beyond x_n all probability mass has accumulated (i.e., $F(x) = 1$ for $x > x_n$). In view of this we say that F is the cumulative distribution function of X.

We can also capture all of the known information of a random variable X by defining its probability mass function

$$p(x) = \begin{cases} 0 & \text{if } x \in \mathbb{R} \setminus \{x_1, \ldots, x_n\} \\ \mathbb{P}\{X = x\} & \text{if } x \in \{x_1, \ldots, x_n\}. \end{cases} \tag{3.2}$$

There is an important mathematical link between these two functions, given by the formula

$$F(x) = \sum_{x_i \leq x} \mathbb{P}(X = x_i) = \sum_{x_i \leq x} p(x_i).$$

See Figure 3.1 for an illustration of this result.

Continuous Random Variables

We now turn attention to those random quantities that take on values on a continuum, such as an interval (a, b) or the whole real line $\mathbb{R}$. In contrast to the discrete case, we are not able to count the elements that make up the real line and so we are unable to assign a probability to events of the form $\{X = x\}$. Instead, we usually assign probabilities to events of the form $\{X \leq x\}$, i.e., we assess the likelihood that the outcome of the random variable will be bounded from above. The values of all these probabilities are captured through the distribution function of X; $F(x) = \mathbb{P}[X \leq x]$, for $x \in \mathbb{R}$. The distribution function of any random variable, whether it be discrete or continuous, possesses three defining properties, these are summarized in the following definition.

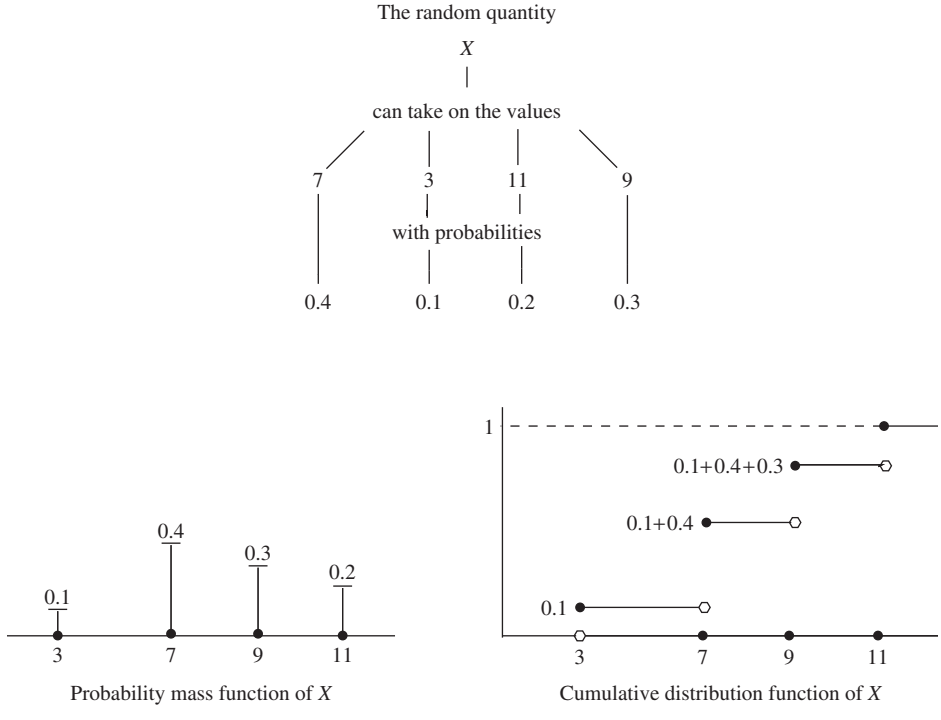


Figure 3.1 Mass and cumulative distribution functions of a discrete random variable.

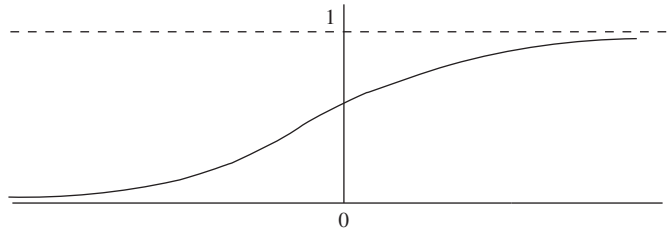
Definition 3.2. *A function* $F : \mathbb{R} \to [0, 1]$ *is the distribution function of some random variable if and only if it has the following properties:*

1. *F is non-decreasing (i.e.,* $F(x) \leq F(y)$ *whenever* $x < y$*).*
2. *F is right-continuous, i.e.,*
$$\lim_{\varepsilon \searrow 0} F(x + \varepsilon) = F(x).$$
3. *F is normalized, i.e.,*
$$\lim_{x \to -\infty} F(x) = 0 \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1.$$

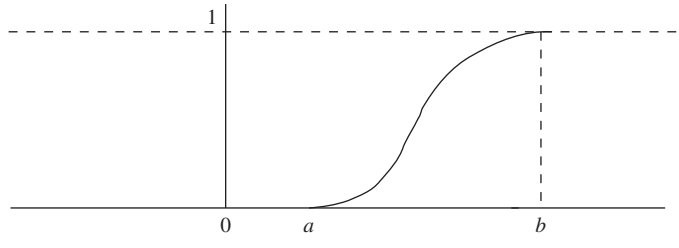
The definition above allows for a range of possible shapes for the distribution function of a random variable:

- It may increase continuously as x varies from $-\infty$ to ∞.
- It may not be defined on the whole of $\mathbb{R}$ but rather on some interval $[a, b]$, thus the probability begins accumulating at $x = a$ and is completely captured by $x = b$.
- It may exhibit periods where the probability remains constant.
- There may be points where the function suddenly jumps in value.

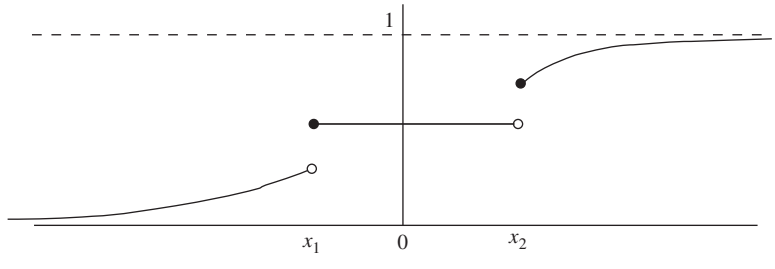
Examples of these possible shapes are displayed in Figure 3.2.



Distribution function – continuous and supported on real line



Distribution function – continuous and supported on [*a*,*b*]



Distribution function – supported on real line with discontinuities

Figure 3.2 Examples of distribution functions of a continuous random variable.

An extremely important use of a distribution function is that it can be used to recover the probabilities of many useful events. For example, if $a < b$ we can write

$$\begin{aligned} F(b) = \mathbb{P}[X \leq b] &= \mathbb{P}[X \leq a] + \mathbb{P}[X \in (a, b]] \\ &= F(a) + \mathbb{P}\left[X \in (a, b]\right], \end{aligned}$$

thus

$$\mathbb{P}\left[X \in (a, b]\right] = F(b) - F(a).$$

The distribution of a continuous random variable can often be described in terms of another function called the probability density function. We provide the following definition:

Definition 3.3. *Let X be a continuous random variable whose distribution F is a continuous function. If there exists a function $p : \mathbb{R} \to \mathbb{R}$ that satisfies*

$$p(x) \geq 0 \quad \text{and} \quad F(x) = \int_{-\infty}^{x} p(u)du \quad \text{for all } x \in \mathbb{R},$$

we say that p is the probability density function of X.

If F is continuously differentiable[1] then we can recover its probability density function via the fundamental theorem of calculus, which tells us that

$$p(x) = F'(x).$$

This indicates that we can characterize a continuous random variable through its probability density function. To strengthen this statement we have the following definition:

Definition 3.4. *A function $p : \mathbb{R} \to \mathbb{R}$ is the probability density function of some continuous random variable X if and only if*

$$p(x) \geq 0 \ \text{ for all } \ x \in \mathbb{R} \quad \text{and} \quad \int_{-\infty}^{\infty} p(u)du = 1.$$

We remark that the evaluation of p at some point $x \in \mathbb{R}$ is not a probability. In particular,

$$p(x) \neq \mathbb{P}\{X = x\}.$$

In order to compute probabilities from p we have to integrate it, for example

$$\mathbb{P}[X \in [a, b]] = F(b) - F(a) = \int_{a}^{b} p(x)dx.$$

Examples of probability density functions are given in Figure 3.3.

3.1.2 Expectation

If X is a discrete random variable with range $\{x_1, \ldots, x_n\}$ then its mathematical expectation or *mean* $\mathbb{E}[X]$ is just the probability–weighted average given by

$$\mathbb{E}[X] = \sum_{i=1}^{n} p_i x_i = \mu_X. \tag{3.3}$$

If X is a continuous random variable then its expectation is defined analogously; discrete probabilities are replaced by the density function of X and the summation is replaced by integration, i.e., we have

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} p(x)xdx = \mu_X. \tag{3.4}$$

[1] A calculus review can be found in the next Chapter.

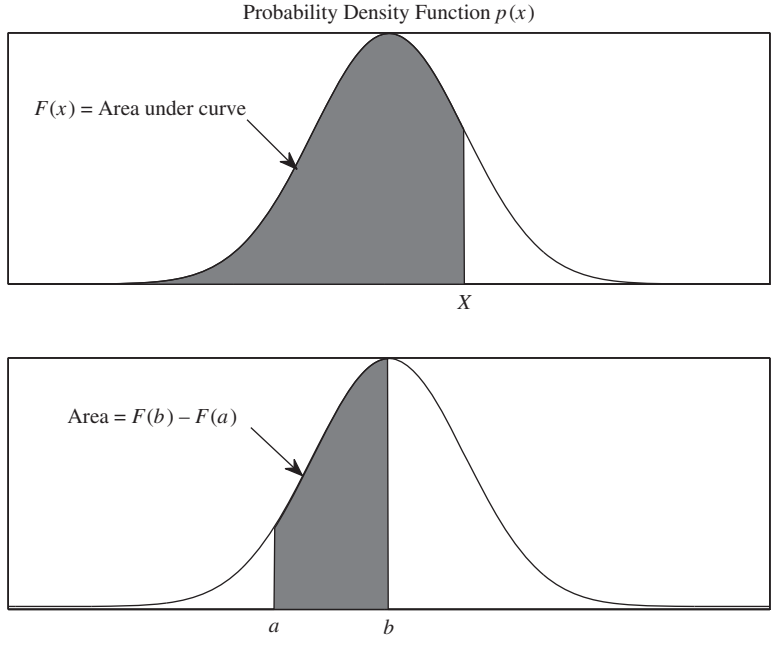


Figure 3.3 $\mathbb{P}[X \leq x] = \int_{-\infty}^{x} p(u)du$ (above) and $\mathbb{P}[X \in [a, b]] = \int_{a}^{b} p(u)du$ (below).

If the integral does not exist, neither does the expectation. In practice, this is rarely the case.

In general, we can view expectation as an operator that acts upon random variables. We list a few of its properties:

- If X and Y are random variables and $\alpha, \beta \in \mathbb{R}$ then

$$\mathbb{E}[\alpha X + \beta Y] = \alpha \mathbb{E}[X] + \beta \mathbb{E}[Y] \quad \text{we say expectation is linear.}$$

- Expectation only acts upon random quantities, thus if X is known then applying expectation has no effect, i.e., $\mathbb{E}[X] = X$.
- If $X \geq 0$ then $\mathbb{E}[X] \geq 0$; we say expectation is sign preserving.
- If X is a random variable and $f : \mathbb{R} \to R$ is a continuous function then the mathematical expectation of the new random variable $f(X)$ is defined by

$$\mathbb{E}[f(X)] = \begin{cases} \sum_{i=1}^{n} p_i f(x_i) & \text{if } X \text{ is discrete;} \\ \int_{-\infty}^{\infty} p(x) f(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

- If X and Y are random variables then the following bound holds:

$$\mathbb{E}\left[XY\right]^2 \leq \mathbb{E}\left[X^2\right] \mathbb{E}\left[Y^2\right]. \tag{3.5}$$

3.1.3 Variance

Given a random variable X its mean μ_X tells us the value that X is likely to hit on average. In addition to this information we also want to measure the extent to which X can deviate

from μ_X. In view of this we consider the quantity $(X - \mu_X)^2$; this is also a random variable and in particular, its value is large if X deviates greatly from μ_X, and much smaller when X is close to μ_X. The expected value of this random variable is thus a useful measure of how much X tends to vary from its expected value. This measure is called the variance of X.

Definition 3.5. *The variance of a random variable X with mean μ_X is defined by*

$$\text{var}(X) = \sigma_X^2 = \mathbb{E}\left[(X - \mu_X)^2\right].$$

The quantity σ_X, i.e., the square root of the variance, is called the standard deviation or volatility of X.

We remark that, by using the properties of the expectation operator, we can derive a useful alternative expression for the variance as follows:

$$\begin{aligned} \text{var}(X) &= \mathbb{E}\left[X^2 - 2\mu_X X + \mu_X^2\right] \\ &= \mathbb{E}\left[X^2\right] - 2\mu_X^2 + \mu_X^2 \\ &= \mathbb{E}\left[X^2\right] - \mu_X^2. \end{aligned} \tag{3.6}$$

3.2 MULTIVARIATE THEORY

Suppose we have a collection of n random variables $\{X_1, \ldots, X_n\}$ then, using tools from the previous chapter, we can capture this information in a random vector by setting

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

If we let $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$, then we write

$$\{\mathbf{X} \leq \mathbf{x}\} \text{ to denote the event } \{X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n\}.$$

Using this terminology we can now formulate the mathematical language we need in order to discuss the probabilistic behaviour of several random variables.

3.2.1 The joint distribution function

We define the n-dimensional joint distribution function of the random vector $\mathbf{X} = (X_1, \ldots, X_n)^T \in \mathbb{R}^n$ as

$$\begin{aligned} F(x_1, \ldots, x_n) = F(\mathbf{x}) &= \mathbb{P}[\mathbf{X} \leq \mathbf{x}] \\ &= \mathbb{P}[X_1 \leq x_1, \ldots, X_n \leq x_n]. \end{aligned}$$

3.2.2 The joint and marginal density functions

When the random variables are continuous and the joint distribution is differentiable in each of its variables we can define the n-dimensional joint probability density function by

$$p(\mathbf{x}) = p(x_1, \ldots, x_n) = \frac{\partial^n F(x_1, \ldots, x_n)}{\partial x_1 \cdots \partial x_n}.$$

Note, a review of differentiation of a multivariate function is given in the next chapter.

In this case we can connect the n-dimensional distribution and density functions of the random vector $\mathbf{X}$ by the identity

$$F(\mathbf{x}) = F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} p(x_1, \ldots, x_n) dx_n \ldots dx_1.$$

Suppose that we are given, or can compute, the n-dimensional distribution, F_{joint} say, of a random vector $\mathbf{X} = (X_1, \ldots, X_n)^T \in \mathbb{R}^n$. We write

$$F_{\text{joint}}(x_1, \ldots, x_n) = \mathbb{P}[X_1 \in (-\infty, x_1), \ldots, X_n \in (-\infty, x_n)].$$

Each random variable is real valued and so, trivially, we know that

$$\mathbb{P}[X_i \in \mathbb{R}] = 1 \quad \text{for } 1 \leq i \leq n.$$

In view of this we can write the distribution of any one of the single random variables, X_i say, as

$$\begin{aligned} F_i(x) &= \mathbb{P}[X_1, \ldots, X_{i-1} \in \mathbb{R}, X_i \leq x, X_{i+1}, \ldots, X_n \in \mathbb{R}] \\ &= \text{limit of } F_{\text{joint}}(x_1, \ldots, x_n) \text{ as } x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \to \infty. \end{aligned}$$

If the joint density function, p_{joint} say, exists then the individual or marginal probability density function of any one of the single random variables X_i can be obtained by fully integrating out all other contributions, that is

$$\underbrace{\underbrace{\int_{\mathbb{R}} \cdots \int_{\mathbb{R}}}_{(n-1)\text{times}} p_{\text{joint}}(x_1, \ldots x_{i-1}, x, x_{i+1}, \ldots x_n) dx_n \ldots dx_{i+1} dx_{i-1} \ldots dx_1}_{= p_i(x) \text{ the marginal density of the random variable } X_i}.$$

3.2.3 The notion of independence

Suppose we have two random variables X_1 and X_2 with density functions p_1 and p_2 respectively. If we now bring these variables together we can compute the joint density function p_{joint} of the 2-dimensional vector $(X_1, X_2)^T$. The joint density is designed to capture the potential effect that the outcome of one random variable can have on the shape of the density function of the other. If there is no effect then we say that the random variables X_1

and X_2 are independent. Mathematically, this happens when the joint distribution is simply the product of the two marginals, i.e., when

$$p_{\text{joint}}(x_1, x_2) = p_1(x_1)p_2(x_2) \tag{3.7}$$

and, equivalently, when their joint distribution F_{joint} is the product of the individual distributions,

$$\begin{aligned} F_{\text{joint}}(x_1, x_2) &= \mathbb{P}[X_1 \leq x_1, X_2 \leq x_2] \\ &= \mathbb{P}[X_1 \leq x_1]\mathbb{P}[X_2 \leq x_2] = F_1(x_1)F_2(x_2). \end{aligned}$$

This notion extends to a set of many random variables, i.e., we say that the random variables $X_1, \ldots, X_n$ are mutually independent if their joint density function factorizes as

$$p_{\text{joint}}(x_1, x_2, \ldots, x_n) = p_1(x_1)p_2(x_2)\cdots p_n(x_n). \tag{3.8}$$

3.2.4 The notion of conditional dependence

We now examine the general case where the outcome of one random variable, X_2 say, changes the probability structure of X_1. In this case we say that X_1 is conditionally dependent upon the outcome of X_2 and we capture this via the so-called conditional density function defined by

$$p_1^{(\text{cond})}(x_1|X_2 = x_2) = \frac{p_{\text{joint}}(x_1, x_2)}{p_2(x_2)}.$$

Thus, with no other probabilistic information, the density of X_1 is given by p_1. However, the outcome of X_1 will typically depend upon many external influences. In particular, if the random variable X_2 exerts an influence on X_1 then the density function p_1 can be refined to $p_1^{(\text{cond})}$, the conditional density, which captures the probabilistic structure of X_1 based on the outcome of X_2.

We note that if the random variables are independent then this conditional refinement doesn't change anything since, using (3.7), we have

$$p_1^{(\text{cond})}(x_1|X_2 = x_2) = \frac{p_{\text{joint}}(x_1, x_2)}{p_2(x_2)} = \frac{p_1(x_1)p_2(x_2)}{p_2(x_2)} = p_1(x_1).$$

3.2.5 Covariance and correlation

Let X_1 and X_2 denote a pair of continuous random variables whose joint density is given by p_{joint}. We can recover the marginal densities of both X_1 and X_2 via the identities

$$p_1(x_1) = \int_{\mathbb{R}} p_{\text{joint}}(x_1, u)du \quad \text{and} \quad p_2(x_2) = \int_{\mathbb{R}} p_{\text{joint}}(u, x_2)du.$$

We can apply (3.4) to compute the expected values of these variables as

$$\mu_1 = \int_{\mathbb{R}} \int_{\mathbb{R}} x_1 p_{\text{joint}}(x_1, u) du dx_1$$
$$\text{and} \quad \mu_2 = \int_{\mathbb{R}} \int_{\mathbb{R}} x_2 p_{\text{joint}}(u, x_2) du dx_2.$$

Similarly, we can compute the variance of X_1 and X_2 as

$$\sigma_1^2 = \int_{\mathbb{R}} \int_{\mathbb{R}} (x_1 - \mu_1)^2 p_{\text{joint}}(x_1, u) du dx_1$$
$$\text{and} \quad \sigma_2^2 = \int_{\mathbb{R}} \int_{\mathbb{R}} (x_2 - \mu_2)^2 p_{\text{joint}}(u, x_2) du dx_2.$$

These calculations show how the joint density can be used to recover the familiar mean and variance of an individual random variable. We are also interested in measuring how the pair of variables interact with each other. For this purpose we consider the random variable $(X_1 - \mu_1)(X_2 - \mu_2)$ and we make the following simple observations:

- $(X_1 - \mu_1)(X_2 - \mu_2) > 0$ tends to indicate that X_1 and X_2 move in harmony, i.e., both move above or below their respective means.
- $(X_1 - \mu_1)(X_2 - \mu_2) < 0$ tends to indicate that X_1 and X_2 oppose each other, i.e., they move in opposite directions relative to their means.

We define the covariance between X_1 and X_2 to be the expected value of $(X_1 - \mu_1)(X_2 - \mu_2)$, i.e., we set

$$\begin{aligned} \sigma_{12} &= \text{cov}(X_1, X_2) \\ &= \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= \int_{\mathbb{R}} \int_{\mathbb{R}} (x_1 - \mu_1)(x_2 - \mu_2) p_{\text{joint}}(x_1, x_2) dx_1 dx_2. \end{aligned} \tag{3.9}$$

The covariance measure is highly useful in applications and we collect together some of its properties.

- Using the properties of the expectation operator we can derive the following alternative expression:

$$\begin{aligned} \sigma_{12} = \text{cov}(X_1, X_2) &= \mathbb{E}\left[X_1 X_2 - X_1 \mu_2 - X_2 \mu_1 + \mu_1 \mu_2\right] \\ &= \mathbb{E}\left[X_1 X_2\right] - \mu_2 \mathbb{E}[X_1] - \mu_1 \mathbb{E}[X_2] + \mu_1 \mu_2 \\ &= \mathbb{E}\left[X_1 X_2\right] - \mu_1 \mu_2. \end{aligned} \tag{3.10}$$

- Covariance is symmetric, i.e., $\sigma_{12} = \text{cov}(X_1, X_2) = \text{cov}(X_2, X_1) = \sigma_{21}$.
- The covariance of a random variable with itself is its variance, i.e., $\text{cov}(X, X) = \mathbb{E}[(X - \mu)^2]$.
- If X_1 and X_2 are independent then they have zero covariance. To demonstrate this, consider the following:

$$\begin{aligned}\operatorname{cov}(X_1, X_2) &= \int_{\mathbb{R}}\int_{\mathbb{R}} (x_1-\mu_1)(x_2-\mu_2)\, p_{\text{joint}}(x_1, x_2)\, dx_1 dx_2 \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} (x_1-\mu_1)(x_2-\mu_2)\, \underbrace{p_1(x_1)p_2(x_2)}_{\text{by independence}}\, dx_1 dx_2 \\ &= \int_{\mathbb{R}}\int_{\mathbb{R}} x_1 x_2 p_1(x_1) p_2(x_2) dx_1 dx_2 - \mu_1\mu_2 \quad \text{by (3.10)} \\ &= \underbrace{\int_{\mathbb{R}} x_1 p_1(x_1) dx_1}_{=\mu_1} \underbrace{\int_{\mathbb{R}} x_2 p_2(x_2) dx_2}_{=\mu_2} - \mu_1\mu_2 = 0.\end{aligned}$$

- The above development establishes that if X_1 and X_2 are independent then $\mathbb{E}[X_1X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$. In fact, one can use precisely the same argument to show that, more generally, if X_1 and X_2 are independent then

$$\mathbb{E}[f_1(X_1)f_2(X_2)] = \mathbb{E}[f_1(X_1)]\mathbb{E}[f_2(X_2)] \tag{3.11}$$

  where $f_1 = f_1(x_1)$ and $f_2 = f_2(x_2)$ are any two functions for which $\mathbb{E}[f_2(X_2)]$ and $\mathbb{E}[f_1(X_1)]$ are well defined.
- The magnitude of the covariance between random variables X_1 and X_2 is bounded above by the product of their standard deviations, that is

$$|\sigma_{12}| \leq \sigma_1\sigma_2. \tag{3.12}$$

  This follows directly from (3.5) by setting $X = X_1 - \mu_1$ and $Y = X_2 - \mu_2$.
- In view of the covariance bound (3.12) it is useful to de-scale the covariance measure by dividing through by $\sigma_1\sigma_2$. The result, defined by

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1\sigma_2}, \tag{3.13}$$

  is called the correlation coefficient between X_1 and X_2 and has the useful property that its value is confined to the interval $[-1, 1]$.
- We say that two random variables are uncorrelated if their correlation coefficient (and hence their covariance) is zero. We can conclude that independent random variables are uncorrelated, however the converse is not true; an uncorrelated pair of random variables are not necessarily independent.

3.2.6 The mean vector and covariance matrix

Suppose we have an n-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$. We define its mean vector by

$$\mathbf{e} = \mathbb{E}\left[\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}\right] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix} \in \mathbb{R}^n.$$

We can display the covariance information of these random variables in an $n \times n$ covariance matrix defined by

$$\mathbf{V} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}, \tag{3.14}$$

or, equivalently, we have $\mathbf{V} \in \mathbb{R}^{n \times n}$ such that

$$(\mathbf{V})_{kl} = \sigma_{kl} = \mathbb{E}[(X_k - \mu_k)(X_l - \mu_l)] \quad (1 \leq k, l \leq n). \tag{3.15}$$

Appealing to our linear algebra toolbox we observe that the outer product of the mean-adjusted random vector $\mathbf{X} - \mathbf{e}$ with itself is given by $(\mathbf{X} - \mathbf{e})(\mathbf{X} - \mathbf{e})^T \in \mathbb{R}^{n \times n}$, such that

$$[(\mathbf{X} - \mathbf{e})(\mathbf{X} - \mathbf{e})^T]_{kl} = (X_k - \mu_k)(X_l - \mu_l) \quad (1 \leq k, l \leq n).$$

Comparing this to (3.15) we notice that the covariance matrix $\mathbf{V}$ can be neatly expressed as the expectation of this outer product, i.e.,

$$\mathbf{V} = \mathbb{E}[(\mathbf{X} - \mathbf{e})(\mathbf{X} - \mathbf{e})^T]. \tag{3.16}$$

3.2.7 Linear combinations of random variables

Suppose we have a random vector $\mathbf{X} = (X_1, \dots, X_n)^T \in \mathbb{R}^n$; we can build a new random variable by taking a linear combination of the elements that make up $\mathbf{X}$. Specifically, we let $\mathbf{w} = (w_1, \dots w_n)^T \in \mathbb{R}^n$ and define

$$X = \sum_{j=1}^{n} w_j X_j = \mathbf{w}^T \mathbf{X}.$$

The expectation of the new random variable $\mathbf{X}$ is given by

$$\mu = \mathbb{E}[X] = \mathbb{E}[\mathbf{w}^T \mathbf{X}] = \mathbf{w}^T \mathbb{E}[\mathbf{X}] = \mathbf{w}^T \mathbf{e} = \sum_{j=1}^{n} w_j \mu_j. \tag{3.17}$$

Furthermore, its variance is given by

$$\begin{aligned} 0 \leq \sigma^2 = \mathbb{E}[(X - \mu)^2] &= \mathbb{E}\Big[\Big(\mathbf{w}^T (\mathbf{X} - \mathbf{e})\Big)\Big(\mathbf{w}^T (\mathbf{X} - \mathbf{e})\Big)^T\Big] \\ &= \mathbb{E}[\mathbf{w}^T (\mathbf{X} - \mathbf{e})(\mathbf{X} - \mathbf{e})^T \mathbf{w}] \\ &= \mathbf{w}^T \mathbb{E}[(\mathbf{X} - \mathbf{e})(\mathbf{X} - \mathbf{e})^T] \mathbf{w} \\ &= \mathbf{w}^T \mathbf{V} \mathbf{w}. \end{aligned} \tag{3.18}$$

The above analysis reveals that the variance of a linear combination of random variables is given by the quadratic form based on the non-negative definite covariance matrix $\mathbf{V}$, i.e.,

$$\sigma^2 = \text{var}(\sum_{j=1}^{n} w_j X_j) = \text{var}(\mathbf{w}^T \mathbf{X}) = \mathbf{w}^T \mathbf{V} \mathbf{w} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij}.$$

The simplest possible case of a linear combination occurs when we have just two independent random variable X and Y and we consider their sum $Z = X + Y$. If we let F_X and F_Y denote the respective distribution functions of X and Y, then it can be shown, see Section 15.12 in Cramér (1966), that F_Z, the distribution of their sum, is given by

$$F_Z(z) = \int_{\mathbb{R}} F_X(z-u) dF_Y(u) = \int_{\mathbb{R}} F_Y(z-u) dF_X(u). \tag{3.19}$$

3.3 THE NORMAL DISTRIBUTION

In order to make progress in probability theory it is useful to become acquainted with some of the more popular families of distribution/density functions. These families are fixed by specifying a closed-form representation for either the probability mass function in the discrete case, or the probability density function in the continuous case. We will encounter many different families throughout this book, however perhaps the most important is the normal distribution for a continuous random variable.

Definition 3.6. *A continuous random variable X with mean μ and variance σ^2 is said to be normally distributed (written as $X \sim N(\mu, \sigma^2)$) if its probability density function is given by*

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \tag{3.20}$$

where $\exp(\cdot)$ *denotes the exponential function defined by*

$$\exp(x) := \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} \cdots, \tag{3.21}$$

for any $x \in \mathbb{R}$.

We note that the distribution of a normal random variable X is completely defined by its mean and its variance and is given by

$$F(x) = \mathbb{P}(X \leq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right) du.$$

Unfortunately, the above integral has no closed-form solution; numerical methods are required for its evaluation. In order to avoid this we can make a simple change of variable and set

$$Z = \frac{X-\mu}{\sigma}.$$

This shifted and scaled random variable has zero mean and unit variance; we say that it has the standard normal distribution and we write $Z \sim N(0, 1)$. In this standardized case we reserve a special notation; the standard normal density function is given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right),$$

and the standard normal distribution function is

$$\Phi(z) = \mathbb{P}[Z \leq z] = \int_{-\infty}^{z} \varphi(u)du.$$

Fortunately, there exist statistical tables which enable the user to evaluate the function $\Phi(z) \in [0, 1]$ for a given z and also to evaluate the inverse function $\Phi^{-1}(\alpha)$ for a given $\alpha \in [0, 1]$.

The standard normal distribution and its tabulated values can be used to compute the distribution of any normally distributed random variable, this follows since

$$\begin{aligned} F(x) = \mathbb{P}[X \leq x] &= \mathbb{P}\left[Z\sigma + \mu \leq x\right] \quad \left(\text{where } Z = \frac{X-\mu}{\sigma} \sim N(0,1)\right) \\ &= \mathbb{P}\left[Z \leq \frac{x-\mu}{\sigma}\right] = \Phi\left(\frac{x-\mu}{\sigma}\right). \end{aligned}$$

The Multivariate Normal Distribution

We have described how a normally distributed random variable is characterized by the fact that its density function is given by (3.20). We note that this can equivalently be written as

$$p(x) = (2\pi)^{-\frac{1}{2}}(\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)(\sigma^2)^{-1}(x-\mu)\right). \tag{3.22}$$

We can extend this notion to higher dimensions, where we talk of a random vector having a multivariate normal distribution. This concept is captured in the following definition:

Definition 3.7. *An n-dimensional random vector of $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is said to have the multivariate normal distribution if, given any non-zero coefficient vector $\boldsymbol{w} = (w_1, \ldots, w_n)^T \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$, the resulting linear combination*

$$X = \mathbf{w}^T \boldsymbol{X} = w_1 X_1 + \cdots + w_n X_n$$

is a normally distributed random variable.

Let $\mathbf{X}$ denote an n-dimensional random vector whose corresponding mean vector and covariance matrix are given by

$$\mathbf{e} = \mathbb{E}[\mathbf{X}] \quad \text{and} \quad \mathbf{V} = \mathbb{E}[(\mathbf{X}-\mathbf{e})(\mathbf{X}-\mathbf{e})^T].$$

If $\mathbf{X}$ has the multivariate normal distribution then, using (3.17) and (3.18), we can deduce that

$$X = \mathbf{w}^T\mathbf{X} \sim N(\mathbf{w}^T\mathbf{e}, \mathbf{w}^T\mathbf{V}\mathbf{w}).$$

In order to summarize this information we will frequently write $\mathbf{X} \sim N(\mathbf{e}, \mathbf{V})$ to indicate that the random vector $\mathbf{X}$ has the multivariate distribution.

If $\mathbf{X} \sim N(\mathbf{e}, \mathbf{V})$ and its covariance matrix $\mathbf{V}$ is positive definite then the probability density function for $\mathbf{X}$ is defined by

$$p(\mathbf{x}) = (2\pi)^{-\frac{n}{2}}(\det(\mathbf{V}))^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{e})^T\mathbf{V}^{-1}(\mathbf{x}-\mathbf{e})\right). \tag{3.23}$$

We notice how this expression acts as the higher-dimensional analogue of the single-variable case; the univariate quadratic expression, based on the inverse of σ^2, in the exponential (3.22) is replaced by a quadratic form based on $\mathbf{V}^{-1}$, the inverse of the covariance matrix.

Essential Mathematics for Market Risk Management
by Simon Hubbert

4 Optimization Tools

Optimization theory is the branch of mathematics where methods are developed for delivering the best possible solution to some fixed mathematical problem. A financial risk manager faces such problems on a day-to-day basis, e.g., calculating the minimum risk associated with a portfolio of financial assets or determining the best way to interpret financial data. In this chapter we introduce the essential background and methodology from this area which will enable a practitioner to tackle these problems successfully.

4.1 BACKGROUND CALCULUS

Calculus is the study of how functions change and, as such, is an essential tool for solving problems in a wide range of applications in science, engineering and finance. It is assumed that the reader will have some familiarity with calculus however, for convenience, we use this section to briefly review some fundamentals; the presentation here is inspired and influenced by the excellent numerical analysis lecture notes of B.J.C. Baxter (2010).

4.1.1 Single-variable functions

We kick-start our review in the familiar univariate setting, where we think of function $f : U \to \mathbb{R}$ as a rule which, for any real number $x \in U$, determines a unique real number denoted by $f(x)$. We say that the set U is the domain of the function and in most cases we can think of U as the whole real line or a simple interval. A function is said to be continuous if it has no jumps, i.e., we can draw its graph without having to lift our pencil from the paper. A more mathematical definition is to say that f is continuous at a point $a \in U$ if

$$\lim_{h \to 0} f(a + h) = f(a),$$

i.e., as the variable $x = a + h$ approaches a (from the left or the right) the function approaches its value $f(a)$. If f is continuous at every $a \in U$ we say that f is continuous (on U).

We should be aware from our high-school mathematics that the derivative of f at some particular point $a \in U$ represents the gradient of the tangent line that touches the graph of f at a. More formally, this is defined as the limiting value of

$$\frac{f(a + h) - f(a)}{h} \quad \text{as} \quad h \to 0,$$

which, assuming this limit exists, we write as

$$\frac{df}{dx}(a) = f'(a) = \lim_{h \to 0} \frac{f(a + h) - f(a)}{h}. \tag{4.1}$$

If f has a derivative at every point in U then we say that f is a differentiable function (on U). It may be possible to continue taking derivatives, in which case we speak of the degree of differentiability. Specifically, we say that f is m times differentiable if it is differentiable, and, in addition, there exist functions $f^{(2)}, \ldots, f^{(m)}$ such that

$$f^{(k)}(a) = \lim_{h \to 0} \frac{f^{(k-1)}(a+h) - f^{(k-1)}(a)}{h} \quad \text{for} \quad 2 \leq k \leq m \quad \text{and all} \quad a \in U.$$

We remark that the *primed* notation is commonly used to denote low order derivatives, i.e., we often write f' for $f^{(1)}$, f'' for $f^{(2)}$, and so on. We say that a function $f : \mathbb{R} \to \mathbb{R}$ is smooth if it is infinitely differentiable. A remarkable result from mathematical analysis states that almost every smooth function we encounter in practice can be expanded in the form of a Taylor series:

$$f(a+h) = \sum_{m=0}^{\infty} \frac{h^m}{m!} f^{(m)}(a), \quad h \in \mathbb{R} \tag{4.2}$$

or, alternatively, it is often useful to express this as

$$f(a+h) = f(a) + f'(a)h + \frac{1}{2} f''(a)h^2 + O(h^3), \tag{4.3}$$

where $O(h^3)$ denotes a term that is dominated by a constant multiple of h^3. In particular, if we are examining $f(x)$ where x is close to a (i.e., when $h = x - a$ is small) then we can effectively ignore the higher-order term and employ the quadratic approximation

$$p(x) = f(a) + f'(a)(x-a) + \frac{1}{2} f''(a)(x-a)^2. \tag{4.4}$$

Indeed, if x is very close to a then we may only need a linear approximation

$$l(x) = f(a) + f'(a)(x-a). \tag{4.5}$$

4.1.2 Multivariable functions

We now suppose that f is a function that is defined on $\mathbb{R}^n$ where $n > 1$. For instance, f may denote the value of a portfolio depending upon the n-dimensional vector of financial variables $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$. In analogy to the univariate case, we say that f is continuous at $\mathbf{a} \in \mathbb{R}^n$ if

$$\lim_{\mathbf{h} \to \mathbf{0}} f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}),$$

i.e., as the variable $\mathbf{x} = \mathbf{a} + \mathbf{h}$ approaches $\mathbf{a}$ (from any direction) the function approaches its value $f(\mathbf{a})$.

If we are given a function $f(x_1, \ldots, x_n)$ that is continuous on the whole of $\mathbb{R}^n$ then we could choose to hold all but one of these variables fixed and, as a result, obtain a univariate function which we can (potentially) differentiate. Specifically, if we let $k \in \{1, 2, \ldots, n\}$ we

say that the kth partial derivative of f is the univariate derivative of f with respect to x_k only, while the remaining $n-1$ variables are all held fixed. Mathematically, we say that

$$\frac{\partial f}{\partial x_k}(\mathbf{a})$$

is the kth partial derivative of f at the point $\mathbf{a} = (a_1, \ldots, a_n)^T \in \mathbb{R}^n$ and its value is given by

$$\frac{\partial f}{\partial x_k}(\mathbf{a}) = \lim_{h \to 0} \left[\frac{f(a_1, \ldots, a_{k-1}, a_k + h, a_{k+1}, \ldots, a_n) - f(a_1, \ldots, a_n)}{h} \right].$$

If the above limit is well defined for every k and for all $\mathbf{a} \in \mathbb{R}^n$, we say that f is continuously differentiable (on $\mathbb{R}^n$). We store the values of these n partial derivatives (evaluated at some $\mathbf{a} \in \mathbb{R}^n$) in the gradient vector defined by

$$\nabla f(\mathbf{a}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{a}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{a}) \end{pmatrix} \in \mathbb{R}^n. \tag{4.6}$$

Now, since each partial derivative of f is itself a function of n variables then, as with the univariate setting, we can continue to differentiate and thereby generate the higher-order mixed derivatives of f. For our purposes we need only expose ourselves to the second-order derivatives, defined by

$$\frac{\partial^2 f}{\partial x_l \partial x_k} = \frac{\partial}{\partial x_l} \left(\frac{\partial f}{\partial x_k} \right) \quad \text{where} \quad (1 \leq k, l \leq n).$$

We note that there are n^2 second-order derivatives of f and, for convenience, we store the values they take at some point $\mathbf{a} \in \mathbb{R}^n$ in the $n \times n$ second-derivative matrix $D^2 f(\mathbf{a}) \in \mathbb{R}^{n \times n}$, defined by

$$\left(D^2 f(\mathbf{a}) \right)_{kl} = \frac{\partial^2 f}{\partial x_l \partial x_k}(\mathbf{a}) \quad \text{for} \quad 1 \leq k, l \leq n. \tag{4.7}$$

We shall always assume that all of the mixed second-order derivatives are continuous functions on $\mathbb{R}^n$, in which case it can be shown that

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_l \partial x_k} = \frac{\partial^2 f(\mathbf{x})}{\partial x_k \partial x_l} \quad \text{for} \quad 1 \leq k, l \leq n,$$

and so the second-derivative matrix $D^2 f$ is always symmetric.

In this higher-dimensional setting it turns out that there is a rather nice analogue of the Taylor expansion (4.3), namely:

$$f(\mathbf{a} + \mathbf{h}) = f(\mathbf{a}) + \nabla f(\mathbf{a})^T \mathbf{h} + \frac{1}{2} \mathbf{h}^T D^2 f(\mathbf{a}) \mathbf{h} + O(\|\mathbf{h}\|^3). \tag{4.8}$$

A rigorous proof of the higher-dimensional Taylor formula (4.8) is beyond the scope of this book, however, the following sketch should serve as a convincing justification. First of all we recall the univariate setting and let $D \equiv \frac{d}{dx}$ so that $D^m f(x) = f^{(m)}(x)$. Using this notation we can rewrite (4.2) as

$$f(a+h) = \sum_{m=0}^{\infty} \frac{(hD)^m}{m!} f(a).$$

This reminds us of the Taylor expansion (about zero) of the exponential function (3.21) and leads us to write the Taylor expansion as $f(a+h) = \exp(hD) f(a)$. If we extend this representation to a function $f(\mathbf{x})$ of n variables, then we find that

$$\begin{aligned} f(\mathbf{a}+\mathbf{h}) &= \exp(h_1\partial_1)\cdots\exp(h_n\partial_n) f(\mathbf{a}) \\ &= \exp(h_1\partial_1 + \cdots + h_n\partial_n) f(\mathbf{a}), \end{aligned}$$

where $\partial_k = \frac{\partial}{\partial x_k}$ for $k = 1, \ldots, n$. Now, using (3.21), we find

$$\begin{aligned} f(\mathbf{a}+\mathbf{h}) &= \exp(h_1\partial_1 + \cdots + h_n\partial_n) f(\mathbf{a}) \\ &= \sum_{m=0}^{\infty} \frac{(h_1\partial_1 + \cdots + h_n\partial_n)^m}{m!} f(\mathbf{a}) \\ &= f(\mathbf{a}) + \underbrace{\left(h_1 \frac{\partial f}{\partial x_1}(\mathbf{a}) + \cdots + h_n \frac{\partial f}{\partial x_n}(\mathbf{a}) \right)}_{\nabla f(\mathbf{a})^T \mathbf{h}} \\ &\quad + \frac{1}{2} (h_1\partial_1 + \cdots + h_n\partial_n)^2 f(\mathbf{a}) + \cdots \end{aligned}$$

Now, the squared term above can be written as

$$\begin{aligned} (h_1\partial_1 + \cdots + h_n\partial_n)^2 f(\mathbf{a}) &= \sum_{j=1}^{n} h_j\partial_j \sum_{k=1}^{n} h_k\partial_k f(\mathbf{a}) \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} h_j h_k \partial_j\partial_k f(\mathbf{a}) \\ &= \sum_{j=1}^{n}\sum_{k=1}^{n} h_j h_k \frac{\partial^2 f}{\partial_j\partial_k}(\mathbf{a}) \\ &= \mathbf{h}^T D^2 f(\mathbf{a}) \mathbf{h}. \end{aligned}$$

4.2 OPTIMIZING FUNCTIONS

An important application of multivariable calculus occurs when our aim is to minimize a function of n variables. We shall say that the function f has a *local minimum* at $\mathbf{x} = \mathbf{a}$ if

$$f(\mathbf{a} + h\mathbf{u}) \geq f(\mathbf{a}) \quad \text{for every unit vector } \mathbf{u}, \tag{4.9}$$

when $h > 0$ is sufficiently small.

Before we continue our investigation, the following remarks are in order:

- If (4.9) holds for all values of $h > 0$ then we say that $\mathbf{x} = \mathbf{a}$ is the *global minimum* of f.
- If we flip the $\leq$ sign in (4.9) to $\geq$ then we have the definition of a *local maximum* of f.
- If we strengthen the inequality signs $\leq$ ($\geq$) in (4.9) to $<$ ($>$) then we say that the resulting point $\mathbf{x} = \mathbf{a}$ is a *strict local minimum (maximum)*.

We kick-start our investigation in the univariate setting ($n = 1$) where we recall (from our high-school mathematics) that a function $f(x)$ has a local minimum at $x = a$ if $f'(a) = 0$ and $f''(a) > 0$. We now consider the analogous conditions for $n > 1$.

Proposition 4.1. *If $\nabla f(\boldsymbol{a}) \neq 0$ then $f(\boldsymbol{x})$ does not have a local minimum (or maximum) at $\boldsymbol{x} = \boldsymbol{a}$.*

Proof. Consider the multivariate linear approximation

$$l(\mathbf{x}) = f(\mathbf{a}) + \mathbf{g}^T(\mathbf{x} - \mathbf{a})$$

to $f(\mathbf{x})$, where $\mathbf{g} = \nabla f(\mathbf{a})$. For any $h > 0$ and any unit vector $\mathbf{u}$ we have

$$l(\mathbf{a} + h\mathbf{u}) = f(\mathbf{a}) + h\mathbf{g}^T\mathbf{u}.$$

In particular, we can set $\mathbf{u} = -\mathbf{g}/\|\mathbf{g}\|$ to obtain

$$l(\mathbf{a} + h\mathbf{u}) = f(\mathbf{a}) - \|\mathbf{g}\|h < f(\mathbf{a}).$$

Now, since $f(\mathbf{a} + h\mathbf{u}) = l(\mathbf{a} + h\mathbf{u}) + O(h^2)$ we can deduce that

$$f(\mathbf{a} + h\mathbf{u}) < f(\mathbf{a})$$

for all sufficiently small $h > 0$. Thus, f does not possess a local minimum at $\mathbf{x} = \mathbf{a}$. If we choose $\mathbf{u} = \mathbf{g}/\|\mathbf{g}\|$ then the same argument shows that there is no local maximum at $\mathbf{x} = \mathbf{a}$. □

In view of the above result we can conclude that $\nabla f(\mathbf{a}) = 0$ is a necessary condition for $f(\mathbf{x})$ to possess a local minimum (or maximum) at $\mathbf{x} = \mathbf{a}$. It turns out that, as in the univariate case, a necessary condition for a local minimum involves the second-derivative information.

Proposition 4.2. *If $\nabla f(\boldsymbol{a}) = 0$ and the second-derivative matrix $D^2 f(\boldsymbol{a})$ is positive definite then $f(\boldsymbol{x})$ has a strict local minimum (or maximum) at $\boldsymbol{x} = \boldsymbol{a}$.*

Proof. Consider the multivariate quadratic approximation

$$p(\mathbf{x}) = f(\mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T D^2 f(\mathbf{a})(\mathbf{x} - \mathbf{a}).$$

Then, for any $h > 0$ and any unit vector $\mathbf{u}$, we obtain

$$p(\mathbf{a} + h\mathbf{u}) = f(\mathbf{a}) + \frac{h^2}{2} \underbrace{\mathbf{u}^T D^2 f(\mathbf{a})\mathbf{u}}_{>0} > f(\mathbf{a}),$$

and thus, arguing as before, we deduce that

$$f(\mathbf{a} + h\mathbf{u}) > f(\mathbf{a}),$$

for all sufficiently small $h > 0$. □

4.2.1 Unconstrained quadratic functions

In order to illustrate the theory we have developed, we turn our attention to the problem of minimizing a quadratic form. We should be familiar with the simple one-dimensional case where the functional form is given by

$$q(x) = \frac{1}{2}ax^2 + bx + c, \qquad \text{where } a, b, c \in \mathbb{R} \text{ and } a \neq 0.$$

It is often useful to view this function in an alternative form by performing a simple algebraic operation known as completing the square. Specifically, it is easy to check that

$$q(x) = \frac{1}{2}a\left(x + \frac{b}{a}\right)^2 + c - \frac{b^2}{2a}.$$

The turning point of q is determined by solving

$$q'(x^*) = a\left(x^* + \frac{b}{a}\right) = 0 \quad \Rightarrow \quad x^* = -b/a. \tag{4.10}$$

Furthermore, the turning point is deemed to be a maximum or a minimum by examining the sign of its curvature, i.e.,

$$\text{if } q''(x^*) = a > 0 \text{ then } x^* \text{ is a minimum;}$$

$$\text{if } q''(x^*) = a < 0 \text{ then } x^* \text{ is a maximum.}$$

In higher dimensions the analogue of the quadratic function is given by

$$q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c, \tag{4.11}$$

where $\mathbf{A} \in \mathbb{R}^{n\times n}$ is invertible, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. We note that for any square matrix $\mathbf{A}$, we can write

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T \underbrace{\left(\frac{\mathbf{A}+\mathbf{A}^T}{2}\right)}_{\text{symmetric matrix}} \mathbf{x}$$

and thus, for this reason, we consider the matrix $\mathbf{A}$ to be symmetric.

The notion of completing the square extends to this higher-dimensional setting, where the aim is to find a vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ and a constant $\beta \in \mathbb{R}$ such that

$$q(\mathbf{x}) = \frac{1}{2}(\mathbf{x}+\boldsymbol{\alpha})^T\mathbf{A}(\mathbf{x}+\boldsymbol{\alpha}) + \beta.$$

Expanding this we find that

$$q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \boldsymbol{\alpha}^T\mathbf{A}\mathbf{x} + \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{A}\boldsymbol{\alpha} + \beta.$$

Now matching the term in $\mathbf{x}$ above with the corresponding term in (4.11), we find that

$$\mathbf{A}\boldsymbol{\alpha} = \mathbf{b} \Rightarrow \boldsymbol{\alpha} = \mathbf{A}^{-1}\mathbf{b} \quad \text{(given that } \mathbf{A} \text{ is invertible)}. \tag{4.12}$$

Now, in the same fashion, we can match the constant term to find that

$$\begin{aligned}\beta &= c - \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{A}\boldsymbol{\alpha}\\ &= c - \left(\mathbf{A}^{-1}\mathbf{b}\right)^T\mathbf{A}\mathbf{A}^{-1}\mathbf{b}\\ &= c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}\end{aligned}$$

and so, in conclusion, (4.11) can be written as

$$q(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x}+\mathbf{A}^{-1}\mathbf{b}\right)^T\mathbf{A}\left(\mathbf{x}+\mathbf{A}^{-1}\mathbf{b}\right) + c - \mathbf{b}^T\mathbf{A}^{-1}\mathbf{b}. \tag{4.13}$$

Note the analogy with the univariate case (4.10), which can be written as

$$q(x) = \frac{1}{2}\left(x + a^{-1}b\right)a\left(x + a^{-1}b\right) + c - \frac{1}{2}ba^{-1}b.$$

Differentiation of the n-dimensional quadratic form is analogous to the univariate case too; specifically, we have

$$\begin{aligned}\nabla q(\mathbf{x}) = \mathbf{A}\left(\mathbf{x}+\mathbf{A}^{-1}\mathbf{b}\right) = \mathbf{A}\mathbf{x} + \mathbf{b} \quad &\text{compare with } q'(x) = ax + b;\\ \text{and } \nabla^2 q(\mathbf{x}) = \mathbf{A} \quad &\text{compare with } q''(x) = a.\end{aligned}$$

The process of finding the minimum (or maximum) of the quadratic form is also analogous to the univariate case; using Propositions 4.1 and 4.2, we set the gradient to zero and then examine the second-derivative information to determine the nature of the turning point. Specifically, we have

$$\nabla q(\mathbf{x}^*) = \mathbf{A}\left(\mathbf{x}^* + \mathbf{A}^{-1}\mathbf{b}\right) = \mathbf{0} \Rightarrow \mathbf{x}^* = -\mathbf{A}^{-1}\mathbf{b}.$$

In addition, $\mathbf{x}^*$ is deemed to be a minimum or a maximum based upon the following criteria:

$$\begin{aligned} &\text{if } \nabla^2 q(\mathbf{x}^*) = \mathbf{A} \text{ is positive definite then } \mathbf{x}^* \text{ is a minimum;} \\ &\text{if } -\nabla^2 q(\mathbf{x}^*) = -\mathbf{A} \text{ is positive definite then } \mathbf{x}^* \text{ is a maximum.} \end{aligned} \tag{4.14}$$

4.2.2 Constrained quadratic functions

We have shown how to optimize the n-dimensional quadratic form by searching the whole of $\mathbb{R}^n$ for the solution vector. We now address the same problem, except we restrict the search region by imposing some linear constraints. Specifically, we aim to solve

$$\text{minimize } q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$$

subject to $m < n$ linear constraints given by

$$\mathbf{P}\mathbf{x} = \begin{pmatrix} p_{11} & \cdots & \cdots & p_{1n} \\ \vdots & & & \vdots \\ p_{m1} & \cdots & \cdots & p_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} d_1 \\ \vdots \\ d_m \end{pmatrix} = \mathbf{d}. \tag{4.15}$$

We assume that the constraints are linearly independent. We can express this condition more succinctly if we express the transpose of the constraint matrix in terms of its m column vectors, i.e.,

$$\mathbf{P}^T = (\mathbf{p}_1 \cdots \mathbf{p}_m) \quad \text{where} \quad \mathbf{p}_k = \begin{pmatrix} p_{k1} \\ \vdots \\ \vdots \\ p_{kn} \end{pmatrix} \in \mathbb{R}^n \quad 1 \le k \le m.$$

The assumption of linearly independent constraints is then equivalent to demanding that the column vectors $\{\mathbf{p}_1, \ldots, \mathbf{p}_m\}$ are linearly independent, and this holds if and only if

$$\text{the only vector } \boldsymbol{\lambda} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix} \in \mathbb{R}^m \text{ that satisfies } \mathbf{P}^T\boldsymbol{\lambda} = \mathbf{0} \text{ is } \boldsymbol{\lambda} = \mathbf{0}. \tag{4.16}$$

To solve this problem we introduce an m-dimensional vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m)^T$ and construct the so-called Lagrange function

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c - \boldsymbol{\lambda}^T(\mathbf{d} - \mathbf{P}\mathbf{x}).$$

The Lagrange function simply involves subtracting multiples of each of the constraints from the original quadratic form. We note that $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = q(\mathbf{x})$ whenever $\mathbf{x}$ satisfies the constraints.

To investigate the potential stationary values of our problem we take the gradient of the Lagrangian with respect to $\mathbf{x}$ and $\boldsymbol{\lambda}$ and set it equal to zero, i.e.,

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{A}\mathbf{x}^* + \mathbf{b} + \mathbf{P}^T\boldsymbol{\lambda}^* = \mathbf{0},$$
$$\nabla_{\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{d} - \mathbf{P}\mathbf{x}^* = \mathbf{0}.$$

This rearranges to

$$\mathbf{A}\mathbf{x}^* + \mathbf{P}^T\boldsymbol{\lambda}^* = -\mathbf{b},$$
$$\mathbf{P}\mathbf{x}^* = \mathbf{d},$$

or, in matrix–vector form,

$$\begin{pmatrix} \mathbf{A} & \mathbf{P}^T \\ \mathbf{P} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x}^* \\ \boldsymbol{\lambda}^* \end{pmatrix} = \begin{pmatrix} -\mathbf{b} \\ \mathbf{d} \end{pmatrix}. \tag{4.17}$$

It can be shown that if this system has a unique solution then $\mathbf{x}^*$ is the location of either a maximum or minimum of the constrained problem; see Nash and Sofer (1996). In view of this we have the following result:

Lemma 4.3. *The matrix*

$$\begin{pmatrix} \boldsymbol{A} & \boldsymbol{P}^T \\ \boldsymbol{P} & \boldsymbol{0} \end{pmatrix}$$

is invertible if either $\boldsymbol{A}$ or $-\boldsymbol{A}$ is positive definite on the $(n-m)$-dimensional subspace of $\mathbb{R}^n$ defined by

$$\mathcal{Z} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{P}\boldsymbol{x} = \boldsymbol{0}\}.$$

Proof. In view of (2.11) we can demonstrate that the matrix is invertible under the conditions of the lemma by showing they imply that

$$\begin{pmatrix} \mathbf{A} & \mathbf{P}^T \\ \mathbf{P} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} \mathbf{x} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

The matrix system gives

$$\mathbf{A}\mathbf{x} + \mathbf{P}^T\boldsymbol{\lambda} = \mathbf{0},$$
$$\mathbf{P}\mathbf{x} = \mathbf{0}.$$

We note that $\mathbf{Px} = \mathbf{0}$ implies $\mathbf{x} \in \mathcal{Z}$, and thus

$$\begin{aligned} 0 &= \mathbf{x}^T\mathbf{Ax} + \mathbf{x}^T\mathbf{P}^T\boldsymbol{\lambda} \\ &= \mathbf{x}^T\mathbf{Ax} + \left(\boldsymbol{\lambda}^T \underbrace{\mathbf{Px}}_{=\mathbf{0}}\right)^T = \mathbf{x}^T\mathbf{Ax}. \end{aligned}$$

Since we assume that either $\mathbf{A}$ or $-\mathbf{A}$ is positive definite on $\mathcal{Z}$, we can only conclude $\mathbf{x} = \mathbf{0}$.

Armed with this information, the first equation now reads $\mathbf{P}^T\boldsymbol{\lambda} = \mathbf{0}$; this equation can only have $\boldsymbol{\lambda} = \mathbf{0}$ as a solution since we have assumed that the columns of $\mathbf{P}^T$ are linearly independent, see (4.16). This completes the proof. □

If the conditions of the above lemma hold then we are able to conclude:

- The linear system (4.17) has a unique solution $(\mathbf{x}^*, \boldsymbol{\lambda}^*)^T \in \mathbb{R}^{n+m}$.
- The solution $\mathbf{x}^*$ corresponds to the location of a stationary point of our constrained problem (4.15).

In addition, it can be shown that

$$\begin{aligned} &\text{if } \mathbf{A} \text{ is positive definite on } \mathcal{Z} \text{ then } \mathbf{x}^* \text{ is a minimum;} \\ &\text{if } -\mathbf{A} \text{ is positive definite on } \mathcal{Z} \text{ then } \mathbf{x}^* \text{ is a maximum.} \end{aligned} \tag{4.18}$$

The theory we have developed in this section will be employed directly in the next chapter, where we focus upon selecting the financial portfolio which holds the minimum amount of risk whilst maintaining a certain level of expected return.

4.3 OVER-DETERMINED LINEAR SYSTEMS

In this section we revisit the problem of solving a general linear system of equations which, when written in matrix–vector form, has the representation

$$\mathbf{Ax} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1l} \\ a_{21} & a_{22} & \cdots & a_{2l} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{ml} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_l \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \vdots \\ y_m \end{pmatrix} = \mathbf{y}.$$

In the special case where the number of equations matches the number of unknowns, i.e., $m = l$, we already know (from the theory developed in Chapter 2) that a unique solution vector $\mathbf{x} \in \mathbb{R}^l$ exists provided the matrix $\mathbf{A}$ is non-singular. Unfortunately, in many real-world problems we often have more equations than we have unknowns, i.e., $m > l$, and we are faced with what is known as an over-determined system.

We cannot expect to find a unique solution to an over-determined system and so, instead, we look for a good approximation. Specifically, we consider vectors $\mathbf{x}$ which approximately satisfy the linear system, i.e., such that $\mathbf{Ax} \approx \mathbf{y}$. In order to quantify the approximation we

measure the distance $\|\mathbf{Ax} - \mathbf{y}\|$ of $\mathbf{Ax}$ from its target $\mathbf{y}$. Our aim is to find the best possible approximation and this translates into solving the following minimization problem:

$$\text{minimize } \left\{ \|\mathbf{Ax} - \mathbf{y}\| : \mathbf{x} \in \mathbb{R}^l \right\}.$$

The position of the minimum point does not change if we minimize the square of the distance instead, thus we consider the following equivalent version of the problem:

$$\begin{aligned}\text{minimize } \Big\{ \|\mathbf{Ax} - \mathbf{y}\|^2 &= (\mathbf{Ax} - \mathbf{y})^T (\mathbf{Ax} - \mathbf{y}) \\ &= \mathbf{x}^T \mathbf{A}^T \mathbf{Ax} - 2\mathbf{x}^T \mathbf{A}^T \mathbf{y} + \mathbf{y}^T \mathbf{y} : \mathbf{x} \in \mathbb{R}^l \Big\}.\end{aligned}$$

If we set

$$\mathbf{\Gamma} = \mathbf{A}^T \mathbf{A} \in \mathbb{R}^{l \times l} \text{ and } \mathbf{b} = \mathbf{A}^T \mathbf{y} \in \mathbb{R}^l,$$

then the problem can be posed as

$$\text{minimize } q(\mathbf{x}) = \mathbf{x}^T \mathbf{\Gamma x} - 2\mathbf{b}^T \mathbf{x} + \|\mathbf{y}\|^2, \text{ for } \mathbf{x} \in \mathbb{R}^l. \tag{4.19}$$

The solution relies on minimizing the quadratic form $q(\mathbf{x})$, thus we seek the vector $\mathbf{x}^*$ where the gradient of q is zero, i.e., we solve

$$\nabla_{\mathbf{x}} q(\mathbf{x}^*) = 2(\mathbf{\Gamma x}^* - \mathbf{b}) = \mathbf{0}.$$

Thus, the vector $\mathbf{x}^*$ satisfies

$$\mathbf{\Gamma x}^* = \mathbf{b}.$$

We note that $\mathbf{\Gamma}$ is a non-negative definite matrix since, for any $\mathbf{x} = (x_1, \ldots, x_l)^T \in \mathbb{R}^l$, we have

$$\mathbf{x}^T \mathbf{\Gamma x} = (\mathbf{Ax})^T (\mathbf{Ax}) = \|\mathbf{Ax}\|^2 \geq 0.$$

Furthermore, if we let $\{\mathbf{a}_1, \ldots, \mathbf{a}_l\}$ denote the m-dimensional column vectors of $\mathbf{A}$ then we can rewrite the above quadratic form as

$$\mathbf{x}^T \mathbf{\Gamma x} = \|\mathbf{Ax}\|^2 = \Big\| \sum_{i=1}^{l} x_i \mathbf{a}_i \Big\|^2 \geq 0.$$

In particular, we can conclude that if the column vectors of $\mathbf{A}$ are linearly independent then $\mathbf{\Gamma}$ is positive definite; this follows because, in this case, we know

$$\begin{aligned}\Big\| \sum_{i=1}^{l} x_i \mathbf{a}_i \Big\|^2 = 0 &\Leftrightarrow \sum_{i=1}^{l} x_i \mathbf{a}_i = \mathbf{0} \\ &\Leftrightarrow x_1 = \cdots = x_l = 0, \text{ i.e., } \mathbf{x} = \mathbf{0}.\end{aligned}$$

The following result captures all of our observations.

Theorem 4.4. *Let $\boldsymbol{A}$ denote an $m \times l$ matrix with more rows than columns, i.e., $m > l$. If the column vectors of $\boldsymbol{A}$ are linearly independent then the vector $\boldsymbol{x}^*$ defined by*

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{A}^T\boldsymbol{y} \tag{4.20}$$

is the unique minimizer of the quadratic form

$$q(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|^2. \tag{4.21}$$

We close our discussion of this problem with a little extra background information.

- The solution $\mathbf{x}^*$ given by (4.16) is often called the linear least squares solution.
- The matrix $\boldsymbol{\Gamma} = \mathbf{A}^T\mathbf{A}$ is often called the normal matrix and the equations (4.20) are known as the normal equations. In practice, the solution $\mathbf{x}^*$ is hardly ever calculated by solving the normal equations directly; this route can be shown to be very unstable. Thankfully, there exist efficient algorithms that provide the least-squares solution in a computationally stable manner.

4.4 LINEAR REGRESSION

A common goal facing scientists and engineers is to gain insight into the nature of some random variable, denoted by Y say. We usually think of Y as the random variable that represents the outcome of an experiment. The experimenter will identify a collection of random factors $X_1, \ldots, X_l$ that are thought to influence the outcome of the experiment. The factors are chosen to be directly observable, i.e., their values are easy to obtain. To illustrate with a simple example we could imagine we need to investigate the random variable

$$Y \mapsto y = \text{number of times an individual visits the doctor per year.}$$

To clarify the notation; we use the capital letter Y to describe the experiment and the lower case y to describe a potential outcome of the experiment, thus in this example y can, theoretically, take on a value from the set $\{0, 1, 2, \ldots\}$.

There are a number of measurable factors that could conceivably influence the outcome of Y, for instance we could consider

$$\begin{aligned}
X_1 \mapsto x_1 &= \text{age of the individual;} \\
X_2 \mapsto x_2 &= \text{weight of the individual;} \\
X_3 \mapsto x_3 &= \text{alcohol consumption of the individual per week;} \\
X_4 \mapsto x_4 &= \text{number of hours of exercise undertaken per week;} \\
&\vdots \\
&\text{etc.}
\end{aligned}$$

Once equipped with a collection of influential factors the experimenter typically proposes an approximate linear relationship

$$y \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_l x_l,$$

or equivalently

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_l x_l + \varepsilon, \tag{4.22}$$

where ε is a random but non-observable error term.

The investigator believes that (4.22) is correct and has the task of determining the coefficients $\beta_0, \beta_1, \ldots, \beta_l$. In order to do this, data have to be collected. We consider two cases:

- If Y represents an experiment then it can be repeated many times; on each occasion the outcome y is noted together with the values of the underlying factors.
- If Y represents a random process which evolves through time then data are gathered by noting the past realization of Y and the corresponding values of the factors.

In either case the result is that we have the relevant data, namely the kth outcome/realization $= y_k$

$$\text{together with } \underbrace{(x_1^{(k)}, x_2^{(k)}, \ldots, x_l^{(k)})}_{\text{corresponding factor values}} \quad \text{for } k = 1, 2, \ldots, m \ (m >> l).$$

We collect this information in a realization vector $\mathbf{y}$ and a factor matrix $\mathbf{X}$ defined by

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_m \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & x_1^{(1)} & \cdots & x_l^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_l^{(2)} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 1 & x_1^{(m)} & \cdots & x_l^{(m)} \end{pmatrix} \in \mathbb{R}^{m \times (l+1)}.$$

In view of the linear relationship (4.22) we can form the following over-determined linear system:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4.23}$$

where

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_l \end{pmatrix} \in \mathbb{R}^{l+1} \quad \text{and} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_m \end{pmatrix} \in \mathbb{R}^m$$

denote the vectors of unknown beta coefficients and unobservable errors respectively. It can be a difficult task to find the precise coefficient vector $\boldsymbol{\beta}$ and so we aim to explore how

we can generate a useful estimate $\widehat{\boldsymbol{\beta}}$. To accommodate this we make the following initial assumptions:

$$\begin{aligned}&1.\ \text{The } l+1 \text{ column vectors of } \mathbf{X} \text{ are linearly independent.}\\ &2.\ \text{The vector } \boldsymbol{\varepsilon} \text{ of residual errors satisfies}\\ &\begin{cases}\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0} & \text{zero mean;}\\ \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2\mathbf{I}_m & \text{constant variance but uncorrelated.}\end{cases}\end{aligned} \tag{4.24}$$

In this framework we can use our previous work to find the vector $\widehat{\boldsymbol{\beta}}$ which minimizes the sum of squares of the error components. Specifically, we know that the statement

$$\widehat{\boldsymbol{\beta}} \text{ minimizes } \sum_{i=1}^{m} \varepsilon_i^2$$

is equivalent to

$$\widehat{\boldsymbol{\beta}} \text{ minimizes } \|\varepsilon\|^2 = \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|^2.$$

We can now evoke Theorem 4.16 to conclude that the unique solution to this problem is given by

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{4.25}$$

The whole process outlined above can be summarized as follows:

- Collect data from experiments or from past history.
- Use the data to form the over-determined linear system (4.23).
- Using assumption 1 of (4.24) compute the least-squares estimate $\widehat{\boldsymbol{\beta}}$ (4.25) of the true $\boldsymbol{\beta}$.

This process is known in statistics as a regression run. Let us investigate the outcome of the run in more detail. Specifically, we shall address the following questions:

1. How good is the fitted regression model?
2. Given that the residual error vector satisfies $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2\mathbf{I}_m$, how do we estimate the variance σ^2?
3. How good is the least-squares estimate of $\boldsymbol{\beta}$?

Goodness of Fit

To answer the first question we need to develop some mathematical theory. We kick this process off by observing that, given the estimated parameter vector $\widehat{\boldsymbol{\beta}}$ (4.25), the corresponding vector of fitted values $\widehat{\mathbf{y}}$ is given by

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{4.26}$$

To make our development a little cleaner we let

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \quad \text{and so} \quad \widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}. \tag{4.27}$$

We observe that $\mathbf{H}$ is a symmetric $m \times m$ matrix and, as we shall now reveal, it possesses several nice mathematical properties.

- $\mathbf{H}^2 = \mathbf{H}$.
  We see this by direct calculation:

$$\begin{aligned}\mathbf{H}^2 &= \mathbf{X}\underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}}_{\mathbf{I}_{l+1}}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \\ &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}.\end{aligned} \tag{4.28}$$

- $(\mathbf{I}_m - \mathbf{H})^2 = \mathbf{I}_m - \mathbf{H}$.
  This follows directly from

$$(\mathbf{I}_m - \mathbf{H})^2 = \mathbf{I}_m^2 - 2\mathbf{H} + \mathbf{H}^2 = \mathbf{I}_m - \mathbf{H}. \tag{4.29}$$

- $\mathbf{H}(\mathbf{I}_m - \mathbf{H}) = \mathbf{0}_m$ (the $m \times m$ matrix consisting entirely of zeroes).
  Once more, a simple calculation gives

$$\mathbf{H}(\mathbf{I}_m - \mathbf{H}) = \mathbf{H} - \mathbf{H}^2 = \mathbf{0}_m. \tag{4.30}$$

- Only $l + 1$ columns of $\mathbf{H}$ are linearly independent.
  To establish this fact we let $\{\lambda_1, \ldots, \lambda_m\}$ denote the eigenvalues of $\mathbf{H}$ and let $\mathbf{\Gamma}$ denote the matrix whose columns are the corresponding eigenvectors of $\mathbf{H}$. We now turn to Theorem 2.4 to conclude that the spectral decomposition of $\mathbf{H}$ is given by

$$\mathbf{H} = \mathbf{\Gamma}\mathbf{D}\mathbf{\Gamma}^T \text{ where } \mathbf{D} = \text{diag } (\lambda_1, \ldots, \lambda_m).$$

  Now, using (4.28) we also have that

$$\mathbf{H} = \mathbf{H}^2 = \mathbf{\Gamma}\mathbf{D}\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{D}\mathbf{\Gamma}^T = \mathbf{\Gamma}\mathbf{D}^2\mathbf{\Gamma}^T.$$

  Combining these findings we have that

$$\mathbf{D} = \mathbf{D}^2 \Rightarrow \lambda_i^2 = \lambda_i \quad i = 1, \ldots, m,$$

  and this implies that each eigenvalue of $\mathbf{H}$ is either zero or one. As a result of this we can conclude that the number of linearly independent columns of $\mathbf{H}$ is equal to the number of its unit eigenvalues and, according to (2.22), this is given by the trace of $\mathbf{H}$. The following development establishes that $\mathbf{H}$ consists of precisely $l + 1$ linearly independent columns:

$$\begin{aligned}\sum_{i=0}^{m}\lambda_i &= \text{Trace}(\mathbf{H}) \\ &= \text{Trace}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\ &= \text{Trace}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}) \quad \text{using (2.23)} \\ &= \text{Trace}(\mathbf{I}_{l+1}) = l + 1.\end{aligned}$$

  We remark that as a consequence we can, in a similar fashion, also conclude that only $m - (l + 1)$ columns of $\mathbf{I}_m - \mathbf{H}$ are linearly independent.

The mathematical properties that we have established above enable us to provide a geometrical interpretation of the matrix $\mathbf{H}$. Specifically, for any $\mathbf{y} \in \mathbb{R}^m$, we can write

$$\mathbf{y} = \underbrace{\mathbf{H}\mathbf{y}}_{=\widehat{\mathbf{y}}} + \underbrace{(\mathbf{I}_m - \mathbf{H})\mathbf{y}}_{=\boldsymbol{\varepsilon}}.$$

Furthermore, using the properties of $\mathbf{H}$ we can show that

$$\begin{aligned}\widehat{\mathbf{y}}^T\boldsymbol{\varepsilon} &= \mathbf{y}^T\mathbf{H}(\mathbf{I}_m - \mathbf{H})\mathbf{y} \\ &= \mathbf{y}^T\mathbf{H}\mathbf{y} - \mathbf{y}^T\underbrace{\mathbf{H}^2}_{=\mathbf{H}}\mathbf{y} = 0,\end{aligned}$$

i.e., the fitted vector $\widehat{\mathbf{y}}$ is orthogonal to the residual error vector $\boldsymbol{\varepsilon}$. In more detail, we say that $\widehat{\mathbf{y}}$ is the orthogonal projection of the true vector $\mathbf{y}$ onto the $(l+1)$-dimensional space, $\mathcal{R}_{\mathbf{H}}$ say, that is spanned by the $l+1$ linearly independent column vectors of $\mathbf{H}$. The residual vector $\boldsymbol{\varepsilon}$ belongs to the orthogonal complement of $\mathcal{R}_{\mathbf{H}}$.

In view of this discovery we can take the Euclidean norm of $\mathbf{y} = \widehat{\mathbf{y}} + \boldsymbol{\varepsilon}$ to find that

$$\begin{aligned}\|\mathbf{y}\|^2 = \mathbf{y}^T\mathbf{y} &= (\widehat{\mathbf{y}} + \boldsymbol{\varepsilon})^T(\widehat{\mathbf{y}} + \boldsymbol{\varepsilon}) \\ &= \widehat{\mathbf{y}}^T\widehat{\mathbf{y}} + \underbrace{\widehat{\mathbf{y}}^T\boldsymbol{\varepsilon}}_{=0} + \underbrace{\boldsymbol{\varepsilon}^T\widehat{\mathbf{y}}}_{=0} + \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} \\ &= \|\widehat{\mathbf{y}}\|^2 + \|\boldsymbol{\varepsilon}\|^2.\end{aligned}$$

We recall that the residual error vector $\boldsymbol{\varepsilon}$ has zero mean and so, if we let $\widehat{\mathbf{e}}_{\mathbf{y}}$ denote an estimate for the mean of $\mathbf{y}$, we can also write that

$$\|\mathbf{y} - \widehat{\mathbf{e}}_{\mathbf{y}}\|^2 = \|\widehat{\mathbf{y}} - \widehat{\mathbf{e}}_{\mathbf{y}}\|^2 + \|\boldsymbol{\varepsilon}\|^2. \tag{4.31}$$

We recognize both of the above expressions as versions of Pythagoras' theorem. The second expression (4.31) is particularly illuminating as it states that the variability of $\mathbf{y}$ measured by $\|\mathbf{y} - \widehat{\mathbf{e}}_{\mathbf{y}}\|^2$ is captured in two parts:

- a component that is explained by the proposed factors measured by $\|\widehat{\mathbf{y}} - \widehat{\mathbf{e}}_{\mathbf{y}}\|^2$;
- a component that cannot be explained by the factors measured by $\|\boldsymbol{\varepsilon}\|^2$.

Naturally, we would consider our fitted linear model

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \ldots + \widehat{\beta}_l x_l + \varepsilon$$

to be a success if the proposed factors were found to capture the bulk of the variability of y. In view of this we consider the quantity

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\|\widehat{\mathbf{y}} - \widehat{\mathbf{e}}_{\mathbf{y}}\|^2}{\|\mathbf{y} - \widehat{\mathbf{e}}_{\mathbf{y}}\|^2} = 1 - \frac{\|\boldsymbol{\varepsilon}\|^2}{\|\mathbf{y} - \widehat{\mathbf{e}}_{\mathbf{y}}\|^2}. \tag{4.32}$$

In statistical terminology the above measure is more commonly known as the *coefficient of determination* for the regression and its value, by definition, belongs to the unit interval $[0, 1]$. We note here some of its features:

- If $R^2 = 1$ then all of the observed data points $(x_1^{(k)}, \ldots, x_l^{(k)}|y_k)$ satisfy

$$y_k = \widehat{\beta}_0 + \widehat{\beta}_1 x_1^{(k)} + \cdots + \widehat{\beta}_l x_l^{(k)} \quad k = 1, \ldots, m,$$

  i.e., they all lie on the line predicted by the regression.
- If $R^2 = 0$ then the random variable Y and the proposed explanatory random variables $X_1, \ldots, X_l$ are in fact uncorrelated.
- The value of R^2 clearly depends upon the data set that is used to run the regression. A different data set will potentially yield a different value of R^2. In view of this the coefficient itself can be viewed as a random variable in its own right.

In conclusion, the value of R^2 provides us with a measure of the goodness of fit of a particular regression; a value that is close to one provides evidence of a good fit.

Estimating the Variance of the Residual Error

In order to address this question we consider the following helpful result:

Lemma 4.5. *Let $\boldsymbol{y} \in \mathbb{R}^n$ be a random vector whose mean and covariance matrix are given by*

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{e_y} \quad \text{and} \quad \mathbb{E}[(\boldsymbol{y} - \boldsymbol{e_y})(\boldsymbol{y} - \boldsymbol{e_y})^T] = \boldsymbol{V}$$

respectively. Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a fixed symmetric matrix, then

$$\mathbb{E}[\boldsymbol{y}^T \boldsymbol{A} \boldsymbol{y}] = \text{Trace}(\boldsymbol{AV}) + \boldsymbol{e}_y^T \boldsymbol{A} \boldsymbol{e}_y. \tag{4.33}$$

Proof. We begin by observing that

$$\mathbf{y}^T \mathbf{A} \mathbf{y} = (\mathbf{y} - \mathbf{e}_\mathbf{y})^T \mathbf{A} (\mathbf{y} - \mathbf{e}_\mathbf{y}) + 2\mathbf{e}_\mathbf{y}^T \mathbf{A} \mathbf{y} - \mathbf{e}_\mathbf{y}^T \mathbf{A} \mathbf{e}_\mathbf{y}.$$

Using this we find that

$$\begin{aligned} \mathbb{E}[\mathbf{y}^T \mathbf{A} \mathbf{y}] &= \mathbb{E}[(\mathbf{y} - \mathbf{e}_\mathbf{y})^T \mathbf{A} (\mathbf{y} - \mathbf{e}_\mathbf{y}) + 2\mathbf{e}_\mathbf{y}^T \mathbf{A} \mathbf{y} - \mathbf{e}_\mathbf{y}^T \mathbf{A} \mathbf{e}_\mathbf{y}] \\ &= \mathbb{E}[(\mathbf{y} - \mathbf{e}_\mathbf{y})^T \mathbf{A} (\mathbf{y} - \mathbf{e}_\mathbf{y})] + 2\mathbf{e}_\mathbf{y}^T \mathbf{A} \mathbf{e}_\mathbf{y} - \mathbf{e}_\mathbf{y}^T \mathbf{A} \mathbf{e}_\mathbf{y} \\ &= \mathbb{E}[(\mathbf{y} - \mathbf{e}_\mathbf{y})^T \mathbf{A} (\mathbf{y} - \mathbf{e}_\mathbf{y})] + \mathbf{e}_\mathbf{y}^T \mathbf{A} \mathbf{e}_\mathbf{y}. \end{aligned} \tag{4.34}$$

Now since a scalar is its own trace, we have

$$\begin{aligned} \mathbb{E}[(\mathbf{y} - \mathbf{e}_\mathbf{y})^T \mathbf{A} (\mathbf{y} - \mathbf{e}_\mathbf{y})] &= \mathbb{E}\Big[\text{Trace}\Big((\mathbf{y} - \mathbf{e}_\mathbf{y})^T \mathbf{A} (\mathbf{y} - \mathbf{e}_\mathbf{y})\Big)\Big] \\ &= \mathbb{E}\Big[\text{Trace}\Big(\mathbf{A}(\mathbf{y} - \mathbf{e}_\mathbf{y})(\mathbf{y} - \mathbf{e}_\mathbf{y})^T\Big)\Big] \\ &= \text{Trace}\Big(\mathbf{A}\mathbb{E}\Big[(\mathbf{y} - \mathbf{e}_\mathbf{y})(\mathbf{y} - \mathbf{e}_\mathbf{y})^T\Big]\Big) \\ &= \text{Trace}(\mathbf{AV}), \end{aligned}$$

where the second line of the above development follows from (2.23). Now, substituting this result into (4.34) proves the lemma. □

We recall that the residual error vector $\boldsymbol{\varepsilon}$ can be expressed as

$$\boldsymbol{\varepsilon} = (\mathbf{I}_m - \mathbf{H})\mathbf{y}.$$

Using the properties of the matrix $\mathbf{H}$ we can conclude that

$$\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = \mathbf{y}^T(\mathbf{I}_m - \mathbf{H})\mathbf{y}.$$

The mean vector of $\mathbf{y}$ is given by

$$\mathbb{E}[\mathbf{y}] = \mathbf{e}_\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

and its covariance matrix is given by

$$\mathbb{E}[(\mathbf{y} - \mathbf{e}_\mathbf{y})(\mathbf{y} - \mathbf{e}_\mathbf{y})^T] = \mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2\mathbf{I}_m.$$

Now applying the result of the lemma with $\mathbf{A} = \mathbf{I}_m - \mathbf{H}$ we find that

$$\mathbb{E}[\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}] = \sigma^2 \text{ Trace } (\mathbf{I}_m - \mathbf{H}) + \widehat{\boldsymbol{\beta}}^T\mathbf{X}^T(\mathbf{I}_m - \mathbf{H})\mathbf{X}\widehat{\boldsymbol{\beta}}. \tag{4.35}$$

We make the following observations:

- We have already established that Trace $(\mathbf{I}_m - \mathbf{H}) = m - l - 1$.
- Using (4.26) and (4.27) we note that
$$\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$$
and thus, appealing to the properties of $\mathbf{H}$, we find that the second term in (4.35) is zero.

As a result, we can conclude that (4.35) becomes

$$\mathbb{E}[\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}] = (m - l - 1)\sigma^2,$$

and so we propose that the variance of the residual error terms be estimated via the following formula:

$$\widehat{\sigma}^2 = \frac{1}{m - l - 1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}). \tag{4.36}$$

The Least-squares Estimate of $\boldsymbol{\beta}$

It is clear that, in general, different runs of a regression will provide different estimates of $\boldsymbol{\beta}$. In order to investigate the properties of the least-squares estimate $\widehat{\boldsymbol{\beta}}$ (4.25), we present the following helpful relationship:

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \quad \text{by (4.25)} \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}) \quad \text{by (4.23)} \\
&= \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}}_{=\mathbf{I}_n}\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon} \\
&= \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon}.
\end{aligned}$$

Thus we have

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\varepsilon} \tag{4.37}$$

and, due to the mean assumption (4.24), we can immediately deduce that

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}. \tag{4.38}$$

In statistical terminology we say that $\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of the true $\boldsymbol{\beta}$. This is a desirable property of any statistical estimator and is a concept we will examine in more detail in Chapter 16.

Essential Mathematics for Market Risk Management
by Simon Hubbert

5 Portfolio Theory I

We open this chapter with the following simple investment problem. An investor has set aside a certain amount of wealth denoted by W and decides to invest this in a group of risky assets from the stock market, i.e., he has decided to build an investment portfolio. The problem is deciding just how the wealth should be distributed amongst the constituent assets, see Figure 5.1.

This is not a straightforward task; we know from the familiar charts in the financial press that asset prices evolve randomly and there is no way of knowing their future values. In theory, an asset can take on any one of a range of positive future values. In this chapter we shall use elements of the previous three chapters to establish a simple mathematical framework within which we can set up and solve the above problem.

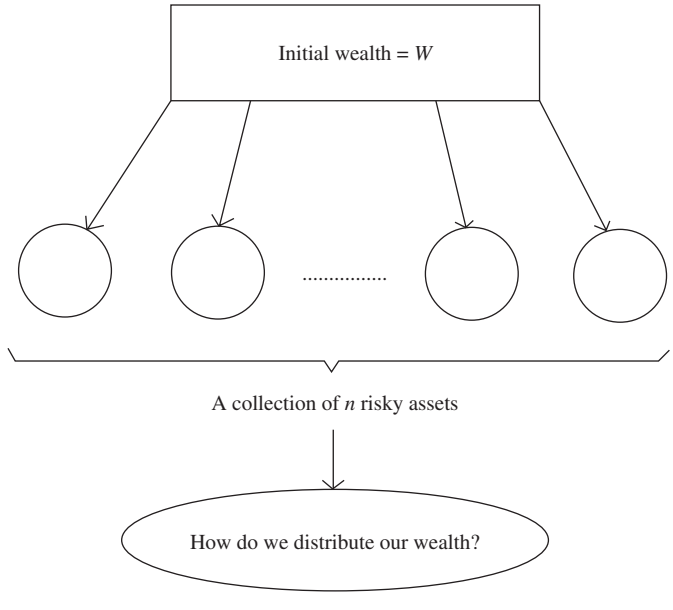


Figure 5.1 The task of building an investment.

5.1 MEASURING RETURNS

In order to express our financial investment problems mathematically, we need to identify and describe their basic features. To begin with we shall consider the simple case of a single asset. We know the price of the asset today, at time t, but its value at a future date, at time $t + \tau$ say, is unknown. There are essentially two commonly used approaches for measuring

the τ-day random return on a given asset:

$$r(t:\tau) = \begin{cases} r_{\text{std}}(t:\tau) = \frac{S(t+\tau)-S(t)}{S(t)} & \text{the standard return;} \\ r_{\log}(t:\tau) = \log\left(\frac{S(t+\tau)}{S(t)}\right) & \text{the log return.} \end{cases} \tag{5.1}$$

5.1.1 A comparison of the standard and log returns

The practical question of which return rate to use, standard or logarithmic, is a subjective one and depends upon the context of the problem.

Simple Relationships and Observations

As a starting point for a comparison we observe that we can rearrange (5.1) to express the price ratio of the asset as

$$\frac{S(t+\tau)}{S(t)} = 1 + r_{\text{std}}(t:\tau). \tag{5.2}$$

Taking logs provides us with the relationship between the two measures

$$\begin{aligned} r_{\log}(t:\tau) &= \log(1 + r_{\text{std}}(t:\tau)) \\ \text{and } \; r_{\text{std}}(t:\tau) &= \exp(r_{\log}(t:\tau)) - 1. \end{aligned} \tag{5.3}$$

We choose the return rate, standard or logarithmic, as our main random variable of interest. We can easily convert from rates (i.e., percentages) into prices (i.e., monetary units) by applying the following formulae:

$$S(t+\tau) = \begin{cases} S(t)(1 + r_{\text{std}}(t:\tau)) & \text{for the standard return;} \\ S(t)\exp(r_{\log}(t:\tau)) & \text{for the log return.} \end{cases} \tag{5.4}$$

We note that the standard rate is bounded below by -1 this places a constraint upon the form of the underlying distribution function. In particular, we are unable to fit standard return data to many of the popular distributions, e.g., the normal distribution. The range of the log return rate, on the other hand, is the whole real line, thus it is much more amenable to a parametric fit.

Another useful way of investigating the difference between the two measures is to use the series expansion for the exponential function (3.21) to yield

$$r_{\text{std}}(t:\tau) = r_{\log}(t:\tau) + \underbrace{\frac{\left(r_{\log}(t:\tau)\right)^2}{2!} + \frac{\left(r_{\log}(t:\tau)\right)^3}{3!} + \cdots}_{\text{high-order terms}}$$

We remark that if the log return is small then we can ignore the high-order terms in the above expansion and conclude that the two measures are approximately the same, and so it should make little difference which one is employed. The log return is zero if there is no change in the price and the measure tends to remain small over short time intervals, e.g., over one day, when (generally) there is little variation in the price. In view of this it is

common to accept the approximation

$$r_{\text{std}}(t:\tau) \approx r_{\log}(t:\tau) \text{ for } \tau \leq 1 \text{ day}.$$

Investigating Linearity

The return rate is flexible and is perhaps the best way to analyse and compare different investments. For instance, suppose an investor decides he wants to invest a portion of his wealth $\$W$ say, in a stock whose value today is given by $S(t)$. In theory, the investor buys α shares of the asset, where α is a number that satisfies $W = \alpha S(t)$. We can then use the random return rates to describe the future of this investment as

$$\begin{array}{ccc} \text{today } t & \rightarrow & \text{future } t+\tau \\ \$W & \rightarrow & \begin{cases} \$W \cdot (1 + r_{\text{std}}(t:\tau)) & \text{standard;} \\ \$W \exp(r_{\log}(t:\tau)) & \text{logarithmic.} \end{cases} \end{array}$$

An appealing property of the standard return rate is that it is linear. To illustrate what this means, we assume that our investor constructs a portfolio of n assets $\{S_1, \ldots, S_n\}$ by partitioning his wealth so that

$$W_i \text{ is invested in } S_i \text{ for } i = 1, \ldots, n.$$

The τ-day standard portfolio return rate is defined by

$$\begin{aligned} r_{\text{std}}^{(p)}(t:\tau) &= \frac{(\text{future value at } t+\tau) - (\text{initial value})}{\text{initial value}} \\ &= \frac{\sum_{i=1}^n W_i (1 + r_{\text{std}}^{(i)}(t:\tau)) - \sum_{i=1}^n W_i}{\sum_{i=1}^n W_i} \\ &= \frac{\sum_{i=1}^n W_i r_{\text{std}}^{(i)}(t:\tau)}{W} \\ &= \sum_{i=1}^n w_i r_{\text{std}}^{(i)}(t:\tau), \end{aligned} \tag{5.5}$$

where

$$w_i = \frac{W_i}{W} \text{ for } i = 1, \ldots, n \text{ are portfolio weights} : \sum_{i=1}^n w_i = 1.$$

A useful consequence of this is the following neat expression for the change in value of the portfolio:

$$\begin{aligned} \Delta V &= (\text{future value at } t+\tau) - (\text{initial value}) \\ &= W \cdot \sum_{i=1}^n w_i r_{\text{std}}^{(i)}(t:\tau) \\ &= W \cdot r_{\text{std}}^{(p)}(t:\tau). \end{aligned} \tag{5.6}$$

If logarithmic return rates are used then, using (5.3), we can rewrite (5.5) as

$$\begin{aligned} \exp\Big(r_{\log}^{(p)}(t:\tau)\Big) &= \sum_{i=1}^{n} w_i \exp\Big(r_{\log}^{(i)}(t:\tau)\Big) \\ \Rightarrow r_{\log}^{(p)}(t:\tau) &= \log\Big[\sum_{i=1}^{n} w_i \exp\Big(r_{\log}^{(i)}(t:\tau)\Big)\Big]. \end{aligned} \tag{5.7}$$

In this case the portfolio return cannot be expressed as the weighted linear combination of the log returns. This does not pose a great problem because we can use a combination of (5.6) and (5.3) to write

$$\begin{aligned} \Delta V &= (\text{future value at } t+\tau) - (\text{initial value}) \\ &= W \cdot \left(\sum_{i=1}^{n} w_i (\exp(r_{\log}^{(i)}(t:\tau)) - 1)\right) \\ &= W \left(\sum_{i=1}^{n} w_i \exp(r_{\log}^{(i)}(t:\tau)) - 1\right). \end{aligned}$$

We can then argue, as before, that at small time intervals, when there are only slight deviations in asset prices, we have the approximate linear relationship

$$\begin{aligned} \Delta V &= (\text{future value at } t+\tau) - (\text{initial value}) \\ &\approx W \left(\sum_{i=1}^{n} w_i r_{\log}^{(i)}(t:\tau)\right). \end{aligned}$$

Comparing this to (5.6) we can conclude that, at small time intervals, we have

$$r_{\text{std}}^{(p)}(t:\tau) \approx \sum_{i=1}^{n} w_i r_{\log}^{(i)}(t:\tau) \quad \text{for } \tau \le 1 \text{ day}.$$

Longer-Period Returns

In the case where $\tau > 1$ day, it is useful to represent the price ratio of an asset as a product of daily ratios as follows:

$$\frac{S(t+\tau)}{S(t)} = \underbrace{\frac{S(t+1)}{S(t)}\frac{S(t+2)}{S(t+1)}\cdots\frac{S(t+\tau-1)}{S(t+\tau-2)}\frac{S(t+\tau)}{S(t+\tau-1)}}_{\text{product of future 1-day ratios}}. \tag{5.8}$$

Using this identity, together with (5.2), we can express the τ-day standard return rate as

$$r_{\text{std}}(t:\tau) = \prod_{\Delta=0}^{\tau-1} (1 + r_{\text{std}}(t+\Delta:1)) - 1, \tag{5.9}$$

and the τ-day log return rate as

$$\begin{aligned} r_{\log}(t:\tau) &= \log\left(\frac{S(t+1)}{S(t)}\frac{S(t+2)}{S(t+1)}\cdots\frac{S(t+\tau-1)}{S(t+\tau-2)}\frac{S(t+\tau)}{S(t+\tau-1)}\right) \\ &= \underbrace{\sum_{\Delta=0}^{\tau-1}\log\left(\frac{S(t+\Delta+1)}{S(t+\Delta)}\right)}_{\text{sum of future 1-day log returns}} \\ &= \sum_{\Delta=0}^{\tau-1} r_{\log}(t+\Delta:1). \end{aligned} \tag{5.10}$$

The multi-period log formula (5.10) is an appealing one; if we are able to model the evolution of the daily log return then the model for the multi-period return can easily be constructed. We shall return to this property later in the book.

In Summary

There is no universal agreement regarding which measure of return should be used, both have their strengths and weaknesses. We collect together our main observations.

- **Standard**
  The standard return possesses the linearity property (5.5) and this makes it a good candidate for portfolio analysis. The perceived drawback of the standard return is that we have to be careful when making assumptions regarding its distribution function; it is bounded below by -1 and this rules out many of the popular choices such as the normal distribution.
- **Logarithmic**
  The log return does not possess the linearity property, although an approximate linear relationship exists over a small holding period. One of the most interesting properties of the log return is that a longer-period return can be expressed as a sum of future daily returns. Under the right conditions this property can be exploited; insight into the evolution of daily returns can be used so solve problems involving multi-period returns.
  Unlike the standard return, the log return can theoretically take any value on $\mathbb{R}$ and so can easily be fit to a whole host of popular probability distributions. This fact is used by many academics and practitioners who aim to model the way a stock price evolves through time.

5.2 SETTING UP THE OPTIMAL PORTFOLIO PROBLEM

A financial portfolio is a fundamental investment, it is manufactured from a collection of basic financial assets and its composition depends upon the investors' preferences and requirements. In mathematical terms our intention is to build a portfolio from a set of n assets denoted by $\{S_1, \ldots, S_n\}$. Given that we intend to hold our portfolio for a total τ-days, the corresponding standard return rates for this period are given by

$$\{r_1, \ldots, r_n\} \quad \text{where} \quad r_i = \frac{S_i(t+\tau) - S_i(t)}{S_i(t)}, \quad i = 1, \ldots, n,$$

and thus, for a collection of portfolio weights $\{w_1, \ldots, w_n\}$ that satisfy

$$\sum_{i=1}^{n} w_i = 1, \tag{5.11}$$

the corresponding τ-day portfolio return is given by

$$r_p = \sum_{i=1}^{n} w_i r_i. \tag{5.12}$$

We can employ vector notation and write

$$\mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix}, \mathbf{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \text{ and } \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{5.13}$$

then equations (5.11) and (5.12) can be expressed as

$$r_p = \mathbf{w}^T \mathbf{r} \quad \text{and} \quad \mathbf{w}^T \mathbf{1} = 1 \quad \text{respectively.}$$

We remark that some of the entries of the weight vector $\mathbf{w}$ may be negative; this corresponds to a short-selling strategy for the assets in question.

The portfolio is completely determined by its portfolio weights: the potential reward and the (unavoidable) risk attached to a given portfolio can be altered by a modification of the portfolio weights. Our aim is to find the weights that somehow provide the optimal balance between risk and expected reward.

- **Reward**
  To measure the potential reward we can use the expected return rate on the portfolio given by

$$\mu_p = \mathbb{E}[r_p] = \sum_{i=1}^{n} w_i \mathbb{E}[r_i] = \sum_{i=1}^{n} w_i \mu_i.$$

  In matrix vector form we write

$$\mu_p = \mathbf{w}^T \mathbf{e} \quad \text{where } \mathbf{e} = \mathbb{E}[\mathbf{r}] = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}. \tag{5.14}$$

- **Risk**
  The variance of the portfolio return serves as a measure of portfolio risk. To compute this we need the covariance information of all the return rates. We recall from (3.16) that the covariance matrix can be expressed as

$$\mathbf{V} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \ldots & \sigma_{nn} \end{pmatrix} = \mathbb{E}[(\mathbf{r} - \mathbf{e})(\mathbf{r} - \mathbf{e})^T].$$

The variance of the portfolio with weight vector $\mathbf{w}$ is then given by

$$\sigma_p^2 = \mathbf{w}^T \mathbf{V} \mathbf{w} = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij}. \tag{5.15}$$

Every portfolio weight vector (whose component weights sum to one) defines a feasible portfolio and for each feasible portfolio we can compute its expected return μ_p and its portfolio volatility σ_p. In order to investigate the way these values depend upon the portfolio weights we can display the information in (σ, μ)-space, i.e., a space where we can plot the coordinates (σ_p, μ_p) for a given feasible portfolio p. We begin by performing a simple experiment with the following steps:

1. Generate a feasible weight vector $\mathbf{w}_p \in \mathbb{R}^n$.
2. Plug $\mathbf{w}_p$ into equations (5.14) and (5.15) to generate the expected return μ_p and volatility σ_p for the feasible portfolio p.
3. Plot the point (σ_p, μ_p) in (σ, μ)-space and repeat many times for many different weight vectors.

As the above experiment evolves we begin to see a distinct area of (σ, μ)-space that is occupied by feasible portfolios, this area is bullet shaped and appears to have a very definite boundary (see Figure 5.2). If we take a closer look at Figure 5.2 we see that we have randomly generated four portfolios, each with different feasible weights, for which the expected return is 7%. If, as investors, we are happy with 7% as a potential rate of return then we would always choose the portfolio with the least risk. In a nutshell we want to be on the perceived boundary or frontier of the diagram. We want to achieve the 7% with minimum risk!

Our task can now be defined. Let's suppose our investor is aiming to achieve a return on his portfolio of $100 \times \mu\%$, then in order to do this feasibly (i.e., with a portfolio whose

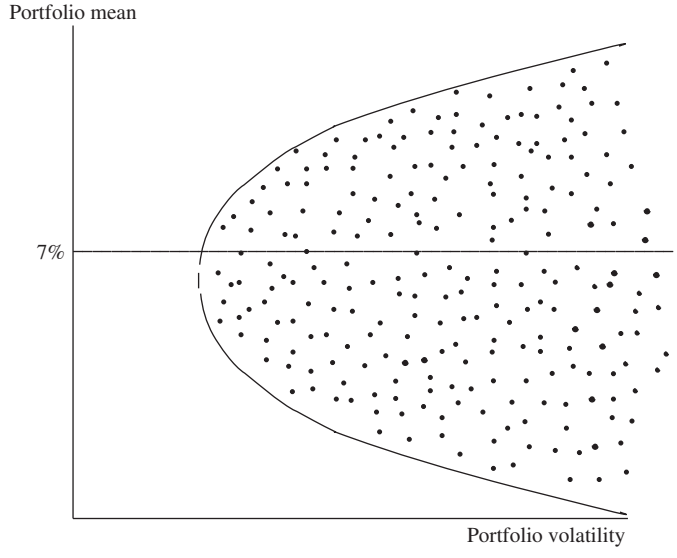


Figure 5.2 Filling (σ, μ)-space with feasible portfolios.

weights sum to one) and with the least risk we must find the portfolio weight vector that solves the following optimization problem:

$$\begin{aligned} \text{minimize } & \frac{1}{2}\mathbf{w}^T\mathbf{V}\mathbf{w} \\ \text{subject to } & \mathbf{w}^T\mathbf{e} = \mu_{\mathbf{p}} \quad \text{expected return matches desired rate} \\ & \mathbf{w}^T\mathbf{1} = 1 \quad \text{achieved with a feasible portfolio.} \end{aligned} \tag{5.16}$$

We remark that scaling the variance by a factor of $1/2$ does not affect the location of the optimal solution; it is merely to ensure a cleaner mathematical solution.

We can write the two constraints in matrix form as

$$\mathbf{P}\mathbf{w} = \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_n \\ 1 & 1 & \cdots & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} \mu \\ 1 \end{pmatrix} = \mathbf{d},$$

and so our optimal portfolio problem can be stated neatly as

$$\text{minimize } \frac{1}{2}\mathbf{w}^T\mathbf{V}\mathbf{w} \quad \text{subject to} \quad \mathbf{P}\mathbf{w} = \mathbf{d}. \tag{5.17}$$

5.3 SOLVING THE OPTIMAL PORTFOLIO PROBLEM

The mathematical formulation of the optimal portfolio problem (5.17) dates back to the 1950s and is attributed to Harry Markowitz who, as a PhD student at the University of Chicago, investigated and ultimately solved the problem in the course of his doctoral studies. Markowitz' solution has the remarkable (and very appealing) property that it can be visualized; thus anyone with a passing interest in financial investments can see, at a glance, the smallest amount of risk needed to achieve a given expected return. In short, Markowitz derived the precise equation of the frontier of Figure 5.2 and his work has had a huge impact on modern finance; he was amongst the first to apply scientific tools to solve hard problems and has subsequently inspired many more researchers to develop mathematical techniques for financial applications.

The Markowitz framework relies upon the following two mild assumptions:

- **A1**. The expected returns are not all the same value; this ensures that the two rows of $\mathbf{P}$ are linearly independent.
- **A2**. The covariance matrix $\mathbf{V}$ is positive definite.

The solution itself uses a Lagrangian technique from the optimization toolbox of Chapter 4; specifically we seek the solution $(\mathbf{w}_p, \boldsymbol{\lambda}^*)^T \in \mathbb{R}^{n+2}$ to the linear system

$$\begin{pmatrix} \mathbf{V} & \mathbf{P}^T \\ \mathbf{P} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{w}_p \\ \boldsymbol{\lambda}^* \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{d} \end{pmatrix},$$

i.e.,

$$\mathbf{V}\mathbf{w}_p + \mathbf{P}^T\boldsymbol{\lambda}^* = \mathbf{0} \;\Rightarrow\; \mathbf{V}\mathbf{w}_p = -\lambda_1^*\mathbf{e} - \lambda_2^*\mathbf{1}, \tag{5.18}$$

where $\boldsymbol{\lambda}^* = (\lambda_1^*, \lambda_2^*)^T$ is the vector of Lagrange multipliers and $\mathbf{w}_p$ is the optimal weight vector which satisfies

$$\mathbf{P}\mathbf{w}_p = \mathbf{d} \;\Rightarrow\; \begin{cases} \mathbf{e}^T\mathbf{w}_p = \mu_p, \\ \mathbf{1}^T\mathbf{w}_p = 1. \end{cases} \tag{5.19}$$

Since $\mathbf{V}$ is positive definite it is invertible, thus we can solve (5.18) to give

$$\mathbf{w}_p = -\lambda_1^*\mathbf{V}^{-1}\mathbf{e} - \lambda_2^*\mathbf{V}^{-1}\mathbf{1}. \tag{5.20}$$

We can then substitute this expression into the constraint equations (5.19) to give

$$\begin{aligned} -(\mathbf{e}^T\mathbf{V}^{-1}\mathbf{e})\lambda_1^* - (\mathbf{e}^T\mathbf{V}^{-1}\mathbf{1})\lambda_2^* &= \mu_p, \\ -(\mathbf{1}^T\mathbf{V}^{-1}\mathbf{e})\lambda_1^* - (\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1})\lambda_2^* &= 1. \end{aligned} \tag{5.21}$$

For a cleaner presentation we follow the approach of Huand and Litzenberger (1988) and set

$$\begin{pmatrix} B & A \\ A & C \end{pmatrix} = \begin{pmatrix} \mathbf{e}^T\mathbf{V}^{-1}\mathbf{e} & \mathbf{e}^T\mathbf{V}^{-1}\mathbf{1} \\ \mathbf{1}^T\mathbf{V}^{-1}\mathbf{e} & \mathbf{1}^T\mathbf{V}^{-1}\mathbf{1} \end{pmatrix}, \tag{5.22}$$

then the constraint equations can be rewritten in matrix form as

$$-\begin{pmatrix} B & A \\ A & C \end{pmatrix}\begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix} = \begin{pmatrix} \mu_p \\ 1 \end{pmatrix}. \tag{5.23}$$

We know from Chapter 2 that a unique solution for λ_1^* and λ_2^* exists if and only if the determinant of the matrix is non-zero. The following claim establishes that this is indeed the case.

Claim 5.1. *The determinant*

$$D = \mathrm{Det}\begin{pmatrix} B & A \\ A & C \end{pmatrix} = BC - A^2 \tag{5.24}$$

is positive and hence non-zero.

Proof. Since $\mathbf{V}^{-1}$ is positive definite it has a Choleski decomposition (see Theorem 2.6)

$$\mathbf{V}^{-1} = \mathbf{R}\mathbf{R}^T, \quad \text{where} \quad \mathbf{R} \in \mathbb{R}^{n\times n} \text{ is lower triangular.}$$

Using this we can consider the vectors

$$\mathbf{u} = \mathbf{R}^T\mathbf{e} \quad \text{and} \quad \mathbf{v} = \mathbf{R}^T\mathbf{1}.$$

We note that, from these choices, we can deduce

$$\mathbf{u}^T\mathbf{v} = \mathbf{e}^T\mathbf{R}\mathbf{R}^T\mathbf{1} = \mathbf{e}^T\mathbf{V}^{-1}\mathbf{1} = A,$$
$$\mathbf{u}^T\mathbf{u} = \mathbf{e}^T\mathbf{R}\mathbf{R}^T\mathbf{e} = \mathbf{e}^T\mathbf{V}^{-1}\mathbf{e} = B,$$
$$\mathbf{v}^T\mathbf{v} = \mathbf{1}^T\mathbf{R}\mathbf{R}^T\mathbf{1} = \mathbf{1}^T\mathbf{V}^{-1}\mathbf{1} = C.$$

We now evoke the Cauchy–Schwarz inequality, a geometric result which tells us that for any two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ we have

$$(\mathbf{u}^T\mathbf{v})^2 \leq (\mathbf{u}^T\mathbf{u})(\mathbf{v}^T\mathbf{v})$$

with equality if and only if $\mathbf{u}$ and $\mathbf{v}$ are linearly dependent. Applying this to our vectors yields

$$A^2 \leq BC.$$

We recall that we have assumed that the expected returns of our assets are not all the same, this means that $\mathbf{e}$ and $\mathbf{1}$ are linearly independent and thus, so are $\mathbf{u}$ and $\mathbf{v}$. We can conclude then that the inequality above is strict, that is

$$D = BC - A^2 > 0.$$

□

We can now establish that the Lagrange multipliers λ_1^* and λ_2^* are given by

$$\begin{pmatrix} \lambda_1^* \\ \lambda_2^* \end{pmatrix} = -\frac{1}{D}\begin{pmatrix} C & -A \\ -A & B \end{pmatrix}\begin{pmatrix} \mu_p \\ 1 \end{pmatrix},$$

and so

$$\lambda_1^* = \frac{A - C\mu_p}{D} \quad \text{and} \quad \lambda_2^* = \frac{A\mu_p - B}{D}. \tag{5.25}$$

The weight vector $\mathbf{w}_p$ whose corresponding feasible portfolio p provides an expected return of $100 \times \mu_p\%$ with minimum risk is now available; we simply substitute the Lagrange multipliers above into equation (5.20) to discover that

$$\begin{aligned} \mathbf{w}_p &= \left(\frac{C\mu_p - A}{D}\right)\mathbf{V}^{-1}\mathbf{e} + \left(\frac{B - A\mu_p}{D}\right)\mathbf{V}^{-1}\mathbf{1} \\ &= \frac{1}{D}\left(B\mathbf{V}^{-1}\mathbf{1} - A\mathbf{V}^{-1}\mathbf{e}\right) + \frac{1}{D}\left(C\mathbf{V}^{-1}\mathbf{e} - A\mathbf{V}^{-1}\mathbf{1}\right)\mu_p. \end{aligned} \tag{5.26}$$

To simplify this expression we let

$$\mathbf{g} = \frac{1}{D}\left(B\mathbf{V}^{-1}\mathbf{1} - A\mathbf{V}^{-1}\mathbf{e}\right) \quad \text{and} \quad \mathbf{h} = \frac{1}{D}\left(C\mathbf{V}^{-1}\mathbf{e} - A\mathbf{V}^{-1}\mathbf{1}\right), \tag{5.27}$$

and so we can write the optimal vector of portfolio weights as

$$\mathbf{w}_p = \mathbf{g} + \mathbf{h}\mu_p. \tag{5.28}$$

We note that the vectors **g** and **h** in formulae (5.27) depend only upon the covariance information contained in the matrix **V** and the vector of expected returns **e**. Crucially these vectors are independent of the desired level of expected return μ_p and thus the formula (5.28) provides the optimal weight vector for the whole range of expected returns. This means that the portfolio whose weight vector is given by (5.28) is guaranteed to be the feasible portfolio that provides an expected return $100 \times \mu\%$ with minimum risk. This is clearly an extremely useful discovery, however, it is only part of the story; what is missing is the actual value of the risk attached to these optimal portfolios. The fact that the risk is minimized is not enough for a potential investor; the risk needs to be quantified. The variance of the optimal portfolio corresponding to an expected return of $100 \times \mu\%$ is given by

$$\begin{aligned}
\sigma^2 &= \mathbf{w}^T \mathbf{V} \mathbf{w} \\
&= \mathbf{w}^T \mathbf{V} (\mathbf{g} + \mathbf{h}) \\
&= \mathbf{w}^T \mathbf{V} \left(\frac{1}{D} \left[B\mathbf{V}^{-1}\mathbf{1} - A\mathbf{V}^{-1}\mathbf{e} \right] + \mu \left[C\mathbf{V}^{-1}\mathbf{e} - A\mathbf{V}^{-1}\mathbf{1} \right] \right) \\
&= \frac{1}{D} \mathbf{w}^T \left([B\mathbf{1} - A\mathbf{e}] + \mu \left[C\mathbf{e} - A\mathbf{1} \right] \right) \quad \text{[follows since } \mathbf{V}^{-1}\mathbf{V} = \mathbf{I}_n \text{]} \\
&= \frac{1}{D} \left(C\mu^2 - 2A\mu + B \right) \quad \text{[follows since } \mathbf{w}^T\mathbf{e} = \mu \text{ and } \mathbf{w}^T\mathbf{1} = 1 \text{]} \\
&= \frac{C}{D} \left(\mu^2 - \frac{2A}{C}\mu + \frac{B}{C} \right) \\
&= \frac{C}{D} \left[\left(\mu - \frac{A}{C} \right)^2 + \frac{BC - A^2}{C^2} \right] \\
&= \frac{C}{D} \left(\mu - \frac{A}{C} \right)^2 + \frac{1}{C} \quad \text{[follows since } BC - A^2 = D \text{]}.
\end{aligned} \tag{5.29}$$

The strength of the above formula is that it is completely general. An investor whose aim is to construct a portfolio with an expected return of μ_p can use the formula to deduce that the least amount of risk involved in hitting this target is given by

$$\sigma_p^2 = \frac{C}{D} \left(\mu_p - \frac{A}{C} \right)^2 + \frac{1}{C}. \tag{5.30}$$

If the investor is comfortable with this level of risk then his required portfolio weights are given by (5.28).

We can display the relationship between risk and expected return in (σ, μ)-space. A minor rearrangement of (5.29) shows that the risk–reward coordinates of any optimal portfolio are related by

$$\frac{\sigma^2}{1/C} - \frac{(\mu - A/C)^2}{D/C^2} = 1. \tag{5.31}$$

We recognize that this formula describes a familiar curve studied in high-school geometry; it represents a hyperbola with centre $(0, A/C)$, asymptotes

$$\mu = \pm\sqrt{\frac{D}{C}}\sigma + \frac{A}{C} \tag{5.32}$$

and vertex

$$V = \left(\sqrt{\frac{1}{C}}, \frac{A}{C}\right). \tag{5.33}$$

We note that this simple curve is precisely the boundary of the region of feasible portfolios, it is commonly called the optimal frontier (see Figure 5.3).

It is remarkable that a seemingly complicated problem has such an elegant solution. A portfolio manager need only compute the constants A, B, C and D and, with this information alone, the whole optimal frontier can be plotted and visualized. A potential investor can consult the optimal frontier to find, at a glance, the level of expected return that suits his appetite for risk; he can then use formula (5.28) to determine the required composition of his desired portfolio.

We remark that the problem we have solved assumes that short selling of assets is allowed and is unrestricted. We have already observed, in Chapter 1, that short selling is a high-risk strategy and, for this reason, it is common for short selling to be restricted; indeed, in September 2008, in response to the global credit crisis, the UK and USA imposed a temporary ban on short selling in an attempt to stabilize their markets. In the case where

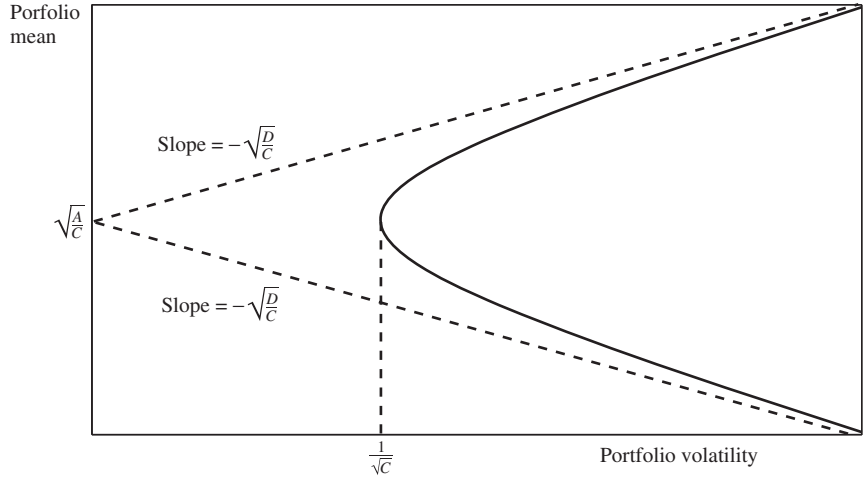


Figure 5.3 The general optimal frontier.

restrictions on short selling are enforced, the corresponding optimal portfolio problem takes the form

$$\begin{aligned} &\text{minimize } \frac{1}{2}\mathbf{w}^T\mathbf{V}\mathbf{w} \quad \text{subject to} \quad \mathbf{P}\mathbf{w} = \mathbf{d} \\ &\qquad\qquad\qquad\qquad \text{and} \quad w_i \geq l_i \text{ for } i = 1, \ldots, n. \end{aligned} \tag{5.34}$$

This new problem can be solved by using an extension of the Lagrange function approach, however its solution, in general, will not have a neat closed form. Instead, the solution can be found by employing an appropriate numerical optimization algorithm; the reader who is interested in discovering how such algorithms are constructed is advised to consult Nash and Sofer (1996) and/or Gill, Murray and Wright (1982).

Essential Mathematics for Market Risk Management
by Simon Hubbert

6 Portfolio Theory II

In this chapter we reveal a collection of remarkable discoveries which arise as a direct consequence of our mathematical solution to the optimal portfolio problem. For instance, we will show that:

- Only two optimal portfolios are needed to provide an entire investment service to mean–variance optimizers!
- For a given optimal portfolio p there is only one other optimal portfolio $z(p)$ say, that satisfies

$$\text{correl}(r_p, r_{z(p)}) = 0.$$

  Furthermore, a simple geometrical argument provides us with an easy way of locating $z(p)$ from p on the efficient frontier.
- When mean–variance optimizers are allowed to borrow and lend at the risk-free rate, the optimal portfolio frontier transforms from the familiar bullet shape to the shape of an arrow head.

6.1 THE TWO-FUND INVESTMENT SERVICE

In solving the optimal portfolio problem we have derived that the optimal weight vector for the least risky portfolio achieving an expected return of μ is given by (5.28). Suppose, for a moment, that we live in a world where only two frontier portfolios are available; p_1 providing an expected return μ_1 and p_2 providing an expected return μ_2. Our analysis tells us that, since these portfolios lie on the frontier, their respective weight vectors are given by

$$\mathbf{w}_1 = \mathbf{g} + \mathbf{h}\mu_1 \quad \text{and} \quad \mathbf{w}_2 = \mathbf{g} + \mathbf{h}\mu_2.$$

On the face of it the opportunity to invest in just two frontier portfolios seems like a harsh constraint. However, if an investor requires the frontier portfolio that provides an expected return μ ($\neq \mu_1$ or μ_2) then he can use the following strategy:

- Find the unique $\alpha \in \mathbb{R}$ such that

$$\mu = \alpha\mu_1 + (1-\alpha)\mu_2.$$

- Combine p_1 and p_2 using the weights $w_1 = \alpha$ and $w_2 = 1-\alpha$ to form a new portfolio, p say.
- The weight vector that defines portfolio p is given by

$$\begin{aligned}\mathbf{w}_p = \alpha\mathbf{w}_1 + (1-\alpha)\mathbf{w}_2 &= \alpha(\mathbf{g} + \mathbf{h}\mu_1) + (1-\alpha)(\mathbf{g} + \mathbf{h}\mu_2) \\ &= \mathbf{g} + \mathbf{h}(\alpha\mu_1 + (1-\alpha)\mu_2) \\ &= \mathbf{g} + \mathbf{h}\mu.\end{aligned}$$

  We recognize this as the optimal weight vector of the frontier portfolio that delivers an expected return μ.

We appear to have made a dramatic discovery; one pair of distinct frontier portfolios is all that is needed to generate all frontier portfolios. Unfortunately, in the real world, it is too much to expect that this result should hold true. The real financial markets are complex and we should be aware that all scientific approaches involve certain simplifying assumptions. In our case we have assumed the following:

- All investors assess the risk and potential reward of a portfolio by its variance and expected return.
- All investors agree on the levels of expected return of individual assets and the values of their pairwise correlation, i.e., they all work with the same expected return vector **e** and covariance matrix **V**.
- Each investment is assumed to be held for a period of τ days.

Clearly these assumptions deviate from true economic reality and so we see why care must be taken when applying theoretical results to real-world problems. Nevertheless, the two fund theorem has influenced the market for financial investments. In particular, many individuals prefer to invest their wealth in mutual funds (companies which offer a professionally managed investment scheme) rather than building their own portfolio from scratch.

6.2 A MATHEMATICAL INVESTIGATION OF THE OPTIMAL FRONTIER

6.2.1 The minimum variance portfolio

If we consider the optimal frontier, Figure 5.3, we see that the least volatile portfolio corresponds to the turning point of the hyperbola. We call this the minimum variance portfolio and denote it by p^*. It provides an expected return of

$$\mu_{p^*} = \frac{A}{C} \tag{6.1}$$

with minimum volatility

$$\sigma_{p^*} = \frac{1}{\sqrt{C}}. \tag{6.2}$$

We note that for every level of volatility $\sigma > \sigma_{p^*}$ there are precisely two frontier portfolios, one positioned on the lower limb of the hyperbola and one on the upper limb. An investor will disregard the lower limb of frontier portfolios as their expected returns are inferior.

Definition 6.1. *Consider an optimal portfolio p with an expected return of μ_p. We use the following terminology:*

- *If $\mu_p > A/C$, we say that p is efficient.*
- *If $\mu_p < A/C$, we say that p is inefficient.*

6.2.2 Covariance of frontier portfolios

Let p and q denote any two frontier portfolios which provide expected returns μ_p and μ_q respectively. Using (5.28), the weight vectors corresponding to these portfolios are

$$\mathbf{w}_p = \mathbf{g} + \mathbf{h}\mu_p \quad \text{and} \quad \mathbf{w}_q = \mathbf{g} + \mathbf{h}\mu_q.$$

The covariance between the random rates of return of p and q can be shown to be

$$\sigma_{pq} = \mathbf{w}_p^T \mathbf{V} \mathbf{w}_q = \frac{C}{D}\left(\mu_p - \frac{A}{C}\right)\left(\mu_q - \frac{A}{C}\right) + \frac{1}{C}. \tag{6.3}$$

The derivation of this expression follows the same steps as the calculation for portfolio variance (5.30). Indeed, the expression for variance is recovered by setting $p = q$ in (6.3).

6.2.3 Correlation with the minimum variance portfolio

The weight vector of the minimum variance portfolio p^* is given by setting $\mu = A/C$ in formula (5.28), i.e., we have

$$\mathbf{w}_{p^*} = \mathbf{g} + \mathbf{h}\mu_{p^*} = \mathbf{g} + \frac{A}{C}\mathbf{h}.$$

An interesting property of this portfolio is that the covariance between its random return rate and the random return rate of any other frontier portfolio is constant. In fact it is equal to $\sigma_{p^*}^2$, the variance of p^*. To see this we simply set $q = p^*$ in (6.3) to give

$$\operatorname{cov}(r_p, r_{p^*}) = \frac{C}{D}\underbrace{\left(\mu_p - \frac{A}{C}\right)\left(\mu_{p^*} - \frac{A}{C}\right)}_{=0 \text{ since } \mu_{p^*} = A/C} + \frac{1}{C} = \frac{1}{C} = \sigma_{p^*}^2. \tag{6.4}$$

In addition, we can deduce that correlation between a frontier portfolio and p^* is positive, in fact it is simply the ratio of their volatilities, i.e.,

$$\operatorname{correl}(r_p, r_{p^*}) = \frac{\operatorname{cov}(r_p, r_{p^*})}{\sigma_p \sigma_{p^*}} = \frac{\sigma_{p^*}}{\sigma_p} \in (0, 1). \tag{6.5}$$

6.2.4 The zero-covariance portfolio

We observe that equation (6.5) reveals that there can be no frontier portfolio that is uncorrelated to p^*. This is not true for an arbitrary frontier portfolio p and, in fact, we will show that there exists a unique frontier portfolio $z(p)$ which is uncorrelated to p, that is

$$\operatorname{cov}(r_p, r_{z(p)}) = 0. \tag{6.6}$$

We call $z(p)$ the zero-covariance portfolio of p.

Using (6.3) we can solve the above equation, that is,

$$\operatorname{cov}(r_p, r_{z(p)}) = \frac{C}{D}\left(\mu_p - \frac{A}{C}\right)\left(\mu_{z(p)} - \frac{A}{C}\right) + \frac{1}{C} = 0,$$

which implies

$$\mu_{z(p)} = \frac{A}{C} - \frac{D/C^2}{\left(\mu_p - A/C\right)}. \tag{6.7}$$

This equation defines $\mu_{z(p)}$ uniquely in terms of μ_p and so the frontier portfolio $z(p)$ is itself unique. We note that if p is an efficient portfolio, i.e., $\mu_p - A/C > 0$, then its zero-covariance pair $z(p)$ is inefficient since

$$\mu_{z(p)} - \frac{A}{C} = -\frac{D/C^2}{\left(\mu_p - A/C\right)} < 0.$$

6.3 A GEOMETRICAL INVESTIGATION OF THE OPTIMAL FRONTIER

A straightforward rearrangement of the frontier hyperbola, equation (5.31), provides the following expression for variance as a function of expected return:

$$\sigma^2 = \frac{C}{D}\left(\mu - \frac{A}{C}\right)^2 + \frac{1}{C}. \tag{6.8}$$

6.3.1 Equation of a tangent to an efficient portfolio

For any efficient portfolio p we want to determine the equation of the unique straight line which is tangent to the portfolio frontier (see Figure 6.1). Mathematically, the equation of the tangent at any frontier point (σ_p, μ_p) is given by

$$\mu = \left.\frac{d\mu}{d\sigma}\right|_{(\sigma_p,\mu_p)} \sigma + \text{intercept}. \tag{6.9}$$

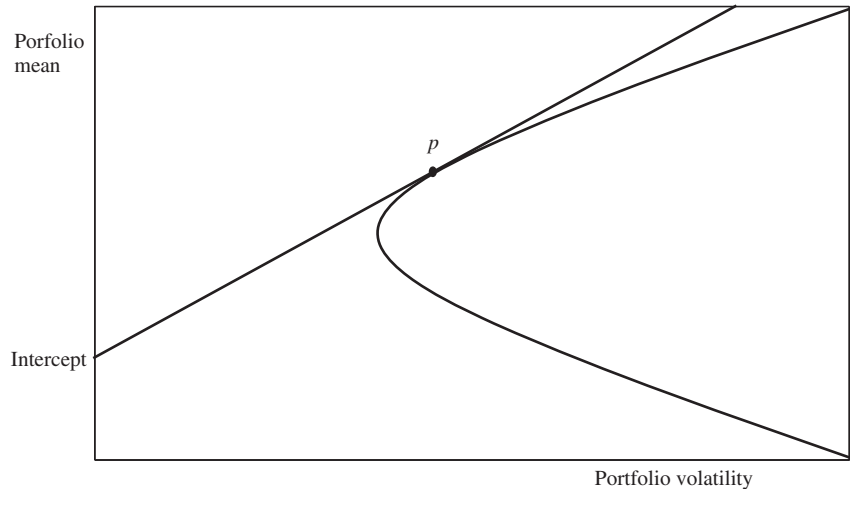


Figure 6.1 Tangent line to an efficient portfolio.

Now, to formally determine the gradient we can differentiate both sides of equation (6.8) with respect to σ as follows:

$$\frac{d}{d\sigma}\sigma^2 = \frac{d\mu}{d\sigma}\cdot\frac{d}{d\mu}\left(\frac{C}{D}\left(\mu - \frac{A}{C}\right)^2 + \frac{1}{C}\right),$$

$$2\sigma = \frac{d\mu}{d\sigma}\left[\frac{2C}{D}\left(\mu - A/C\right)\right],$$

hence

$$\left.\frac{d\mu}{d\sigma}\right|_{(\sigma_p,\mu_p)} = \frac{\sigma_p D}{C(\mu_p - A/C)}. \tag{6.10}$$

We can now determine the μ-intercept of the tangent line by substituting (6.10) and (6.8) into equation (6.9) to give

$$\begin{aligned}
\text{intercept} &= \mu_p - \left.\frac{d\mu}{d\sigma}\right|_{(\sigma_p,\mu_p)}\sigma_p \\
&= \mu_p - \frac{D\sigma_p^2}{C(\mu_p - A/C)} \\
&= \mu_p - \frac{D}{C(\mu_p - A/C)}\left[\frac{C}{D}\left(\mu_p - \frac{A}{C}\right)^2 + \frac{1}{C}\right] \qquad (6.11)\\
&= \mu_p - \left[\mu_p - \frac{A}{C} + \frac{D/C^2}{(\mu_p - A/C)}\right] \\
&= \frac{A}{C} - \frac{D/C^2}{(\mu_p - A/C)}.
\end{aligned}$$

The Gradient Revisited

We have computed the gradient of a tangent line to an efficient portfolio (6.10) by differentiation. An alternative approach would be to use the geometry of the hyperbola, see Figure 6.2, and simply observe that

$$\text{gradient} = \left.\frac{d\mu}{d\sigma}\right|_{(\sigma_p,\mu_p)} = \frac{\mu_p\text{-intercept}}{\sigma_p}. \tag{6.12}$$

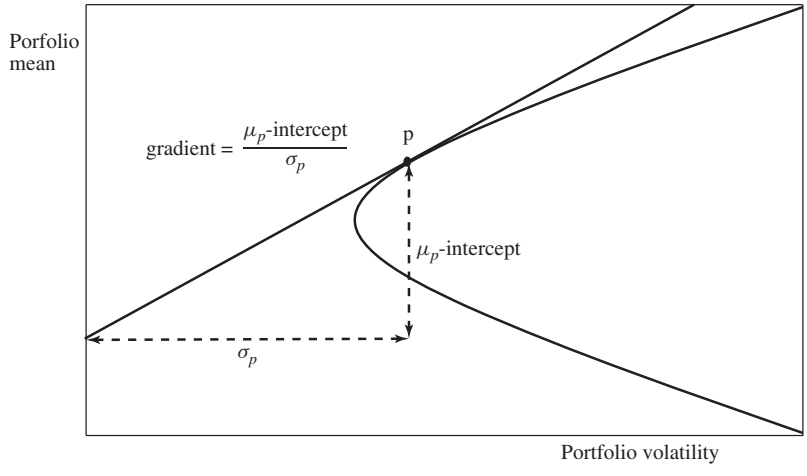


Figure 6.2 The gradient of a tangent line.

We can develop the above equation further by substituting expression (6.11) for the intercept:

$$\begin{aligned}
\text{gradient} &= \frac{\mu_p\text{-intercept}}{\sigma_p} \\
&= \frac{\mu_p - \frac{A}{C} + \frac{D/C^2}{(\mu_p - A/C)}}{\sigma_p} \\
&= \frac{\left(\mu_p - A/C\right)^2 + \frac{D}{C^2}}{\left(\mu_p - A/C\right)\sigma_p} \quad \text{by (3.10) numerator} = \frac{D\sigma_p^2}{C} \\
&= \frac{\sigma_p D}{C(\mu_p - A/C)} \quad \text{and we recover equation (6.10).}
\end{aligned}$$

6.3.2 Locating the zero-covariance portfolio

In the previous section we have shown that for every efficient portfolio p there exists a unique inefficient $z(p)$, the zero-covariance portfolio of p, that satisfies $\text{cov}(r_p, r_{z(p)}) = 0$. We note that the expected return of $z(p)$, given by (6.7), is precisely the intercept of the tangent line to p. This observation links together our mathematical and geometrical investigations and provides us with the following simple way of locating the zero-covariance portfolio $z(p)$ (see Figure 6.3) corresponding to any frontier portfolio p:

- Draw the tangent to the mean–variance hyperbola at the point corresponding to p.
- The μ-intercept of the tangent line represents $\mu_{z(p)}$.
- Draw a horizontal line through this intercept.
- The intersection of this line with the mean–variance hyperbola determines $\sigma_{z(p)}$.

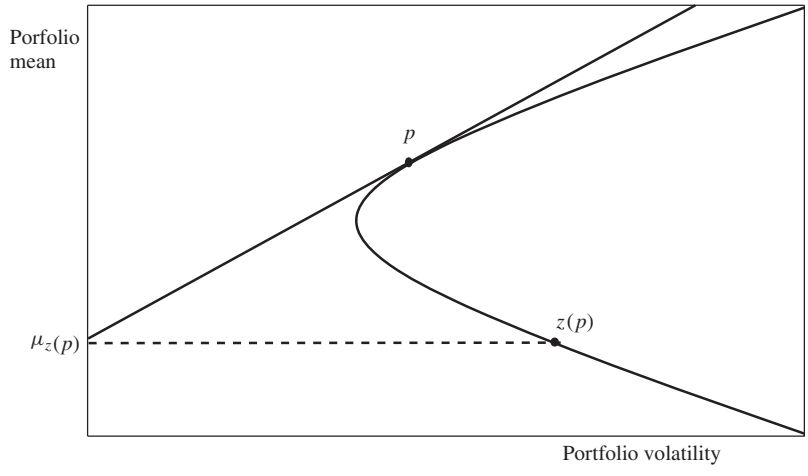


Figure 6.3 Locating the zero-covariance portfolio of an efficient portfolio.

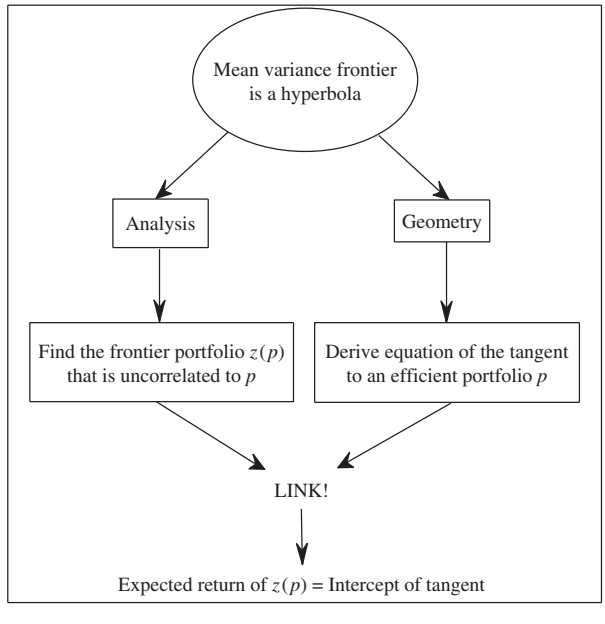


Figure 6.4 How the analytic discoveries are viewed geometrically.

Due to the appealing geometry of our solution we are able to illustrate our theoretical discoveries in (σ, μ)-space. The route is illustrated in the chart displayed in Figure 6.4.

6.4 A FURTHER INVESTIGATION OF COVARIANCE

So far in this chapter we have shown that the covariance between two frontier portfolios can be expressed explicitly via (6.3). We now want to investigate the covariance between

a given efficient frontier portfolio, p say, and any feasible portfolio q. To set the scene we let $\mathbf{w}_q$ denote the weight vector of a feasible portfolio and note that

$$\mathbf{w}_q^T\mathbf{1} = 1 \quad \text{and} \quad \mathbf{w}_q^T\mathbf{e} = \mu_q \quad \text{the expected return on } q.$$

In addition, we let $\mathbf{w}_p$ denote the weight vector of a fixed frontier portfolio. In this case we know from our mean–variance analysis (5.26) that

$$\mathbf{w}_p = \left(\frac{C\mu_p - A}{D}\right)\mathbf{V}^{-1}\mathbf{e} + \left(\frac{B - A\mu_p}{D}\right)\mathbf{V}^{-1}\mathbf{1}.$$

Thus, we can develop an expression for covariance between the two portfolios as follows:

$$\begin{aligned}
\sigma_{pq} &= \mathbf{w}_q^T\mathbf{V}\mathbf{w}_p \\
&= \mathbf{w}_q^T\mathbf{V}\left(\left(\frac{C\mu_p - A}{D}\right)\mathbf{V}^{-1}\mathbf{e} + \left(\frac{B - A\mu_p}{D}\right)\mathbf{V}^{-1}\mathbf{1}\right) \\
&= \frac{C\mu_p - A}{D}\mu_q + \frac{B - A\mu_p}{D} \quad \text{since} \quad \begin{cases} \mathbf{w}_q^T\mathbf{V}\mathbf{V}^{-1}\mathbf{1} = 1; \\ \mathbf{w}_q^T\mathbf{V}\mathbf{V}^{-1}\mathbf{e} = \mu_q. \end{cases}
\end{aligned}$$

Rearranging this formula we can write

$$\mu_q = \frac{A}{C}\left[\frac{\mu_p - B/A}{\mu_p - A/C}\right] + \sigma_{pq}\left[\frac{D/C}{\mu_p - A/C}\right]. \tag{6.13}$$

It turns out that this expression is disguising a much more satisfying relationship, as the following result illustrates.

Lemma 6.2. *Let p denote an efficient portfolio and $z(p)$ denote its zero-covariance pair. The following identities hold:*

$$\frac{A}{C}\left[\frac{\mu_p - B/A}{\mu_p - A/C}\right] = \mu_{z(p)}$$

and

$$\frac{D/C}{\mu_p - A/C} = \frac{\mu_p - \mu_{z(p)}}{\sigma_p^2}.$$

Proof. The following development establishes the first identity:

$$\begin{aligned}
\frac{A}{C}\left[\frac{\mu_p - B/A}{\mu_p - A/C}\right] &= \frac{A}{C}\left[\frac{(\mu_p - A/C) - (B/A - A/C)}{\mu_p - A/C}\right] \\
&= \frac{A}{C} - \frac{A}{C}\left[\frac{\frac{BC - A^2}{AC}}{\mu_p - A/C}\right] \quad (\text{recall } D = BC - A^2) \\
&= \frac{A}{C} - \frac{D/C^2}{\mu_p - \frac{A}{C}} = \mu_{z(p)} \quad (\text{by (6.7)}).
\end{aligned}$$

To establish the second identity we rearrange (6.8) to give the simple (but effective) identity

$$1 = \frac{1}{\sigma_p^2}\left[\frac{1}{C} + \frac{(\mu_p - \frac{A}{C})^2}{D/C}\right].$$

We can use this to establish the result, specifically we write

$$\begin{aligned} 1 \cdot \frac{D/C}{(\mu_p - A/C)} &= \frac{1}{\sigma_p^2} \cdot \left[\frac{1}{C} + \frac{(\mu_p - \frac{A}{C})^2}{D/C}\right] \cdot \frac{D/C}{(\mu_p - A/C)} \\ &= \frac{1}{\sigma_p^2}\left(\frac{D/C^2}{\mu_p - \frac{A}{C}} + \mu_p - \frac{A}{C}\right) \\ &= \frac{1}{\sigma_p^2}\left(\mu_p - \underbrace{\left[\frac{A}{C} - \frac{D/C^2}{\mu_p - \frac{A}{C}}\right]}_{=\mu_{z(p)}}\right) \\ &= \frac{1}{\sigma_p^2}(\mu_p - \mu_{z(p)}). \end{aligned}$$

□

A Surprising Linear Relationship

As a direct result of Lemma 6.2 we can establish that the expected return on any feasible portfolio can be written as a linear combination of expected returns of a fixed frontier portfolio p and of its zero-covariance pair $z(p)$. Specifically, we have

$$\mu_q = \mu_{z(p)} + \frac{\sigma_{pq}}{\sigma_p^2}\left(\mu_p - \mu_{z(p)}\right). \tag{6.14}$$

In particular, if we let

$$\beta_{pq} = \frac{\sigma_{pq}}{\sigma_p^2}, \tag{6.15}$$

then we can rewrite (6.14) as

$$\mu_q = (1 - \beta_{pq})\mu_{z(p)} + \beta_{pq}\mu_p. \tag{6.16}$$

Thus we have discovered that, given any fixed efficient portfolio p, we can express the expected return of a more general feasible portfolio q as a convex combination of μ_p and $\mu_{z(p)}$; i.e., a linear combination whose coefficients ($1 - \beta_{pq}$ and β_{pq} in this case) sum to one.

This relationship is an important one. Indeed, in order to derive the famous CAPM we need only establish that there exists a natural efficient portfolio p to be used in (6.16). Once this special portfolio is revealed, the true CAPM follows immediately. The key step

in reaching this goal is to reconsider our original optimal portfolio problem, only this time we shall include a risk-free asset and allow unlimited lending and borrowing at this rate.

6.5 THE OPTIMAL PORTFOLIO PROBLEM REVISITED

So far we have implicitly assumed that our financial portfolio consists entirely of risky assets. In reality, we will also have the opportunity to borrow and lend cash at the risk-free rate. The inclusion of a risk-free asset in the list of all possible assets is necessary to obtain realism. It is our aim to investigate how the bullet-shaped portfolio frontier changes when we consider risk-free borrowing and lending. To build the risk-free asset into our framework we use the index $i = 0$ and let r_0 denote the known risk-free rate of return available to us. We assume that we can borrow and lend unlimited amounts at this rate and this assumption triggers an important difference to our previous optimal portfolio problem. Specifically, we can now drop the condition that the portfolio weights $w_1, w_2, \ldots, w_n$ must add up to 1. In fact, we now employ the following strategy:

- Choose **any** combination of weights $w_1, w_2, \ldots, w_n$ to invest in the n risky assets.
- If $\sum_{i=1}^{n} w_i < 1$, then we are **below budget** and so we can set

$$w_0 = 1 - \sum_{i=1}^{n} w_i > 0$$

  and lend this proportion of our wealth at the risk-free rate r_0.
- If $\sum_{i=1}^{n} w_i > 1$, then we are **over budget** and so we can set

$$w_0 = 1 - \sum_{i=1}^{n} w_i < 0,$$

  and borrow this proportion of our wealth at the risk-free rate r_0.

We are still faced with the same problem of how to choose the optimal weights for such a portfolio. We let p_+ denote any portfolio with a risk-free component. The random rate of return is now given by

$$\begin{aligned} r_{p_+} &= \underbrace{w_0 r_0}_{\text{non-random}} + \underbrace{w_1 r_1 + w_2 r_2 + \ldots + w_n r_n}_{\text{random}} \\ &= \left(1 - \sum_{i=1}^{n} w_i\right) r_0 + \sum_{i=1}^{n} w_i r_i \\ &= r_0 + \sum_{i=1}^{n} w_i (r_i - r_0). \end{aligned}$$

The expected rate of return is given by

$$\mu_{p_+} = \mathbb{E}[r_0] + \sum_{i=1}^{n} w_i \mathbb{E}[r_i - r_0] = r_0 + \sum_{i=1}^{n} w_i (\mu_i - r_0)$$

or, in vector notation,

$$\mu_{p_+} = r_0 + \mathbf{w}^T(\mathbf{e} - r_0\mathbf{1}).$$

The formula for portfolio variance remains unchanged, i.e., we have

$$\sigma_{p_+}^2 = \mathbf{w}^T\mathbf{V}\mathbf{w}.$$

For a desired expected return rate μ_{p_+} say, the minimum variance portfolio is then the solution to the following problem:

$$\begin{aligned} &\text{minimize } \frac{1}{2}\mathbf{w}^T\mathbf{V}\mathbf{w} \\ &\text{subject to } (\mathbf{e} - r_0\mathbf{1})^T\mathbf{w} = \mu_{p_+} - r_0. \end{aligned} \tag{6.17}$$

This problem is easier than the case with n risky assets because here we only have one linear constraint. The solution technique is precisely the same, we employ the Lagrangian technique and seek the solution $(\mathbf{w}_{p_+}, \lambda^*)^T \in \mathbb{R}^{n+1}$ to the linear system

$$\begin{pmatrix} \mathbf{V} & (\mathbf{e} - r_0\mathbf{1}) \\ (\mathbf{e} - r_0\mathbf{1})^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{w}_{p_+} \\ \lambda^* \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mu_{p_+} - r_0 \end{pmatrix},$$

i.e.,

$$\mathbf{V}\mathbf{w}_{p_+} + \lambda^*(\mathbf{e} - r_0\mathbf{1}) = \mathbf{0} \quad \Rightarrow \quad \mathbf{V}\mathbf{w}_{p_+} = -\lambda^*(\mathbf{e} - r_0\mathbf{1}) \tag{6.18}$$

and

$$(\mathbf{e} - r_0\mathbf{1})^T\mathbf{w}_{p_+} = \mu_{p_+} - r_0. \tag{6.19}$$

We can solve (6.18) to give

$$\mathbf{w}_{p_+} = -\lambda^*\mathbf{V}^{-1}(\mathbf{e} - r_0\mathbf{1}). \tag{6.20}$$

Substituting this expression into the constraint equation (6.19) gives

$$\lambda^* = -\frac{\mu_{p_+} - r_0}{(\mathbf{e} - r_0\mathbf{1})^T\,\mathbf{V}^{-1}\,(\mathbf{e} - r_0\mathbf{1})}. \tag{6.21}$$

We set

$$H = (\mathbf{e} - r_0\mathbf{1})^T\,\mathbf{V}^{-1}\,(\mathbf{e} - r_0\mathbf{1})\,, \tag{6.22}$$

then the optimal vector of portfolio weights can be written neatly as

$$\mathbf{w}_{p_+} = \left(\frac{\mu_{p_+} - r_0}{H}\right)\mathbf{V}^{-1}\,(\mathbf{e} - r_0\mathbf{1})\,. \tag{6.23}$$

We know that $H > 0$ since $\mathbf{V}^{-1}$ is positive definite and, since we also assume the expected returns are not all the same, the vector $\mathbf{e} - r_0\mathbf{1}$ is non-zero.

We can expand the quadratic form that defines H as follows:

$$\begin{aligned} H &= (\mathbf{e} - r_0\mathbf{1})^T \mathbf{V}^{-1} (\mathbf{e} - r_0\mathbf{1}) \\ &= (\mathbf{1}^T\mathbf{V}^{-1}\mathbf{1})r_0^2 - 2(\mathbf{1}^T\mathbf{V}^{-1}\mathbf{e})r_0 + \mathbf{e}^T\mathbf{V}^{-1}\mathbf{e} \\ &= Cr_0^2 - 2Ar_0 + B \\ &= C\left(r_0^2 - 2\frac{A}{C}r_0 + \frac{B}{C}\right) \\ &= C\left[\left(r_0 - \frac{A}{C}\right)^2 + \frac{BC}{C^2} - \frac{A^2}{C^2}\right] \\ &= C\left[\left(r_0 - \frac{A}{C}\right)^2 + \frac{D}{C^2}\right] \\ &= C\left(r_0 - \frac{A}{C}\right)^2 + \frac{D}{C}. \end{aligned} \tag{6.24}$$

We note that the value of H only depends upon the constants A, B, C, D and the prevailing risk-free rate r_0. In particular, its value is independent of the desired level of expected return and thus, just as with the risky-assets-only problem, we can conclude that the weight vector given by

$$\mathbf{w} = \left(\frac{\mu - r_0}{H}\right)\mathbf{V}^{-1}(\mathbf{e} - r_0\mathbf{1})$$

is guaranteed to define a portfolio that will provide an expected return of $100 \times \mu\%$ with minimum risk. In order to calculate the size of this risk we compute the portfolio variance as follows:

$$\begin{aligned} \sigma^2 &= \mathbf{w}^T\mathbf{V}\mathbf{w} \\ &= \left[\left(\frac{\mu - r_0}{H}\right)\mathbf{V}^{-1}(\mathbf{e} - r_0\mathbf{1})\right]^T \mathbf{V}\left(\frac{\mu - r_0}{H}\right)\mathbf{V}^{-1}(\mathbf{e} - r_0\mathbf{1}) \\ &= \left(\frac{\mu - r_0}{H}\right)^2 (\mathbf{e} - r_0\mathbf{1})^T \mathbf{V}^{-1} \underbrace{\mathbf{V}\mathbf{V}^{-1}}_{=I} (\mathbf{e} - r_0\mathbf{1}) \\ &= \left(\frac{\mu - r_0}{H}\right)^2 \underbrace{(\mathbf{e} - r_0\mathbf{1})^T \mathbf{V}^{-1} (\mathbf{e} - r_0\mathbf{1})}_{=H} \\ &= \frac{(\mu - r_0)^2}{H}. \end{aligned}$$

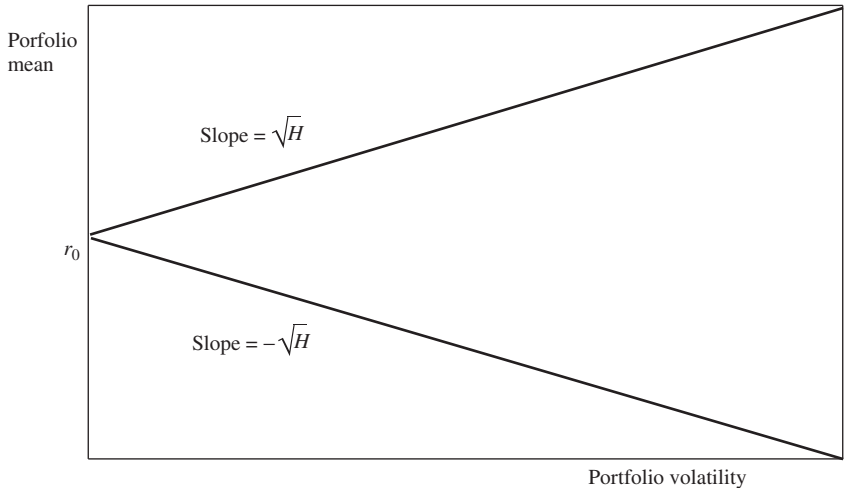


Figure 6.5 Optimal frontier with risk-free asset available.

Taking square roots and rearranging leads to

$$\begin{aligned}\mu &= r_0 \pm \sqrt{H}\sigma \\ &= r_0 \pm \left(\sqrt{C\left(r_0 - \frac{A}{C}\right)^2 + \frac{D}{C}}\right)\sigma \quad \text{(using (6.24))}.\end{aligned} \tag{6.25}$$

Thus, the portfolio frontier this time is much simpler than the hyperbola from before. It is simply composed of two half-lines in (σ, μ)-space, emanating from the point $(0, r_0)$ and with slopes $\sqrt{H}$ and $-\sqrt{H}$ respectively; it is shown in Figure 6.5.

Essential Mathematics for Market Risk Management
by Simon Hubbert

7

The Capital Asset Pricing Model (CAPM)

Markowitz published his mean–variance approach to the optimal portfolio problem in the mid-1950s and, in retrospect, it is now widely accepted that this work has served as the spark that ignited a new scientific approach to risk management. In this chapter we will continue to work in the mean–variance framework and, specifically, we will derive one of the most famous asset pricing models of modern finance; the so-called Capital Asset Pricing Model (CAPM).

7.1 CONNECTING THE PORTFOLIO FRONTIERS

To kick off our derivation of the CAPM we recall that we have already shown (in the previous two chapters) that the shape of the optimal portfolio frontier depends upon whether or not the portfolio includes risk-free borrowing/lending. Specifically, we have the two forms

$$\text{frontier} \Rightarrow \begin{cases} \frac{\sigma^2}{1/C} - \frac{(\mu - A/C)^2}{D/C^2} = 1 & \text{without risk-free rate;} \\ \mu = r_0 \pm \left(\sqrt{C\left(r_0 - \frac{A}{C}\right)^2 + \frac{D}{C}} \right) \sigma & \text{with risk-free rate.} \end{cases} \tag{7.1}$$

We note that both frontiers involve the constants A, B, C (defined by (5.22)) and $D(= BC - A^2)$.

In this section we shall make the assumption that

$$\text{risk-free rate } r_0 < \frac{A}{C} \text{ expected return of min-variance portfolio}$$

We note that this is justified because we work on the principle that investors only take on risk because they are attracted to potentially high rewards and, as such, an investment that is entirely risk free can only be expected to deliver the smallest reward.

We now turn to our next main goal, which is to investigate whether the two frontiers (7.1) are connected. Specifically, we want to determine whether or not the upper ray of the risk-free frontier connects with the upper limb of the risky hyperbola and if so at which point or points, i.e., we want to form the picture that results from superimposing Figure 6.5 onto Figure 5.3. There are three possibilities:

- The upper ray misses the upper limb, i.e., does not coincide with any efficient portfolio.
- The upper ray touches the upper limb at one unique point, i.e., it forms a unique tangent to some efficient portfolio.
- The upper ray crosses the upper limb at two points, i.e., it crosses exactly two efficient portfolios.

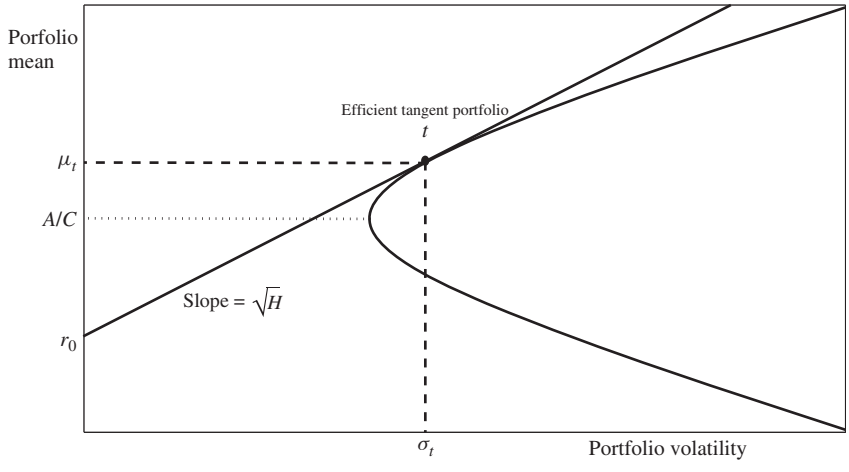


Figure 7.1 The link between our optimal frontiers.

It turns out that the second possibility is true, the upper ray forms a unique tangent to some efficient portfolio t. This result is illustrated in Figure 7.1. We can demonstrate this mathematically by completing two tasks:

1. First, show that there exists an efficient portfolio whose tangent line intercepts the μ axis at r_0. We call this the tangent portfolio and our task is to find its coordinates (σ_t, μ_t).
2. Second, we show that this tangent line is precisely the upper ray of the optimal frontier. Specifically, we use the tangent portfolio's coordinates to establish that

$$\text{gradient of tangent line} = \frac{\mu_t - r_0}{\sigma_t} = \sqrt{H} = \text{gradient of upper ray.}$$

In summary, we have:

Lemma 7.1. *Let $r_0 < A/C$ and let t denote the efficient portfolio of risky assets whose tangent line intercepts the μ axis at r_0. The position of t on the optimal frontier is defined by its coordinates*

$$\mu_t = \frac{A}{C} - \frac{D/C^2}{r_0 - A/C} \tag{7.2}$$

and

$$\sigma_t = \frac{\sqrt{H}}{C\left|r_0 - \frac{A}{C}\right|} = \frac{-\sqrt{H}}{C\left(r_0 - \frac{A}{C}\right)}. \tag{7.3}$$

Furthermore, the gradient of the tangent line through t is $\sqrt{H}$.

Proof. The proof of this result draws on our geometrical investigation of the hyperbola. We recall that the intercept of the tangent to any efficient portfolio is given by equation (6.11).

In our case the intercept is the risk-free rate and thus

$$r_0 = \frac{A}{C} - \frac{D/C^2}{(\mu_t - A/C)}.$$

Rearranging this equation for μ_t gives

$$\mu_t = \frac{A}{C} - \frac{D/C^2}{r_0 - A/C}. \tag{7.4}$$

To identify σ_t we substitute the expression for μ_t into (5.30) to yield

$$\begin{aligned}
\sigma_t^2 &= \frac{D}{C^3}\left[\frac{1}{(r_0 - A/C)^2}\right] + \frac{1}{C} \quad \text{(using (7.4))} \\
&= \frac{1}{C}\left[\frac{D/C}{C\,(r_0 - A/C)^2} + 1\right] \\
&= \frac{1}{C}\left[\frac{D/C + C\,(r_0 - A/C)^2}{C\,(r_0 - A/C)^2}\right] \quad \text{numerator equals } H \text{ by (6.24)} \\
&= \frac{H}{C^2\left(r_0 - \frac{A}{C}\right)^2}.
\end{aligned}$$

Taking square roots, and accounting for the fact that $r_0 - \frac{A}{C} < 0$, we deduce that

$$\sigma_t = \frac{\sqrt{H}}{C\left|r_0 - \frac{A}{C}\right|} = \frac{-\sqrt{H}}{C\left(r_0 - \frac{A}{C}\right)}.$$

To calculate the gradient we evaluate $(\mu_t - r_0)/\sigma_t$. We focus on the numerator, where we find that

$$\begin{aligned}
\mu_t - r_0 &= \frac{A}{C} - \frac{D/C^2}{r_0 - A/C} - r_0 \\
&= -\left(\frac{D/C}{C(r_0 - A/C)} + r_0 - \frac{A}{C}\right) \\
&= -\left(\frac{D/C + C(r_0 - A/C)^2}{C(r_0 - A/C)}\right) \quad \text{numerator equals } H \text{ by (6.24)} \\
&= \frac{-H}{C(r_0 - A/C)}.
\end{aligned}$$

We can use this in conjunction with the expression for σ_t to deduce that

$$\left.\frac{d\mu}{d\sigma}\right|_{(\sigma_t, \mu_t)} = \frac{\mu_t - r_0}{\sigma_t} = \frac{\frac{-H}{C(r_0 - A/C)}}{\frac{-\sqrt{H}}{C(r_0 - A/C).}} = \sqrt{H}.$$

□

The One-Fund Investment Service

Let us distinguish between two types of investor by considering their approaches to portfolio construction. Specifically, we shall define:

- Type A investors are those who wish to build a portfolio by combining risky assets with risk-free borrowing or lending.
- Type B investors are those who decide to build a portfolio of risky assets only.

Type A investors can use our analysis to deduce that the most efficient (i.e., least risky) portfolios that can be constructed are represented, in (σ, μ)-space, by an upwardly sloping straight line that emanates from the risk-free return rate (see Figure 6.5). All type A investors will seek a portfolio on this line. In contrast, we know from Chapter 5 that all type B investors will seek a portfolio which lies on the efficient limb of the optimal hyperbola.

We have shown that the efficient frontiers of type A and type B investors meet at precisely one point; we call this the tangent portfolio and denote it by t. This discovery is particularly important to type A investors; their frontier is a straight line and since only two points are needed to define a line, we can deduce that every efficient type A portfolio is simply a mixture of the tangent portfolio and the risk-free asset:

$$\text{efficient type } A \text{ portfolio} = \alpha r_0 + (1 - \alpha)t \quad \text{for some } \alpha \in \mathbb{R}.$$

This discovery reveals the remarkable fact that only one portfolio, the tangent portfolio, is needed to provide a full investment service to all type A investors. We recall that in Chapter 5, we discovered an equivalent result for type B investors; where any two optimal portfolios of risky assets can generate the entire type B frontier. In view of this discovery, a tantalizing puzzle is to determine precisely the nature of the tangent portfolio.

7.2 THE TANGENT PORTFOLIO

Let us assume we live in a world where the one-fund investment service operates. In this case, following our arguments above, every type A investor would demand a portfolio with the right balance of risk-free rate r_0 and the tangent portfolio t. In particular, the tangent portfolio would be the only risky portfolio in demand. It is this fact that allows us to pin down exactly what the tangent portfolio represents. We work on the principal that the financial markets are in equilibrium, where supply meets demand. Thus, in order to find the in-demand tangent portfolio, we look to the supply side of the risk market.

7.2.1 The market's supply of risky assets

Suppose the entire market consists of n different risky assets, and that:

- Asset i has value S_i.
- There are a total of α_i available shares of asset i (to buy) for $(i = 1, \ldots, n)$.

Based on this we have the total value of asset i is given by $\alpha_i S_i$ and therefore,

$$\text{total value of the market} = \sum_{i=1}^{n} \alpha_i S_i.$$

The market capitalization of each asset is just the proportion it contributes to the risky market as a whole, that is

$$w_i^{(m)} = \frac{\alpha_i S_i}{\sum_{i=1}^n \alpha_i S_i} \quad i = 1, 2, \ldots, n.$$

Clearly the capitalization weights $\{w_i^{(m)} : i = 1, \ldots, n\}$ sum to one, and so the portfolio whose random return rate is given by

$$r_m = \sum_{i=1}^n w_i^{(m)} r_i$$

is feasible. We call this the market portfolio since it represents the supply side of the total risky market. Now, under equilibrium, we have

$$\text{supply} = \text{demand},$$

and so we conclude that

$$\text{the tangent portfolio } t = \text{the market portfolio } m.$$

With this fixed we introduce a new term, we call the upper ray of the frontier the *Capital Market Line* (CML) to signify that it cuts through the market portfolio (see Figure 7.2).

7.3 THE CAPM

We are now perfectly placed to derive the famous Capital Asset Pricing Model. The derivation is quick and the crucial starting point is our revelation (6.16) that, by fixing an efficient

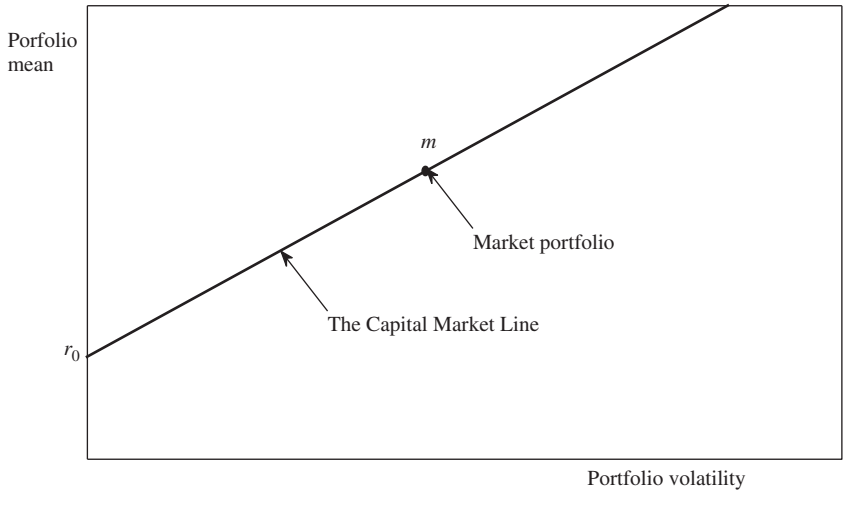


Figure 7.2 The (efficient) capital market line.

portfolio p, we can express the expected return of any feasible portfolio q as

$$\mu_q = (1 - \beta_{pq})\mu_{z(p)} + \beta_{pq}\mu_p, \quad \text{where } \beta_{pq} = \frac{\sigma_{pq}}{\sigma_p^2}.$$

Furthermore, we recall that $\mu_{z(p)}$ (the expected return of the zero-covariance portfolio of p) is precisely the point where the tangent line to p intercepts the μ axis, see Figure 6.4.

In the previous section we have shown that the tangent portfolio has the special property that it can provide a full investment service and so is the only risky portfolio in demand. We have also argued that, under market equilibrium, the tangent portfolio coincides with the market portfolio. In view of this we decide to specialize the above linear relationship by setting $p = m$, the market portfolio, to give

$$\begin{aligned}\mu_q &= (1 - \beta_q)\mu_{z(m)} + \beta_m\mu_m \\ &= (1 - \beta_q)r_0 + \beta_q\mu_m \quad (\text{since } r_0 = \mu_{z(m)} = \text{intercept of the CML}),\end{aligned}$$

where

$$\beta_q = \frac{\sigma_{mq}}{\sigma_m^2} \quad \text{is the beta of } q.$$

This is the portfolio version of CAPM and, after a simple rearrangement it states that, when the market is in equilibrium, the expected excess return of q (in excess of the risk-free rate) is directly proportional to the equivalent expected excess return of the market portfolio, i.e.,

$$\mu_q - r_0 = \beta_q(\mu_m - r_0),$$

where the constant of proportionality is the beta of portfolio q.

We remark that each individual risky asset can be viewed as a feasible portfolio; the feasible vector $\mathbf{w} = (w_1, \ldots, w_n)^T$ whose components are given by

$$w_j = \begin{cases} 0 & \text{if } j \neq i \\ 1 & \text{if } j = i \end{cases}$$

represents the ith risky asset. Thus there is also an asset version of CAPM and it states:

$$\mu_i = r_0 + \beta_i(\mu_m - r_0) \quad \text{where} \quad \beta_i = \frac{\text{cov}(r_i, r_p)}{\sigma_m^2}. \tag{7.5}$$

The quantity β_i denotes the beta of asset i. This proves to be a useful risk measure (as we shall see) since it allows us to gain a feel for how much each individual asset contributes to the overall risk of the portfolio.

7.4 APPLICATIONS OF CAPM

One of the most challenging puzzles of modern finance is to develop accurate mathematical models for the evolution of asset returns. The CAPM provides a simple framework where we can begin to address this challenge. If we let r_m denote the τ-day random return rate of the market portfolio, then the CAPM tells us that the expected return rate of the ith asset

is given by

$$\begin{aligned}\mathbb{E}[r_i] = \mu_i &= r_0 + \beta_i(\mu_m - r_0) \\ &= \mathbb{E}[r_0 + \beta_i(r_m - r_0)].\end{aligned}$$

We see that the beta of the asset determines the extent to which its return is influenced by the expectation of the market. We can conclude that the actual randomness of asset returns can, in part, be explained by fluctuations in the market and as such we postulate a model of the form

$$r_i = r_0 + \beta_i(r_m - r_0) + \varepsilon_i \quad \text{for } i = 1, \ldots, n,$$

where ε_i is a random error term for each asset. We note that the CAPM implies that each ε_i has zero mean. We can gain further insight into the error term by considering its covariance with the market; we have

$$\begin{aligned}\text{cov}(\varepsilon_i, r_m) &= \text{cov}\,(r_i - r_0 - \beta_i(r_m - r_0), r_m) \\ &= \text{cov}(r_i, r_m) - \beta_i \text{cov}(r_m, r_m) \\ &= \text{cov}(r_i, r_m) - \beta_i \sigma_m^2 = 0 \quad \text{(by definition of } \beta_i).\end{aligned}$$

Thus, the market and the error terms are uncorrelated and so we can view ε_i as the component of the asset's return that cannot be explained by the market. We now make the additional assumption that the random components $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are uncorrelated, and we construct our first framework for asset returns.

$$\begin{aligned}\text{The CAPM-based model} &: r_i = r_0 + \beta_i(r_m - r_0) + \varepsilon_i (1 \le i \le n). \\ \text{Consequence 1} &: \mathbb{E}[\varepsilon_i] = 0 (1 \le i \le n). \\ \text{Consequence 2} &: \text{cov}(\varepsilon_i, r_m) = \mathbb{E}[\varepsilon_i r_m] = 0 (1 \le i \le n). \\ \text{Additional assumption} &: \text{cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}[\varepsilon_i \varepsilon_j] = 0 \quad (i \ne j).\end{aligned} \tag{7.6}$$

7.4.1 Decomposing risk

The CAPM-based model we have proposed in (7.6) holds some very appealing properties. In particular, it can be used to decompose the risk of an asset or any feasible portfolio into two components. To reveal this we consider the following general covariance calculation:

$$\begin{aligned}\text{cov}(r_i, r_j) &= \text{cov}(r_i, r_0 + \beta_j(r_m - r_0) + \varepsilon_j) \\ &= \text{cov}(r_i, \beta_j r_m + \varepsilon_j) \quad \text{(since cov (random, certain)} = 0) \\ &= \beta_j \text{cov}(r_i, r_m) + \text{cov}(r_i, \varepsilon_j) \\ &= \beta_j \text{cov}(r_i, r_m) + \text{cov}(r_0 + \beta_i(r_m - r_0) + \varepsilon_i, \varepsilon_j) \\ &= \beta_j \text{cov}(r_i, r_m) + \text{cov}(\beta_i r_m + \varepsilon_i, \varepsilon_j) \\ &= \beta_j \underbrace{\text{cov}(r_i, r_m)}_{=\beta_i \sigma_m^2} + \beta_i \underbrace{\text{cov}(r_m, \varepsilon_j)}_{=0} + \underbrace{\text{cov}(\varepsilon_i, \varepsilon_j)}_{=0 \text{ for } i \ne j}.\end{aligned}$$

Thus we can write

$$\operatorname{cov}(r_i, r_j) = \begin{cases} \beta_i^2 \sigma_m^2 + \operatorname{var}(\varepsilon_i) & i = j; \\ \beta_i \beta_j \sigma_m^2 & i \neq j. \end{cases} \tag{7.7}$$

One of the first things we note from (7.7) is that the riskiness of each individual asset can be decomposed into two distinct components:

- A systematic component which is locked into the market:

$$\text{systematic risk of asset } i := \beta_i^2 \sigma_m^2.$$

- A specific component which is connected to the individual asset itself:

$$\text{specific risk of asset } i := \operatorname{var}(\varepsilon_i).$$

Let us suppose we take any feasible portfolio q whose random return rate is given by

$$r_q = \sum_{i=1}^{n} w_i r_i.$$

According to the portfolio version of CAPM, the expected return of q is given by

$$\mu_q = r_0 + \beta_q (\mu_m - r_0)$$

where the beta of q is given by

$$\begin{aligned} \beta_q = \frac{\operatorname{cov}(r_q, r_m)}{\sigma_m^2} &= \frac{\operatorname{cov}\left(\sum_{i=1}^{n} w_i r_i, r_m\right)}{\sigma_m^2} \\ &= \sum_{i=1}^{n} w_i \frac{\operatorname{cov}(r_i, r_m)}{\sigma_m^2} \\ &= \sum_{i=1}^{n} w_i \beta_i. \end{aligned}$$

We can use the model to discover that the variance of q is given by

$$\begin{aligned} \sigma_q^2 &= \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \operatorname{cov}(r_i, r_j) \\ &= \sigma_m^2 \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \beta_i \beta_j + \sum_{i=1}^{n} w_i^2 \operatorname{var}(\varepsilon_i) \\ &= \sigma_m^2 \left(\sum_{i=1}^{n} w_i \beta_i \right) \left(\sum_{j=1}^{n} w_j \beta_j \right) + \sum_{i=1}^{n} w_i^2 \operatorname{var}(\varepsilon_i) \\ &= \beta_q^2 \sigma_m^2 + \sum_{i=1}^{n} w_i^2 \operatorname{var}(\varepsilon_i). \end{aligned}$$

Thus the risk decomposition for assets has a generalization to portfolios, i.e., the variance of any feasible portfolio can be decomposed into two components:

- A systematic component which is locked into the market:

$$\text{systematic risk of portfolio } q := \beta_q^2 \sigma_m^2.$$

- A specific component which is the square-weighted sum of the specific risks of the individual assets:

$$\text{specific risk of portfolio } q := \sum_{i=1}^{n} w_i^2 \text{var}(\varepsilon_i).$$

Clearly the systematic risk is unavoidable, it is ever-present for all feasible portfolios and its magnitude depends upon the portfolio beta. The specific risk, on the other hand, is dependent upon the portfolio weights and therefore two different portfolios can possess different amounts of specific risk. This poses the following question:

How Much of the Specific Risk can be Reduced?

We would imagine that the best candidate for a portfolio with a small specific risk component would be a frontier portfolio. We shall investigate whether this is the case. We know any efficient portfolio p, by definition, lies on the capital market line, that is its risk–reward coordinates satisfy

$$\sigma_p^2 = \frac{(\mu_p - r_0)^2}{H}. \tag{7.8}$$

The portfolio version of CAPM for p tells us

$$\mu_p - r_0 = \beta_p(\mu_m - r_0). \tag{7.9}$$

Substituting (7.9) into (7.8) gives

$$\sigma_p^2 = \beta_p^2 \left[\frac{(\mu_m - r_0)^2}{H} \right]. \tag{7.10}$$

The market portfolio m is also efficient and so, by (7.8), the term in square brackets is precisely σ_m^2. Thus, we have

$$\sigma_p^2 = \beta_p^2 \sigma_m^2 \text{ for all efficient portfolios } p. \tag{7.11}$$

This tells us that an efficient portfolio (combination of risk-free and market portfolio) on the capital market line has no specific risk. This discovery should provide us with a warning; since it is always possible to diversify away all specific risk, we should not expect any extra reward should we take it on.

Essential Mathematics for Market Risk Management
by Simon Hubbert

8 Risk Factor Modelling

On any journey of discovery it is always rewarding to look back at certain points and admire the achievements, often this process determines how we plot future routes. The same principle applies to our mathematical journey where, so far, we have discovered that a careful analysis of the Markowitz framework (coupled with financial intuition) leads to the famous CAPM and, as a result, this allows us to propose our first model for asset returns; see (7.6). In this chapter we are going to take a much more cavalier approach to modelling asset returns. We will use the CAPM as our inspiration, however, rather than accept that the market portfolio is the key driver of asset returns; as in (7.6), we shall explore the possibility that another more relevant factor (or factors) may be more appropriate. In short, we consider more general linear factor models for asset returns. The reason this approach is described as cavalier is because, unlike CAPM, these models cannot be derived from first principles through the Markowitz framework. However, despite this, these models have proven to be extremely popular in practice and are widely used in the hedge fund industry.

8.1 GENERAL FACTOR MODELLING

We motivate our new investigation by posing the following question:

Do we really expect the market portfolio
to be the driver of all our assets?

Clearly there is good reason to investigate this question, for instance we can consider the following scenarios:

- If our portfolio is composed of random assets taken from the commodities sector of the market, then perhaps a more likely driver would be one of the many exchange-traded commodities indices that are designed to track prices in this sector.
- Similarly, if our portfolio is composed of random assets taken from several sectors of the market, then there may be several driving factors which influence the evolution of the random returns.

To take account of these instincts we propose an alternative approach to modelling. The idea is to generalize the CAPM-based model by allowing more freedom in the choice of driving factors. We consider two possibilities.

1. The single-factor case:
$$r_i = a_i + b_i f + \varepsilon_i \quad (1 \leq i \leq n).$$
Here all assets are driven by a random factor f, this is typically an observable index (or some carefully constructed portfolio) that represents some macro-economic variable. We notice that if $f = r_m$ then we recover the CAPM-based model.

2. The multi-factor case:

$$r_i = a_i + b_{i1} f_1 + b_{i2} f_2 + \cdots + b_{il} f_l + \varepsilon_i \quad i = 1, \ldots, n.$$

Here the assets are driven by l different factors, typically l is small in comparison to n, the number of random returns. The unknown parameters appearing in the model are estimated from a regression run; see Section 3.3.

In both cases there exists, for each asset, a random noise component ε_i. To complete the specification we propose the following CAPM-style assumptions:

$$\begin{aligned} \text{zero mean } \mathbb{E}[\varepsilon_i] &= 0 \quad (1 \le i \le n); \\ \text{no correlation with factors } \mathbb{E}[\varepsilon_i f_j] &= 0 \quad (1 \le i \le n) \quad \text{and} \quad (1 \le j \le l); \\ \text{no inter-correlation } \mathbb{E}[\varepsilon_i \varepsilon_j] &= 0 \quad (i \ne j). \end{aligned} \tag{8.1}$$

8.2 THEORETICAL PROPERTIES OF THE FACTOR MODEL

At the heart of all our investment problems is a set of n risky assets $\{r_1, \ldots, r_n\}$. A general l-factor model for these returns is given by

$$r_i = a_i + \sum_{j=1}^{l} b_{ij} f_j + \varepsilon_i \qquad \text{for} \qquad i = 1, \ldots, n,$$

where the error terms ε_i $(1 \le i \le n)$ and driving factors f_j $(1 \le j \le l)$ satisfy (8.1). In matrix–vector form we can write

$$\begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_n \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1l} \\ b_{21} & b_{22} & \cdots & b_{2l} \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nl} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_l \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \vdots \\ \varepsilon_n \end{pmatrix}, \tag{8.2}$$

that is

$$\mathbf{r} = \mathbf{a} + \mathbf{B}\mathbf{f} + \boldsymbol{\varepsilon}.$$

Under this model we can write the vector of expected returns as

$$\begin{aligned} \mathbf{e} = \mathbb{E}[\mathbf{r}] &= \mathbb{E}[\mathbf{a} + \mathbf{B}\mathbf{f} + \boldsymbol{\varepsilon}] \\ &= \mathbf{a} + \mathbf{B}\mathbf{e}_f, \end{aligned}$$

where

$$\mathbf{e}_f = \mathbb{E}[\mathbf{f}] = \begin{pmatrix} \mu_{f_1} \\ \mu_{f_2} \\ \vdots \\ \mu_{f_l} \end{pmatrix} \in \mathbb{R}^l.$$

Furthermore, we can use (3.16) to write the covariance matrix of the returns as

$$\begin{aligned}\mathbf{V} &= \mathbb{E}[(\mathbf{r}-\mathbf{e})(\mathbf{r}-\mathbf{e})^T] \\ &= \mathbb{E}\Big[\big(\mathbf{B}(\mathbf{f}-\mathbf{e}_f)+\boldsymbol{\varepsilon}\big)\big(\mathbf{B}(\mathbf{f}-\mathbf{e}_f)+\boldsymbol{\varepsilon}\big)^T\Big] \\ &= \mathbb{E}\Big[\mathbf{B}(\mathbf{f}-\mathbf{e}_f)(\mathbf{f}-\mathbf{e}_f)^T\mathbf{B}^T\Big]+\mathbb{E}\Big[\mathbf{B}(\mathbf{f}-\mathbf{e}_f)\boldsymbol{\varepsilon}^T\Big]+\mathbb{E}\Big[\boldsymbol{\varepsilon}^T(\mathbf{f}-\mathbf{e}_f)^T B^T\Big]+\mathbb{E}\Big[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\Big] \\ &= \mathbf{B}\mathbb{E}\Big[(\mathbf{f}-\mathbf{e}_f)(\mathbf{f}-\mathbf{e}_f)^T\Big]\mathbf{B}^T+\mathbf{B}\mathbb{E}\Big[(\mathbf{f}-\mathbf{e}_f)\boldsymbol{\varepsilon}^T\Big]+\mathbb{E}\Big[\boldsymbol{\varepsilon}(\mathbf{f}-\mathbf{e}_f)^T\Big]\mathbf{B}^T+\mathbb{E}\Big[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\Big].\end{aligned}$$

Interpreting the elements on the right-hand side:

- We notice that

$$\mathbb{E}\Big[(\mathbf{f}-\mathbf{e}_f)(\mathbf{f}-\mathbf{e}_f)^T\Big] \tag{8.3}$$

  is, by definition, the covariance matrix $\boldsymbol{\Omega}_f \in \mathbb{R}^{l\times l}$ of the l driving factors $\{f_1,\ldots,f_l\}$.
- The outer product $(\mathbf{f}-\mathbf{e}_f)\boldsymbol{\varepsilon}^T$ is the $l\times n$ matrix

$$\begin{pmatrix} (f_1-\mu_{f_1})\varepsilon_1 & (f_1-\mu_{f_1})\varepsilon_2 & \ldots & \ldots & (f_1-\mu_{f_1})\varepsilon_n \\ (f_2-\mu_{f_2})\varepsilon_1 & (f_2-\mu_{f_2})\varepsilon_2 & \ldots & \ldots & (f_2-\mu_{f_2})\varepsilon_n \\ \vdots & \vdots & & & \vdots \\ (f_l-\mu_{f_l})\varepsilon_1 & (f_l-\mu_{f_l})\varepsilon_2 & \ldots & \ldots & (f_l-\mu_{f_l})\varepsilon_n \end{pmatrix}.$$

  Taking the expectation leads to the zero matrix, due to the assumptions (8.1) we have placed on the noise components ε_i $(1\le i\le n)$. The outer product $(\mathbf{f}-\mathbf{e}_f)^T\boldsymbol{\varepsilon}$ is just the transpose of the above matrix.
- The outer product $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T$ is the $n\times n$ matrix

$$\begin{pmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \ldots & \varepsilon_1\varepsilon_n \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \ldots & \varepsilon_2\varepsilon_n \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \ldots & \varepsilon_n^2 \end{pmatrix}.$$

  We have assumed that the random noise components are uncorrelated, thus taking the expectation leads to the diagonal matrix

$$\boldsymbol{\Lambda} = \begin{pmatrix} \text{var }(\varepsilon_1) & 0 & \ldots & 0 \\ 0 & \text{var }(\varepsilon_2) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \text{var }(\varepsilon_n) \end{pmatrix}.$$

  With this we can show that the covariance matrix has a nice representation

$$\mathbf{V} = \mathbf{B}\boldsymbol{\Omega}_f\mathbf{B}^T + \boldsymbol{\Lambda}.$$

Suppose we build a portfolio of the assets using a weight vector $\mathbf{w} \in \mathbb{R}^n$, then the associated mean and variance are given by

$$\mu_p = \mathbf{w}^T \mathbf{e} = \mathbf{w}^T \mathbf{a} + \mathbf{w}^T \mathbf{B} \mathbf{e}_f \tag{8.4}$$

and

$$\begin{aligned} \sigma_p^2 = \mathbf{w}^T \mathbf{V} \mathbf{w} &= \mathbf{w}^T \mathbf{B} \boldsymbol{\Omega}_f \mathbf{B}^T \mathbf{w} + \mathbf{w}^T \boldsymbol{\Lambda} \mathbf{w} \\ &= \mathbf{w}^T \mathbf{B} \boldsymbol{\Omega}_f \mathbf{B}^T \mathbf{w} + \sum_{i=1}^{n} w_i^2 \text{var } (\varepsilon_i). \end{aligned} \tag{8.5}$$

We note that we have a situation very similar to the CAPM case. Under the factor model we are able to decompose the portfolio risk into two components:

- A systematic component which is locked into the market:

$$\text{systematic risk of the portfolio} := \mathbf{w}^T \mathbf{B} \boldsymbol{\Omega}_f \mathbf{B}^T \mathbf{w}.$$

- A specific component which is the square-weighted sum of the specific risks of the individual assets:

$$\text{specific risk of the portfolio} := \sum_{i=1}^{n} w_i^2 \text{var}(\varepsilon_i) = \mathbf{w}^T \boldsymbol{\Lambda} \mathbf{w}.$$

We note that in the single-factor case the portfolio risk (8.5) collapses to

$$\sigma_p^2 = \left(\sum_{i=1}^{n} w_i b_i \right) \sigma_f^2 + \sum_{i=1}^{n} w_i^2 \text{var}(\varepsilon_i). \tag{8.6}$$

It appears that factor modelling has many similarities with the CAPM model. Let us compare and contrast the two different approaches.

The CAPM World

- Using the insight of mean–variance optimization, the CAPM equation can be derived:

$$\mu_i = r_0 + \beta_i (\mu_m - r_0) (1 \leq i \leq n).$$

- Based on the CAPM equation, the following model is proposed for asset returns:

$$r_i = r_0 + \beta_i (r_m - r_0) + \varepsilon_i (1 \leq i \leq n).$$

- Under this model the portfolio risk can be decomposed into:
  (a) Systematic risk, which will always have to be borne by the investor.
  (b) Specific risk, which can be made to disappear by choosing a frontier portfolio.

The Multi-factor World

- In an attempt to generalize the CAPM view of asset returns we propose the model

$$r_i = a_i + \sum_{j=1}^{l} b_{ij} f_j + \varepsilon_i \qquad \text{for} \qquad i = 1, \ldots, n.$$

- Under this model the portfolio risk can also be decomposed into:

  (a) Systematic risk, which will always have to be borne by the investor.
  (b) Specific risk, which can be reduced through diversification. For instance, if we assume that

$$\sigma^2_{\max} = \max\{\text{var}(\varepsilon_i) : i = 1, \ldots, n\} < \infty$$

  then with the equally weighted portfolio we have

$$\text{specific risk} = \frac{1}{n^2}\sum_{i=1}^{n} \text{var}(\varepsilon_i) \leq \frac{\sigma^2_{\max}}{n} \to 0 \text{ as } n \to \infty.$$

  Thus, the specific risk can be eliminated if we make the theoretical assumption that there are infinitely many risky assets. In practice, there are sufficiently many assets and we can view the reduction through diversification as elimination.

The practicing risk manager who is responsible for monitoring the performance of a large risky portfolio is likely to employ a multi-factor model to reduce the dimension of the problem and to decompose the portfolio risk. A careful selection of relevant driving factors is crucial if the model is to faithfully capture the key properties of the portfolio. From a practical point of view one would expect that the driving factors should represent economic indicators (e.g., industrial production, market indices, oil prices, inflation and interest rate expectations, etc.) that have a plausible impact on future asset prices. This approach is called the fundamental method and its success rests on the shoulders of the practitioner; it is often said that selecting driving factors via the fundamental method is an art rather than a science. The process requires a high level of familiarity with a given market sector, for instance, when modelling the return rate of a stock price, the practitioner should possess (or be able to build) a significant knowledge base of the firm. A further requirement is the ability to access a significant range of concrete economic factors that could potentially drive asset returns. If this is the case then the practitioner's task is as follows:

- Sort through a collection of observable economic factors and rank in order of relevance.
- Decide upon the number of factors to be included in the model for asset return.
- Collect time series data for asset returns and driving factors and run a set of regressions, one for each asset, to derive estimates for the model parameters.

8.3 MODELS BASED ON PRINCIPAL COMPONENT ANALYSIS (PCA)

In some cases it will not be obvious how to build a concrete factor model. The knowledge base regarding the specific assets may not be available or there may be insufficient data to produce reliable parameter estimates. Indeed, we can also argue that the most influential drivers of the return on an asset may not be directly observable. In this situation a more scientific investigation is called for and we present here a popular approach which uses principal component analysis; a data rotation/dimension-reduction tool that is commonly used in applied statistics.

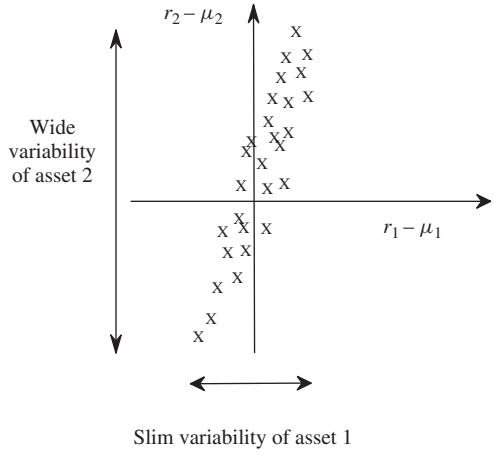


Figure 8.1 Plot of the variation of the random returns r_1 and r_2 about their expected values.

8.3.1 PCA in two dimensions

We motivate the method of principal component analysis (PCA) with a simple two-dimensional example where we are able to visualize the steps. Let us consider two random assets with return rates r_1 and r_2 and expected returns μ_1 and μ_2 respectively. We appeal to the historical time series of both assets and plot the past realizations $(r_1 - \mu_1, r_2 - \mu_2)$ (see Figure 8.1). The plot gives us a nice visual interpretation of how the data varies and it indicates that there is far more variation in asset 2 compared with asset 1.

We note that the reference for our 2-dimensional plot is the familiar one; each point is represented as a certain distance along the horizontal axis and a certain distance along a vertical axis, that is

$$\begin{pmatrix} r_1 - \mu_1 \\ r_2 - \mu_2 \end{pmatrix} = (r_1 - \mu_1)\begin{pmatrix} 1 \\ 0 \end{pmatrix} + (r_2 - \mu_2)\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$
$$= (r_1 - \mu_1)\mathbf{e}_1 + (r_2 - \mu_2)\mathbf{e}_2.$$

Mathematically, we express the vector as a linear combination of the two standard unit vectors $\mathbf{e}_1$ (pointing horizontally) and $\mathbf{e}_2$ (pointing vertically).

We can, of course, express our 2-dimensional data as a linear combination of any pair of orthogonal unit vectors. A typical 2-dimensional unit vector is represented by

$$\mathbf{u} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \quad \text{where } \theta \in [0, 2\pi).$$

If we fix the angle θ then there are two possible unit vectors that are orthogonal to $\mathbf{u}$:

$$\begin{pmatrix} \cos(\theta + \pi/2) \\ \sin(\theta + \pi/2) \end{pmatrix} = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \cos(\theta - \pi/2) \\ \sin(\theta - \pi/2) \end{pmatrix} = \begin{pmatrix} \sin\theta \\ -\cos\theta \end{pmatrix}.$$

The idea is to choose a pair of vectors such that one of the directions captures the dominant component of the data. For example, if we express the data from the plot in Figure 8.1 with

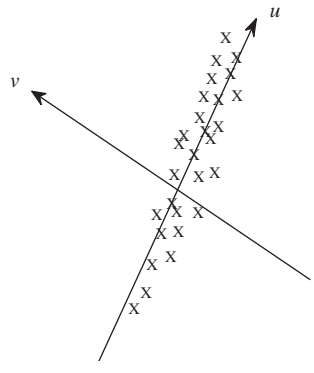


Figure 8.2 Observations of two random return rates (r_1, r_2) with respect to alternative coordinates.

respect to two alternative unit vectors **u** and **v** we can see from Figure 8.2 that most of the data lie close to the **u** axis. This leads us to take on the following mathematical challenge.

How do we find the basis vectors u and v?

To answer this we require the covariance matrix of the two return rates:

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}.$$

The variance of any portfolio $w_1 r_1 + w_2 r_2$ is then given by

$$\mathbf{w}^T \mathbf{V} \mathbf{w} = w_1^2 \sigma_1^2 + w_2^2 \sigma_2^2 + 2 w_1 w_2 \sigma_{12}. \tag{8.7}$$

However, since we restrict the variance calculation to unit vectors,

$$\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix} \quad \text{for some} \quad \theta \in [0, 2\pi),$$

we are able to express the variance as a function of θ, as follows:

$$\begin{aligned} g(\theta) &= \cos^2\theta \sigma_1^2 + \sin^2\theta \sigma_2^2 + 2\cos\theta \sin\theta \sigma_{12} \\ &= \cos^2\theta \sigma_1^2 + \sin^2\theta \sigma_2^2 + \sin 2\theta \sigma_{12} \\ &= \frac{\sigma_1^2 + \sigma_2^2}{2} + \cos 2\theta \left(\frac{\sigma_1^2 - \sigma_2^2}{2} \right) + \sin 2\theta \sigma_{12}. \end{aligned}$$

As θ varies from 0 to 2π the function $g(\theta)$ delivers the variance for all portfolios with unit vector weights. We want to find where this function attains its maximum and its minimum. Differentiating the function gives

$$g'(\theta) = -\sin 2\theta (\sigma_1^2 - \sigma_2^2) + 2\cos 2\theta \sigma_{12}.$$

Setting the derivative equal to zero, we get

$$\tan 2\theta = \frac{2\sigma_{12}}{\sigma_1^2 - \sigma_2^2}. \tag{8.8}$$

If θ^* is a solution to this equation then so is $\theta^* + \pi/2$ since the tangent function is π-periodic. Thus, if we can establish that the variance is at its maximum at θ^* then it can be shown that it reaches its minimum at $\theta^* + \pi/2$. In particular the optimal basis vectors for capturing the variability of the asset returns are

$$\mathbf{u} = \begin{pmatrix} \cos\theta^* \\ \sin\theta^* \end{pmatrix} \text{ and } \mathbf{v} = \begin{pmatrix} \cos(\theta^* + \pi/2) \\ \sin(\theta^* + \pi/2) \end{pmatrix} = \begin{pmatrix} -\sin\theta^* \\ \cos\theta^* \end{pmatrix}.$$

Numerical Illustration

To demonstrate how this theory works we consider a numerical example. Suppose that the covariance matrix of our two-asset returns is given by

$$\mathbf{V} = \frac{1}{100}\begin{pmatrix} 9/4 & 5\sqrt{3}/8 \\ 5\sqrt{3}/8 & 1 \end{pmatrix}. \tag{8.9}$$

The denominator of (8.8) is given by

$$\sigma_1^2 - \sigma_2^2 = \frac{1}{80}.$$

Hence, equation (8.8) becomes

$$\tan 2\theta = \frac{\sqrt{3}/80}{1/80} = \sqrt{3} \rightarrow \theta = \frac{\pi}{6} \quad \text{or} \quad \theta = \frac{2\pi}{3}.$$

To check the nature of the stationary point we compute the second derivative:

$$g''(\theta) = -2\cos 2\theta(\sigma_1^2 - \sigma_2^2) - 4\sin 2\theta\sigma_{12}.$$

Evaluating at $\theta = \pi/6$ we find

$$\begin{aligned} g''\left(\frac{\pi}{6}\right) &= -2 \times \frac{1}{2} \times \frac{1}{80} - 4 \times \frac{\sqrt{3}}{2} \times \frac{5\sqrt{3}}{800} \\ &= \frac{-1-3}{80} = -\frac{4}{80} < 0. \end{aligned}$$

Evaluating at $\theta = 2\pi/3$ we find

$$\begin{aligned} g''\left(\frac{2\pi}{3}\right) &= -2 \times -\frac{1}{2} \times \frac{1}{80} - 4 \times -\frac{\sqrt{3}}{2} \times \frac{5\sqrt{3}}{800} \\ &= \frac{1+3}{80} = \frac{4}{80} > 0. \end{aligned}$$

We can conclude that the direction of most variability for the 2-dimensional data is given by

$$\mathbf{u} = \begin{pmatrix} \cos \pi/6 \\ \sin \pi/6 \end{pmatrix} = \begin{pmatrix} \sqrt{3}/2 \\ 1/2 \end{pmatrix}.$$

Orthogonal to this vector is the minor component

$$\mathbf{v} = \begin{pmatrix} \cos 2\pi/3 \\ \sin 2\pi/3 \end{pmatrix} = \begin{pmatrix} -1/2 \\ \sqrt{3}/2 \end{pmatrix}.$$

The maximum variance on the unit circle is thus given by

$$\begin{aligned} g(\pi/6) &= \frac{3}{4} \times \frac{9/4}{100} + \frac{1}{4} \times \frac{1}{100} + 2 \times \frac{\sqrt{3}}{2} \times \frac{5\sqrt{3}}{800} \\ &= \frac{1}{100}\left(\frac{27}{16} + \frac{4}{16} + \frac{15}{16}\right) = \frac{23}{800} = 0.02875. \end{aligned}$$

The minimum variance on the unit circle is thus given by

$$\begin{aligned} g(2\pi/3) &= \frac{1}{4} \times \frac{9/4}{100} + \frac{3}{4} \times \frac{1}{100} + 2 \times -\frac{\sqrt{3}}{2} \times \frac{5\sqrt{3}}{800} \\ &= \frac{1}{100}\left(\frac{9}{16} + \frac{12}{16} - \frac{15}{16}\right) = \frac{3}{800} = 0.00375. \end{aligned}$$

An Interesting Link

Through this numerical example we have found

$$\text{maximum variance of } \frac{23}{800} \text{ occurs at } \mathbf{u} = \begin{pmatrix} \sqrt{3}/2 \\ 1/2 \end{pmatrix},$$

and

$$\text{minimum variance of } \frac{3}{800} \text{ occurs at } \mathbf{v} = \begin{pmatrix} -1/2 \\ \sqrt{3}/2 \end{pmatrix}.$$

We already know that $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, but even more interesting is the fact that these vectors are simply the eigenvectors of the covariance matrix $\mathbf{V}$ (8.9). Furthermore the extreme values of the variance are the corresponding eigenvalues, i.e., we have

$$\frac{1}{100}\begin{pmatrix} 9/4 & 5\sqrt{3}/8 \\ 5\sqrt{3}/8 & 1 \end{pmatrix}\begin{pmatrix} \sqrt{3}/2 \\ 1/2 \end{pmatrix} = \frac{23}{800}\begin{pmatrix} \sqrt{3}/2 \\ 1/2 \end{pmatrix}$$

and

$$\frac{1}{100}\begin{pmatrix} 9/4 & 5\sqrt{3}/8 \\ 5\sqrt{3}/8 & 1 \end{pmatrix}\begin{pmatrix} -1/2 \\ \sqrt{3}/2 \end{pmatrix} = \frac{3}{800}\begin{pmatrix} -1/2 \\ \sqrt{3}/2 \end{pmatrix}.$$

We know from Chapter 2 that the eigenvectors and eigenvalues of any symmetric matrix are precisely what are needed to construct its spectral decomposition, specifically if we let

$$\mathbf{\Gamma} = \begin{pmatrix} \sqrt{3}/2 & -1/2 \\ 1/2 & \sqrt{3}/2 \end{pmatrix} = \begin{pmatrix} \cos(\pi/6) & -\sin(\pi/6) \\ \sin(\pi/6) & \cos(\pi/6) \end{pmatrix}$$

denote the matrix of eigenvectors of $\mathbf{V}$ and

$$\mathbf{D} = \begin{pmatrix} 23/800 & 0 \\ 0 & 3/800 \end{pmatrix}$$

denote the diagonal matrix of its eigenvalues, then we have

$$\mathbf{V} = \mathbb{E}[(\mathbf{X} - \mathbf{e})(\mathbf{X} - \mathbf{e})^T] = \mathbf{\Gamma D \Gamma}^T.$$

We note that applying the eigenvector matrix $\mathbf{\Gamma}$ to a 2-dimensional vector has the effect of rotating it by $\pi/6$ in an anti-clockwise direction; similarly its inverse $\mathbf{\Gamma}^T$ represents a rotation by $\pi/6$ in a clockwise direction. Armed with this insight we can make the following transformation:

- Re-centre the vector of returns about its mean and then rotate by $\pi/6$ in a clockwise direction.
- Mathematically, this translates as

$$\mathbf{r} = \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} \mapsto \mathbf{f} = \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \mathbf{\Gamma}^T(\mathbf{r} - \mathbf{e}).$$

The components f_1 and f_2 are called the principal components of the data, they have mean zero and are uncorrelated; in fact the covariance matrix of $\{f_1, f_2\}$ coincides with $\mathbf{D}$, the diagonal eigenvalue matrix of $\mathbf{V}$ (the covariance matrix of the original returns). To see this we have the following development:

$$\begin{aligned} \mathbb{E}[\mathbf{f}\mathbf{f}^T] &= \mathbb{E}[\mathbf{\Gamma}^T(\mathbf{r} - \mathbf{e})(\mathbf{r} - \mathbf{e})^T\mathbf{\Gamma}] \\ &= \mathbf{\Gamma}^T\mathbb{E}[(\mathbf{r} - \mathbf{e})(\mathbf{r} - \mathbf{e})^T]\mathbf{\Gamma} \\ &= \mathbf{\Gamma}^T\mathbf{V}\mathbf{\Gamma} = \mathbf{\Gamma}^T\left(\mathbf{\Gamma D \Gamma}^T\right)\mathbf{\Gamma} = \mathbf{D}. \end{aligned} \tag{8.10}$$

We can express the original vector of returns in terms of the principal components as follows:

$$\begin{aligned} \begin{pmatrix} r_1 \\ r_2 \end{pmatrix} &= \mathbf{r} = \mathbf{e} + \mathbf{\Gamma f} \\ &= \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} \sqrt{3} \\ 1 \end{pmatrix} f_1 + \frac{1}{2}\begin{pmatrix} -1 \\ \sqrt{3} \end{pmatrix} f_2, \end{aligned}$$

or equivalently

$$r_1 = \mu_1 + \frac{\sqrt{3}}{2} f_1 - \frac{1}{2} f_2,$$
$$r_2 = \mu_2 + \frac{1}{2} f_1 + \frac{\sqrt{3}}{2} f_2.$$

An important property of the principal components is that together they capture all the variability of the original data, i.e., the following holds:

$$\begin{aligned}\text{variance}(r_1) + \text{variance}(r_2) &= \frac{1}{100}\left(\frac{9}{4} + 1\right) \\ &= \frac{1}{100}\left(\frac{23}{8} + \frac{3}{8}\right) = \text{variance}(f_1) + \text{variance}(f_2).\end{aligned}$$

We notice that the first component f_1 captures the bulk of the variability, in fact $100 \times 23/26 \approx 88.5\%$ of it. We call f_1 the major principal component and f_2 the minor principal component; accounting for only 11.5% of the variability. In view of the dominance of the major component we propose the following single-factor model for asset returns:

$$r_i = a_i + b_i f + \varepsilon_i \quad \text{for} \quad i = 1, 2, \tag{8.11}$$

where $f = f_1$ (i.e., the major principal component is the driving factor) and

$$\begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} \sqrt{3} \\ 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} = \frac{f_2}{2}\begin{pmatrix} -1 \\ \sqrt{3} \end{pmatrix}.$$

We note that (8.11) obeys almost all of the properties of the factor model; the residual error terms have zero mean and each one is uncorrelated with the driving factor f. The only deviation from the factor model is that the ε_1 and ε_2 are correlated, in fact they are negatively correlated with coefficient $\rho = -1$. Under this model we can deduce that

$$\sigma_i^2 = b_i^2 \sigma_f^2 + \text{var}(\varepsilon_i) \quad \text{for} \quad i = 1, 2;$$
$$\text{and} \quad \sigma_{12} = b_1 b_2 \sigma_f^2 - \sqrt{\text{var}(\varepsilon_1)\text{var}(\varepsilon_2)}$$

and hence the risk on a portfolio p with weight vector $\mathbf{w} = (w_1, w_2)^T \in \mathbb{R}^2$ is decomposed as

$$\sigma_p^2 = \underbrace{(w_1 b_1 + w_2 b_2)^2 \sigma_f^2}_{\text{systematic risk}} + \underbrace{w_1^2 \text{var}(\varepsilon_1) + w_2^2 \text{var}(\varepsilon_2) - 2 w_1 w_2 \sqrt{\text{var}(\varepsilon_1)\text{var}(\varepsilon_2)}}_{\text{specific risk}}.$$

We observe that, in comparison to the true single-factor model (8.6), the specific risk component has an additional contribution due to the fact that the random error terms ε_1 and ε_2 are correlated.

In summary, we have shown that, in two dimensions, PCA gives rise to a natural model for the two-asset returns. The PCA captures most of the proposed properties of the original single-factor model; the only difference being in the expression for specific risk. The method has a scientific basis and the calculation procedure allows us to establish precisely how much of the variability of the data is captured by the driving factor. We will now show how the ideas can be extended to higher dimensions.

8.3.2 PCA in higher dimensions

In two dimensions we found the orthogonal vectors $\mathbf{u}$ and $\mathbf{v}$ on the unit circle where the variance attained its maximum and minimum respectively. In higher dimensions we have an n-dimensional vector $\mathbf{r} = (r_1, \ldots, r_n)^T$ of random returns with covariance matrix

$$\mathbf{V} = \mathbb{E}[(\mathbf{r} - \mathbf{e})(\mathbf{r} - \mathbf{e})^T].$$

The analogue of the unit circle in higher dimensions is the unit sphere; mathematically we say that a vector $\mathbf{w} \in \mathbb{R}^n$ lies on the $(n-1)$-dimensional sphere if $\mathbf{w}^T\mathbf{w} = \sum_{i=1}^n w_i^2 = 1$. We begin by finding a point on the sphere where the variance attains its maximum, i.e., we solve

$$\text{maximize } \mathbf{w}^T\mathbf{V}\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T\mathbf{w} = 1.$$

We tackle this problem using the Lagrange multiplier technique, i.e., we define the Lagrange function of the problem as

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^T\mathbf{V}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1).$$

Taking the gradient of this function gives

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, \lambda) = 2\mathbf{V}\mathbf{w} - 2\lambda\mathbf{w},$$
$$\frac{\partial}{\partial\lambda}\mathcal{L}(\mathbf{w}, \lambda) = 1 - \mathbf{w}^T\mathbf{w}.$$

Let $\mathbf{w} := \mathbf{u}_1$ and $\lambda := \lambda_1$ denote the coordinates where the gradient equals zero, then the first-order conditions read

$$\mathbf{V}\mathbf{u}_1 = \lambda_1\mathbf{u}_1 \quad \text{and} \quad \mathbf{u}_1^T\mathbf{u}_1 = 1.$$

We see that the solution $\mathbf{u}_1$ is precisely an eigenvector of the positive definite covariance matrix $\mathbf{V}$. Furthermore, since

$$\mathbf{u}_1^T\mathbf{V}\mathbf{u}_1 = \lambda_1\mathbf{u}_1^T\mathbf{u}_1 = \lambda_1$$

we also discover that the corresponding eigenvalue λ_1 is the value of the maximum variance attained on the sphere.

We have found that the maximum variance on the $(n-1)$-dimensional sphere occurs at $\mathbf{u}_1$. The next step is to consider the $(n-2)$-dimensional sphere that is orthogonal to $\mathbf{u}_1$; we search this lower-dimensional sphere for the point where variance is at its maximum.

Mathematically, we need to solve

$$\text{maximize } \mathbf{w}^T\,\mathbf{V}\mathbf{w} \quad \text{subject to} \quad \mathbf{w}^T\,\mathbf{w} = 1 \quad \text{and} \quad \mathbf{w}^T\mathbf{u}_1 = 0.$$

The Lagrangian in this case is given by

$$\mathcal{L}(\mathbf{w}, \lambda, \nu) = \mathbf{w}^T\mathbf{V}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1) - \nu(\mathbf{w}^T\mathbf{u}_1 - 0),$$

and its gradient gives

$$\begin{aligned}
\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}, \lambda, \nu) &= 2\mathbf{V}\mathbf{w} - 2\lambda\mathbf{w} - \nu\mathbf{u}_1,\\
\frac{\partial}{\partial\lambda}\mathcal{L}(\mathbf{w}, \lambda, \nu) &= 1 - \mathbf{w}^T\mathbf{w},\\
\frac{\partial}{\partial\nu}\mathcal{L}(\mathbf{w}, \lambda, \nu) &= -\mathbf{w}^T\mathbf{u}_1.
\end{aligned}$$

Let $\mathbf{w} := \mathbf{u}_2$, $\lambda := \lambda_2$ and $\nu := \nu_2$ denote the coordinates where the gradient equals zero, then the first-order conditions read

$$\mathbf{V}\mathbf{u}_2 = \lambda_2\mathbf{u}_2 + \frac{1}{2}\nu_2\mathbf{u}_1, \tag{8.12}$$

$$\mathbf{u}_2^T\mathbf{u}_2 = 1 \quad \text{and} \quad \mathbf{u}_2^T\mathbf{u}_1 = 0. \tag{8.13}$$

If we multiply (8.12) on the left by $\mathbf{u}_1$ and use the constraints, we get

$$\mathbf{u}_1^T\mathbf{V}\mathbf{u}_2 = \lambda_2\mathbf{u}_1^T\mathbf{u}_2 + \frac{1}{2}\nu_2\mathbf{u}_1^T\mathbf{u}_1 = \frac{\nu_2}{2}.$$

Equivalently, since the covariance matrix $\mathbf{V}$ is symmetric we can write

$$\mathbf{u}_1^T\mathbf{V}\mathbf{u}_2 = \mathbf{u}_2^T\mathbf{V}\mathbf{u}_1 = \lambda_1\mathbf{u}_2^T\mathbf{u}_1 = 0,$$

thus we deduce that $\nu_2 = 0$ and so (8.12) becomes

$$\mathbf{V}\mathbf{u}_2 = \lambda_2\mathbf{u}_2.$$

We observe that, as before, the point on the $(n-2)$-dimensional sphere (orthogonal to $\mathbf{u}_1$) where the variance attains its maximum is the eigenvector $\mathbf{u}_2$ corresponding to λ_2, the second largest eigenvalue of $\mathbf{V}$. The size of the eigenvalue coincides with the maximal variance since

$$\mathbf{u}_2^T\mathbf{V}\mathbf{u}_2 = \lambda_2\mathbf{u}_2^T\mathbf{u}_2 + \nu_2\mathbf{u}_2^T\mathbf{u}_1 = \lambda_2.$$

This process can be continued; to summarize the results, we let

$$f(\mathbf{w}) = \mathbf{w}^T\mathbf{V}\mathbf{w} \quad \text{for} \quad \mathbf{w} \in \mathbb{R}^n$$

denote the variance function of a portfolio with arbitrary weight vector $\mathbf{w}$ and let $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ denote the eigenvectors of $\mathbf{V}$ corresponding to the ordered eigenvalues

$$\lambda_1 > \lambda_2 > \cdots > \lambda_n > 0.$$

We have shown that:

- The maximum of $f(\mathbf{w})$ over the $(n-1)$-dimensional sphere is attained at the point $\mathbf{w} = \mathbf{u}_1$ and $f(\mathbf{u}_1) = \lambda_1$.
- The maximum of $f(\mathbf{w})$ over the $(n-2)$-dimensional sphere orthogonal to $\{\mathbf{u}_1\}$ is attained at the point $\mathbf{w} = \mathbf{u}_2$ and $f(\mathbf{u}_2) = \lambda_2$.

More generally, it can be shown that:

- The maximum of $f(\mathbf{w})$ over the $(n-k)$-dimensional sphere orthogonal to $\{\mathbf{u}_1, \ldots, \mathbf{u}_{k-1}\}$ is attained at the point $\mathbf{w} = \mathbf{u}_k$ and $f(\mathbf{u}_k) = \lambda_k$, for $k = 3, \ldots, n-1$.
- The minimum of $f(\mathbf{w})$ over the $(n-1)$-dimensional sphere is attained at the point $\mathbf{w} = \mathbf{u}_n$ and $f(\mathbf{u}_n) = \lambda_n$.

In analogy to the 2-dimensional case we have discovered that the eigenvectors of the covariance matrix $\mathbf{V}$ serve as a more illuminating coordinate basis than the usual standard basis; $\mathbf{u}_1$ represents the direction where most of the variability is captured, followed by, in order of importance, $\mathbf{u}_2, \mathbf{u}_3, \ldots$, etc., through to $\mathbf{u}_n$, where the least variation is captured. To continue the analogy, we let

$$\mathbf{\Gamma} = (\mathbf{u}_1 \mathbf{u}_2 \cdots \mathbf{u}_n) = \begin{pmatrix} u_1^{(1)} & u_1^{(2)} & \cdots & u_1^{(n)} \\ u_2^{(1)} & u_2^{(2)} & \cdots & u_2^{(n)} \\ \vdots & \vdots & \ddots & \vdots \\ u_n^{(1)} & u_n^{(2)} & \cdots & u_n^{(n)} \end{pmatrix}$$

denote the matrix of eigenvectors of the covariance matrix $\mathbf{V}$, and recall that its spectral decomposition is given by

$$\mathbf{V} = \mathbb{E}[(\mathbf{r} - \mathbf{e})(\mathbf{r} - \mathbf{e})^T] = \mathbf{\Gamma D \Gamma}^T \quad \text{where} \quad \mathbf{D} := \operatorname{diag}(\lambda_1, \ldots, \lambda_n).$$

This decomposition, when written component-wise, is given by

$$\sigma_{ij} = \sum_{k=1}^{n} u_i^{(k)} u_j^{(k)} \lambda_k. \tag{8.14}$$

We can set $i = j$ in this formula to deduce that

$$\operatorname{var}(r_i) = \sigma_{ii} = \sum_{k=1}^{n} \left(u_i^{(k)}\right)^2 \lambda_k \quad \text{for} \quad i = 1, \ldots, n. \tag{8.15}$$

We transform the vector of returns by re-centring it about its mean followed by an application of $\mathbf{\Gamma}^T$, i.e., we have

$$\mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} \mapsto \mathbf{f} = \mathbf{\Gamma}^T(\mathbf{r} - \mathbf{e}) = \begin{pmatrix} \mathbf{u}_1^T(\mathbf{r} - \mathbf{e}) \\ \vdots \\ \mathbf{u}_n^T(\mathbf{r} - \mathbf{e}) \end{pmatrix} = \begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}.$$

The original returns can be expressed in terms of the transformed returns $\{f_1, \ldots, f_n\}$, called the principal components, via the formula

$$\mathbf{r} = \mathbf{e} + \mathbf{\Gamma f}.$$

The random vector $\mathbf{f}$ of principal components has zero mean and, following the same calculation as (8.10) in the 2-dimensional case, its covariance matrix is the diagonal matrix of eigenvalues, i.e.,

$$\mathbb{E}[\mathbf{f}\mathbf{f}^T] = \mathbf{D} = \text{diag}(\lambda_1, \ldots, \lambda_n),$$

and so the components are uncorrelated. Furthermore, the total variability of the original data is completely captured by the principal components; this follows since

$$\begin{aligned}
\text{var}(r_1) + \cdots + \text{var}(r_n) &= \sum_{i=1}^{n} \sigma_{ii} \quad \text{(sum of diagonal entries of } \mathbf{V}) \\
&= \sum_{i=1}^{n} \left(\sum_{k=1}^{n} \left(u_i^{(k)} \right)^2 \lambda_k \right) \quad \text{(using equation (8.15))} \\
&= \sum_{k=1}^{n} \underbrace{\left(\sum_{i=1}^{n} \left(u_i^{(k)} \right)^2 \right)}_{=\mathbf{u}_k^T \mathbf{u}_k = 1} \lambda_k = \sum_{k=1}^{n} \lambda_k \\
&= \text{var}(f_1) + \cdots + \text{var}(f_n).
\end{aligned}$$

The eigenvalues are supplied in descending order and thus we can calculate that the first l principal components capture

$$\frac{\sum_{i=1}^{l} \lambda_i}{\sum_{i=1}^{n} \lambda_i} \times 100\% \text{ of the total variability.} \tag{8.16}$$

If all principal components are employed then we have an exact relation:

$$\begin{pmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_n \end{pmatrix} + \begin{pmatrix} u_1^{(1)} & \cdots & u_1^{(l)} & \cdots & u_1^{(n)} \\ \vdots & & \vdots & & \vdots \\ u_i^{(1)} & & u_i^{(l)} & & u_i^{(n)} \\ \vdots & & \vdots & & \vdots \\ u_n^{(1)} & \cdots & u_n^{(l)} & \cdots & u_n^{(n)} \end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_i \\ \vdots \\ f_n \end{pmatrix}.$$

In practice it is common to find that the bulk of the variance can be captured by a relatively small number of components. Thus, armed with the eigenvalues of $\mathbf{V}$, the risk manager can use the formula (8.16) to determine an appropriate number l of components

to employ in a factor model. We can manipulate the above relationship to express the model as

$$\begin{pmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_n \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_n \end{pmatrix} + \begin{pmatrix} u_1^{(1)} & \cdots & u_1^{(l)} \\ \vdots & & \vdots \\ u_i^{(1)} & & u_i^{(l)} \\ \vdots & & \vdots \\ u_n^{(1)} & \cdots & u_n^{(l)} \end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_l \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}, \tag{8.17}$$

where

$$\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} u_1^{(l+1)} & \cdots & u_1^{(n)} \\ \vdots & & \vdots \\ u_i^{(l+1)} & & u_i^{(n)} \\ \vdots & & \vdots \\ u_n^{(l+1)} & \cdots & u_n^{(n)} \end{pmatrix} \begin{pmatrix} f_{l+1} \\ \vdots \\ f_n \end{pmatrix}. \tag{8.18}$$

We can conclude that the principal components define a model (the PCA model) for asset returns that enjoys almost all the properties we proposed for a multi-factor model; specifically, its residual error terms, given by (8.18), have zero mean and are uncorrelated with the driving components $\{f_1, \ldots, f_l\}$. We stop short of calling the PCA model a true factor model because the residual errors $\{\varepsilon_1, \ldots, \varepsilon_n\}$ are correlated to each other whereas, for the true factor model, the residual error terms are assumed to be uncorrelated. Nevertheless, the PCA model has many appealing features that make it popular in applications; it efficiently reduces the dimension of the problem, the user can assign an importance ranking to the components using the eigenvalues and, in some cases, the components can be shown to have an economic interpretation.

Essential Mathematics for Market Risk Management
by Simon Hubbert

9 The Value at Risk Concept

So far in this book we have developed the famous mean–variance framework to examine the trade-off between potential reward (measured by the mean of the portfolio return) and the in-built risk (measured by the variance of the portfolio return). These two statistics essentially summarize what we expect to happen to the portfolio on average. In probabilistic terms, if we are able to picture the probability density function of the portfolio returns, then the mean and volatility cover the central part, where most of the probability mass is held (see Figure 9.1).

Today's financial marketplace is a more complex landscape than it was in the late 1950s when the pioneering mean–variance framework was developed by Markowitz. Many of the major markets have seen deregulation, new investment opportunities have arisen, e.g. in the emerging economies of Asia and South America. Investment banks now hold trading portfolios that consist of enormous numbers of positions in a wide range of products. The use of derivative products for speculation and hedging has sky-rocketed. All of these developments present fresh challenges for investors, traders and particularly for risk managers.

In the mid-1990s a whole host of financial disasters hit the headlines and consequently served as a wake-up call for the industry to take action. New regulatory controls have since emerged and perhaps the most important innovation is a formula that determines the minimal capital an investment bank must set aside to act as a buffer for losses due to adverse market risk exposure. All financial institutions are now equipped with teams of quantitative analysts who work closely with the traders; monitoring their risks and working on the development of sophisticated models used to forecast and reduce risk exposures. The central problem that all risk managers must address is:

The Value at Risk Challenge

How much can a financial portfolio potentially lose if it is exposed to an unfavourably rare event?

We devote the rest of this chapter to developing the famous Value at Risk (VaR) measure which directly answers this question. In order to do so it is clear that we now have to move from the central part of the portfolio distribution and give more attention to the tail where large losses can occur (see Figure 9.2).

9.1 A FRAMEWORK FOR VALUE AT RISK

We begin our investigation by setting the scene for what follows. We recall from Chapter 5 that we have two ways of measuring the return rate of a simple financial asset (5.1); either

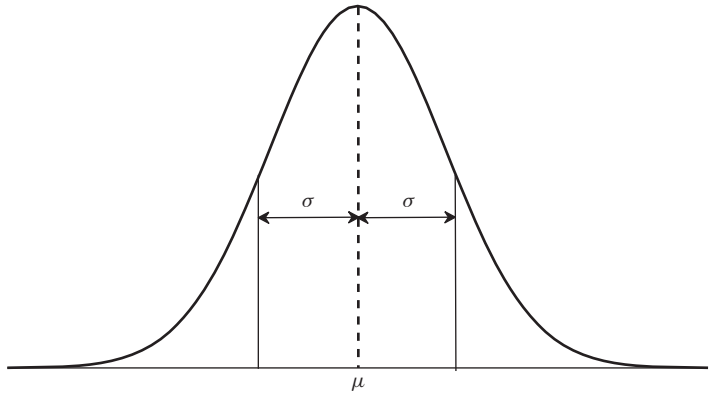


Figure 9.1 Capturing probability weight around the mean with the measure of volatility.

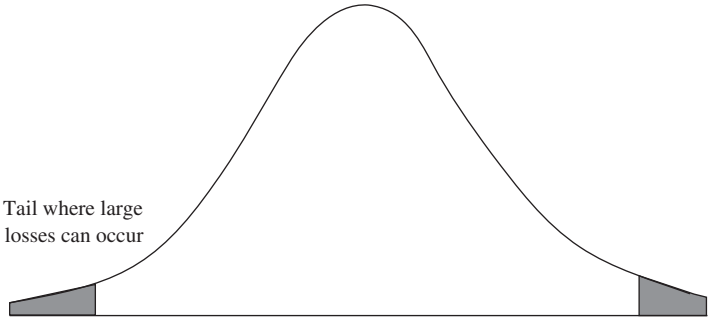


Figure 9.2 The loss tail of the probability density function of portfolio returns.

in standard or logarithmic form:

$$r(t:\tau) = \begin{cases} r_{\text{std}}(t:\tau) = \frac{S(t+\tau)-S(t)}{S(t)} & \text{the standard return;} \\ r_{\log}(t:\tau) = \log\left(\frac{S(t+\tau)}{S(t)}\right) & \text{the log return.} \end{cases}$$

Since we are now focused on the loss our portfolio may suffer, we make a simple adjustment and define the τ-day loss rate as

$$l(t:\tau) = \begin{cases} l_{\text{std}}(t:\tau) = -\frac{S(t+\tau)-S(t)}{S(t)} & \text{the standard loss;} \\ l_{\log}(t:\tau) = -\log\left(\frac{S(t+\tau)}{S(t)}\right) & \text{the log loss.} \end{cases} \tag{9.1}$$

In our development of the mean–variance framework we chose to work with the standard return rate. In contrast, for our investigation into the potential large loss of a portfolio we shall employ the log loss rate; this is the preferred choice of practicing risk managers. Furthermore, since the τ-day log loss can be expressed as a sum of future daily log losses,

i.e., since

$$l(t:\tau) = \sum_{\Delta=0}^{\tau-1} l(t+\Delta:1), \tag{9.2}$$

we shall now place attention on daily log losses and, for convenience, we shall use the simplified notation

$$l(t) = l(t:1) \qquad i = 1, 2, \ldots, n.$$

Suppose that we have a collection of n risky assets, whose daily log loss rates are the random variables $\{l_1(t), \ldots, l_n(t)\}$. Suppose that we create an investment portfolio from these assets with weights $\{w_1, \ldots, w_n\}$. The daily loss random variable that we choose to monitor is the following weighted sum:

$$l_p(t) = \sum_{i=1}^{n} w_i l_i(t). \tag{9.3}$$

We note that $l_p(t)$ does not represent the daily log loss rate of the portfolio since, using (5.7), we have

$$\begin{aligned} \text{daily portfolio log loss } &= -\log\left(\frac{\text{portfolio value at } (t+1)}{\text{portfolio value at } t}\right) \\ &= -\log\Big[\sum_{i=1}^{n} w_i \exp\left(-l_i(t)\right)\Big]. \end{aligned}$$

Comparing this expression to (9.3) it is clear that the linear combination of log losses $l_p(t)$ is much more manageable. Indeed, given that we are only considering daily intervals we can also argue, as in Section 5.1, that

$$\begin{aligned} l_p(t) \approx &\text{ the standard daily rate of portfolio loss} \\ &= -\frac{\big(\text{portfolio value at } (t+1)\big) - \big(\text{portfolio value at } t\big)}{\text{portfolio value at } t}. \end{aligned}$$

Furthermore, we can easily convert the portfolio loss rate $l_p(t)$ into an approximate monetary loss, since

$$\begin{aligned} \text{daily monetary loss} &= -\Big(\big(\text{portfolio value at } (t+1)\big) - \big(\text{portfolio value at } t\big)\Big) \\ &\approx V(t) \times l_p(t), \end{aligned} \tag{9.4}$$

where $V(t)$ denotes the portfolio value at time t. In view of this, we set

$$\mathcal{L}_t = V(t) \sum_{i=1}^{n} w_i l_i(t) \tag{9.5}$$

and take this random variable to represent the approximate loss the portfolio can suffer in a 24-hour period. Now, all possible information regarding $\mathcal{L}_t$ is contained in its distribution function $F : \mathbb{R} \to [0, 1]$ defined by

$$F(x) = \mathbb{P}[\mathcal{L}_t \leq x].$$

A useful way to view this equation is as follows. If we consider a daily portfolio loss of $\$x$ then the distribution tells us that the value $F(x) \in [0, 1]$ is precisely the probability that $\mathcal{L}_t$ will remain below x. We want to unlock this equation. Specifically, we want to supply a probability level and be able to access the corresponding upper bound x on our daily portfolio loss; what we need is the inverse of F.

At this point we know nothing about F apart from that its value increases from zero to one as its argument x moves through values of extreme profit to extreme loss; i.e., from $-\infty$ to ∞. This property is common to all probability distributions (see Definition 3.2), however, as we know from Chapter 3, there are many different functional forms that a distribution can take. The easiest case to deal with is where F increases monotonically across the whole of $\mathbb{R}$. In this case F has a unique inverse, i.e., we can deduce that for any value $\alpha \in [0, 1]$ we can find a unique $x \in \mathbb{R}$, such that $F(x) = \alpha$; we write $x = F^{-1}(\alpha)$. If the distribution function exhibits jumps and/or plateaux then it does not possess a unique inverse and, instead, we define a generalized inverse by

$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}. \tag{9.6}$$

Loosely speaking, $F^{-1}(\alpha)$ is the smallest number x such that $F(x) = \mathbb{P}[\mathcal{L}_t \leq x] \geq \alpha$.

9.1.1 A motivating example

Let us consider the ideal case where the distribution is strictly increasing on $\mathbb{R}$. Suppose, in addition, that F is differentiable and so we know, from Chapter 3, that the derivative of F is precisely the probability density function $p(x)$ of $\mathcal{L}_t$ and we can write

$$F(x) = \int_{-\infty}^{x} p(u)du \Rightarrow F'(x) = p(x).$$

To motivate our future development let us suppose a trader has invested \$1 million in a portfolio of risky assets whose daily loss random variable $\mathcal{L}_t$ is given by (9.5). Suppose that the 90% quantile equates to a loss of $\$150K$, i.e., mathematically we have

$$150{,}000 = F^{-1}(0.9), \quad \text{or} \quad \mathbb{P}[\mathcal{L}_t \leq 150{,}000] = 0.9.$$

See Figure 9.3 for an illustration. There are many (equivalent) ways a risk analyst could interpret this information:

1. With 90% confidence the unknown daily loss will be less than \$150K.
2. There is a 10% chance of losing more than \$150K over a day.
3. Over the next 100 days we can expect to lose more than \$150K on 10 occasions.

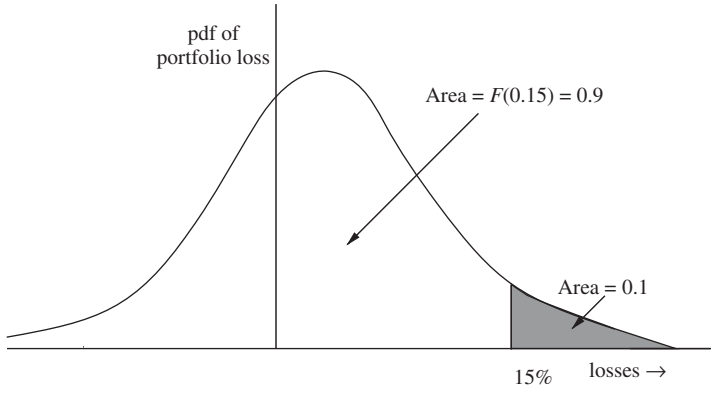


Figure 9.3 Illustration of using the tail statistics.

If the risk manager, who monitors the trader's activity, is confident that there is a sufficient level of risk capital (held in the buffer fund) to cover the expectation that 10 out of 100 losses will exceed \$150K, then there is no reason to call for action. If not, then the risk manager should consider re-balancing the portfolio and/or increasing the buffer fund of risk capital.

9.1.2 Defining value at risk

The preceding example motivates the discussion but it is not how risk managers perform in practice. In reality, it is more likely that a confidence level α will be supplied, usually at 0.99 or 0.95, and the risk manager is given the task of finding the value of the (generalized) inverse function $F^{-1}(\alpha)$ at that confidence level. In other words, in the case of a continuous and strictly increasing loss distribution, finding the unique value x that solves

$$F(x) = \int_{-\infty}^{x} p(u)du = \alpha.$$

We call this special value the daily value at risk and denote it VaR_α. Thus, the computation of the Value at Risk for a portfolio amounts to evaluating the (generalized) inverse distribution function of $\mathcal{L}_t$ at the given confidence level. In view of (9.6), its formal definition is as follows:

Definition 9.1. *Let F denote the distribution function of the daily loss random variable (9.5) for a portfolio of n risky assets. For a given confidence level $\alpha \in [0, 1]$ the daily Value at Risk for the portfolio is the loss that is likely to be exceeded $(1 - \alpha)100\%$ of the time. Mathematically, it is given by*

$$\mathrm{VaR}_\alpha = F^{-1}(\alpha) = \inf_{r \in [0,\infty)} \{F(r) \geq \alpha\}. \tag{9.7}$$

In the case where F is strictly increasing on $\mathbb{R}$, then VaR_α is the unique solution to the equation

$$F(\mathrm{VaR}_\alpha) = \alpha.$$

In the 1990s a trend began to emerge, amongst the leading investment banks, for using VaR as a means of quantifying the market risk exposure of enormous trading portfolios. The creation of such in-house VaR systems is a non-trivial task requiring specialist skills in IT (construction of real-time financial data feeds) and financial engineering (the development and implementation of pricing models and the VaR calculator). It is not surprising that only the top-flight investment banks were able to devote their resources to such a project. A major development occurred in 1994 when JP-Morgan took the bold decision to publish its own VaR methodology and, in addition, made its risk factor data set (i.e., the input to its VaR calculator) freely available to other financial institutions. The JP-Morgan package is known as the RiskMetrics toolbox for VaR and its open release sparked an enormous amount of interest from practitioners and academics alike. The RiskMetrics toolbox crucially gave other smaller financial institutions the ability to measure their exposure to market risk.

The release of RiskMetrics is, in part, responsible for transforming VaR from a concept to a practical tool which is now used on a daily basis in financial institutions across the globe. Academic interest in risk management has since mushroomed and further advances continue to be made. We devote the rest of this chapter to the properties of VaR, the kick-off point of the modern risk management revolution.

9.2 INVESTIGATING VALUE AT RISK

We begin our investigation with the very simple observation that the VaR of any portfolio can be viewed as a function of its portfolio weights, i.e., we can write

$$\mathrm{VaR}_\alpha = \mathrm{VaR}_\alpha(w_1, \ldots, w_n).$$

We now recall that if we are given any continuously differentiable function $f(x_1, \ldots, x_n)$ of n variables we can investigate the effect of a small perturbation in one of its variables by using a first-order Taylor approximation; that is, we have

$$\underbrace{f(x_1, \ldots, x_i + h, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}_{\text{change in } f \text{ due to small change in } x_i}$$
$$\approx h \frac{\partial}{\partial x_i} f(x_1, \ldots, x_i, \ldots, x_n).$$

Applying this to $\mathrm{VaR}_\alpha(w_1, \ldots, w_n)$ we see that

$$\underbrace{\mathrm{VaR}_\alpha(w_1, \ldots, w_i + h, \ldots, w_n) - \mathrm{VaR}_\alpha(w_1, \ldots, w_i, \ldots, w_n)}_{\text{change in } \mathrm{VaR}_\alpha \text{ due to small change in } i\text{th exposure}}$$
$$\approx h \frac{\partial}{\partial w_i} \mathrm{VaR}_\alpha(w_1, \ldots, w_i, \ldots, w_n) = h \frac{\partial \mathrm{VaR}_\alpha}{\partial w_i}.$$

The partial derivative of the VaR function has a special interpretation and we provide the following definition:

Definition 9.2. *Let* $\mathrm{VaR}_\alpha(w_1, \ldots, w_n)$ *denote the* VaR *function for a portfolio of n risky assets with weights* $w_1, \ldots, w_n$. *The Marginal Value at Risk* (MVaR) *of the portfolio with*

respect to the ith exposure is defined to be the ith partial derivative of $\text{VaR}_\alpha(w_1, \ldots, w_n)$. *We write*

$$\text{MVaR}_\alpha^{(i)} = \frac{\partial}{\partial w_i}\text{VaR}_\alpha(w_1, \ldots, w_i, \ldots, w_n) = \frac{\partial \text{VaR}_\alpha}{\partial w_i},$$

We continue the investigation with yet another very simple observation. We recall, from Definition 9.1, that if the loss distribution is continuous and strictly increasing the Value at Risk for the portfolio is the unique solution to

$$\begin{aligned} F(\text{VaR}_\alpha) &= \mathbb{P}\Big[\mathcal{L}_t \leq \text{VaR}_\alpha\Big] \\ &= \mathbb{P}\Big[V(t)\sum_{i=1}^{n} w_i l_i(t) \leq \text{VaR}_\alpha\Big] = \alpha. \end{aligned} \tag{9.8}$$

If we let $\lambda > 0$ and scale up (or down) our positions such that $w_i \mapsto \lambda w_i$ for $i = 1, \ldots, n$ then, accordingly, $\text{VaR}_\alpha \mapsto \lambda \text{VaR}_\alpha$ since (9.8) implies

$$\mathbb{P}\Big[V(t)\sum_{i=1}^{n} \lambda w_i l_i(t) \leq \lambda \text{VaR}_\alpha\Big] = \alpha.$$

In mathematical terms, we have simply shown that the VaR measure is (positively) linearly homogeneous:

$$\text{VaR}_\alpha(\lambda w_1, \ldots, \lambda w_n) = \lambda \text{VaR}_\alpha(w_1, \ldots, w_n) \quad \text{where} \quad \lambda > 0.$$

The following result is Euler's theorem and reveals an interesting property of such functions:

Theorem 9.3. *If* $f(x_1, \ldots, x_n)$ *is linearly homogeneous then*

$$f(x_1, \ldots, x_N) = \sum_{i=1}^{n} x_i \frac{\partial}{\partial x_i} f(x_1, \ldots, x_n).$$

When we apply this result to VaR we have the following useful expression:

$$\text{VaR}_\alpha = \sum_{i=1}^{n} w_i \text{MVaR}_\alpha^{(i)}. \tag{9.9}$$

This additivity property is very helpful for explaining how the VaR_α figure can be broken down into its constituent components.

The rapid promotion of VaR from a simple back-of-an-envelope concept to a vital working tool has caused many academics and practitioners to regard it with suspicion. Value at Risk is a risk measure that appears to have been universally embraced by financial institutions and the regulatory bodies, and as such it has two key purposes:

- To determine the level of risk capital that should be allocated to a buffer fund in order to help absorb unexpected losses.
- To quantify the day-to-day market risk exposure of an institution's large trading portfolio.

These are both important functions and it is worthwhile scrutinizing the suitability of VaR in these two areas.

9.2.1 The suitability of value at risk to capital allocation

Value at Risk is just one of many potential measures that can be used to measure the risk of an unexpected portfolio loss. We can talk of an arbitrary risk measure as a function ρ such that:

$$\text{if } \mathcal{L} \text{ represents the daily monetary loss random variable of portfolio } p$$
$$\text{then } \rho(\mathcal{L}) \text{ quantifies the risk that } p \text{ will suffer an unexpected loss.}$$

Any risk measure that is proposed as a candidate to determine the level of risk capital must satisfy some basic properties. In the late 1990s, Artzner *et al.* (1999) introduced the concept of a coherent risk measure. Specifically, a risk measure ρ is said to be coherent if it satisfies the following four properties:

- Translation invariance.
  This property tells us that if an attempt is made to dampen the loss random variable $\mathcal{L}$ by augmenting the portfolio with a fixed sum of cash then the portfolio risk is reduced accordingly. Mathematically, we write
  $$\rho(\mathcal{L} - \text{cash amount}) = \rho(\mathcal{L}) - \text{cash amount}.$$
  We remark that VaR_α satisfies this property, since
  $$\mathbb{P}\big[\mathcal{L} \leq x\big] \leq \alpha \Rightarrow \mathbb{P}\big[\mathcal{L} - \text{cash amount} \leq x - \text{cash amount}\big] \leq \alpha.$$
- Positive linear homogeneity.
  We have already met this property, which states that if the exposures to all assets in a portfolio are each scaled by a factor $\lambda > 0$ then the portfolio risk scales accordingly. Mathematically, we write
  $$\rho(\lambda\mathcal{L}) = \lambda\rho(\mathcal{L}) \quad \text{for} \quad \lambda > 0.$$
  We have already demonstrated that VaR satisfies this property.
- Monotonicity.
  This property states that if one portfolio is known to be more risky than the other then the risk measure should always correctly distinguish between them. Specifically, if the loss random variable of one portfolio, $\mathcal{L}_1$ say, is bounded from above by $\mathcal{L}_2$, (the daily loss on a different, more risky portfolio), then ρ should assign a higher value to $\mathcal{L}_2$ than $\mathcal{L}_1$. Mathematically, we write
  $$\text{if } \mathcal{L}_1 \leq \mathcal{L}_2 \quad \text{then} \quad \rho(\mathcal{L}_1) \leq \rho(\mathcal{L}_2).$$
  We can demonstrate that VaR enjoys this property by the following argument. Let $\mathcal{L}_1$ and $\mathcal{L}_2$ represent the loss random variables for two portfolios p_1 and p_2 say, and let F_1 and F_2 denote their respective distribution functions. If it is known that $\mathcal{L}_1 \leq \mathcal{L}_2$ then it follows that
  $$F_2(x) \leq F_1(x) \quad \text{for every} \quad x \in \mathbb{R}. \tag{9.10}$$

The VaR for portfolio p_2, at a confidence level $\alpha \in [0, 1]$, is defined to be the smallest number $\text{VaR}_\alpha^{(2)}$ such that

$$\alpha \leq F_2(\text{VaR}_\alpha^{(2)}).$$

In view of (9.10) we can deduce that

$$\alpha \leq F_2(\text{VaR}_\alpha^{(2)}) \leq F_1(\text{VaR}_\alpha^{(2)}). \tag{9.11}$$

Now since the corresponding VaR for p_1 is the smallest number $\text{VaR}_\alpha^{(1)}$ that satisfies

$$\alpha \leq F_1(\text{VaR}_\alpha^{(1)}),$$

we can conclude from (9.11) that

$$\text{VaR}_\alpha^{(1)} \leq \text{VaR}_\alpha^{(2)}$$

and hence VaR satisfies the monotonicity property.

- Sub-additivity.
  This property states that if two different portfolios p_1 and p_2 are combined to make a new portfolio then its risk should not be greater than the sum of the risks of p_1 and p_2. This property is designed to encourage diversification; the risk of the whole trading portfolio should be at most as big as the sum of the risks of any possible breakdown of it. Mathematically we write, if $\mathcal{L}_1$ and $\mathcal{L}_2$ represent the loss random variables for two portfolios p_1 and p_2, then

$$\rho(\mathcal{L}_1 + \mathcal{L}_2) \leq \rho(\mathcal{L}_1) + \rho(\mathcal{L}_2).$$

One of the biggest criticisms regarding the use of VaR is that, in general, it is not sub-additive. We can demonstrate this by examining the following set-up. Let $\mathcal{L}_1$ and $\mathcal{L}_2$ denote the daily loss random variables for two distinct financial portfolios. We shall assume that $\mathcal{L}_1$ and $\mathcal{L}_2$ are independent and share the same distribution and density function:

$$F(x) = 1 - \frac{1}{\sqrt{x}} \quad \text{and} \quad p(x) = \frac{dF(x)}{dx} = \frac{1}{2x^{3/2}} \quad \text{for} \quad x \geq 1.$$

Under this assumption we know that for any confidence level α the corresponding VaR measure is the same for both portfolios, thus we let VaR_α denote this common value that satisfies

$$\mathbb{P}[\mathcal{L}_1 \leq \text{VaR}_\alpha] = \mathbb{P}[\mathcal{L}_2 \leq \text{VaR}_\alpha] = \alpha.$$

Furthermore, given that the losses are independent we can employ (3.19) to deduce that the distribution of their sum is given by

$$\begin{aligned} F_{\mathcal{L}_1+\mathcal{L}_2}(z) &= \mathbb{P}[\mathcal{L}_1 + \mathcal{L}_2 \leq z] \\ &= \int_1^{z-1} F(z-u)dF(u) = \int_1^{z-1} F(z-u)p(u)du \\ &= \int_1^{z-1} \left(1 - \frac{1}{\sqrt{z-u}}\right) \frac{1}{2u^{3/2}} du. \end{aligned}$$

It is straightforward to verify that the derivative of

$$u \mapsto -\frac{1}{\sqrt{u}} + \frac{1}{z}\sqrt{\frac{z-u}{u}}$$

is precisely the above integrand. Thus we can conclude that

$$F_{\mathcal{L}_1+\mathcal{L}_2}(z) = 1 - 2\frac{\sqrt{z-1}}{z}.$$

We can now use the inequality $\sqrt{2z} < 2\sqrt{z-1}$, which is valid for $z > 2$, to deduce that

$$F_{\mathcal{L}_1+\mathcal{L}_2}(z) < 1 - \frac{2}{\sqrt{z}} = F_{2\mathcal{L}_1}(z)(= F_{2\mathcal{L}_2}(z)),\ z > 2.$$

Now, using the argument we used in establishing the monotonicity VaR we can conclude that

$$F_{\mathcal{L}_1+\mathcal{L}_2}(z) < F_{2\mathcal{L}_1}(z) \Rightarrow \mathrm{VaR}_\alpha(2\mathcal{L}_1) < \mathrm{VaR}_\alpha(\mathcal{L}_1 + \mathcal{L}_2).$$

Furthermore, given that VaR is linearly homogeneous we have that

$$\mathrm{VaR}_\alpha(2\mathcal{L}_1) = 2\mathrm{VaR}_\alpha(\mathcal{L}_1) = \mathrm{VaR}_\alpha(\mathcal{L}_1) + \mathrm{VaR}_\alpha(\mathcal{L}_2),$$

and this establishes that VaR does not satisfy the sub-additivity property since

$$\mathrm{VaR}_\alpha(\mathcal{L}_1) + \mathrm{VaR}_\alpha(\mathcal{L}_2) < \mathrm{VaR}_\alpha(\mathcal{L}_1 + \mathcal{L}_2).$$

9.3 TAIL VALUE AT RISK

Another major criticism of VaR is that while it tells us the size of the worst loss we can expect to suffer $100 \times \alpha\%$ of the time, it gives us no information on the magnitude of the loss for the other $100(1-\alpha)\%$ of the time. To investigate this we need to use information from the tail of the loss distribution. Specifically, we need access to the spectrum of all VaR figures for confidence levels greater than α, i.e.,

$$\mathrm{VaR}_u(\mathcal{L}_t) \quad \text{for} \quad \alpha \leq u \leq 1.$$

If we have this information we can now propose an alternative risk measure, which we call the Tail Value at Risk (TVaR).

Definition 9.4. *Let* $\mathrm{VaR}_\alpha(\mathcal{L}_t)$ *denote the Value at Risk measure for a portfolio whose loss random variable is given by (9.5). The corresponding Tail Value at Risk* (TVaR) *for the portfolio is taken to be the average of the* $100(1-\alpha)\%$ *worst losses and is given by*

$$\mathrm{TVaR}_\alpha(\mathcal{L}_t) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(\mathcal{L}_t)du. \tag{9.12}$$

The newly proposed TVaR (sometimes referred to as expected shortfall) has arisen from the VaR framework and so practitioners are likely to be comfortable with the underlying idea. The measure itself is much more informative than plain VaR and, what's more, it can be shown that TVaR is a coherent risk measure Acerbi and Tasche (2002). Thus, the two main criticisms of VaR are corrected with TVaR, making it the first genuine candidate risk measure that could, and perhaps should, supercede VaR.

We remark that, in view of (9.7), the expression (9.12) can be written equivalently as

$$\mathrm{TVaR}_\alpha(\mathcal{L}_t) = \frac{1}{1-\alpha}\int_\alpha^1 F^{-1}(u)du.$$

In the case where the distribution is continuous and strictly increasing we know that, for every $u \in [0,1]$, there is a unique x that satisfies $x = F^{-1}(u)$. In view of this we make the following change of variable:

$$\text{set } x = F^{-1}(u) \Rightarrow F(x) = u \Rightarrow dF(x) = du,$$

and we have

$$\begin{aligned}\mathrm{TVaR}_\alpha(\mathcal{L}_t) &= \frac{1}{1-\alpha}\int_{F^{-1}(\alpha)}^{\infty} x dF(x)\\ &= \frac{\int_{\mathrm{VaR}_\alpha}^{\infty} x dF(x)}{\mathbb{P}[\mathcal{L}_t > \mathrm{VaR}_\alpha]}\\ &= \mathbb{E}[\mathcal{L}_t | \mathcal{L}_t > \mathrm{VaR}_\alpha].\end{aligned} \tag{9.13}$$

Thus, in this case, we see that TVaR coincides with the conditional expectation of the loss random variable with $\{\mathcal{L}_t > \mathrm{VaR}_\alpha\}$ being the conditioning event. It can be shown that (9.13) does not hold in general, for instance, in Acerbi and Tasche (2002), an example of a discontinuous distribution is constructed where TVaR_α does not agree with the conditional expectation.

9.4 SPECTRAL RISK MEASURES

We close this chapter with a couple of interesting observations which point the way towards a new family of risk measures. In anticipation of what follows, we provide the following definition:

Definition 9.5. *A function $\phi : [0,1] \to \mathbb{R}$ is said to be a spectral function if it is non-negative and the area under its curve is equal to one; i.e.,*

$$\phi \text{ is a spectral function } \Leftrightarrow \phi(u) \geq 0 \text{ and } \int_{-\infty}^{\infty} \phi(u)du = 1.$$

Spectral functions come in many different guises, in particular we display in Figure 9.4 some examples that are symmetric about the midpoint and exhibit progressively taller and thinner shoulders. If we take this sequence to its limit we reach a curious function which

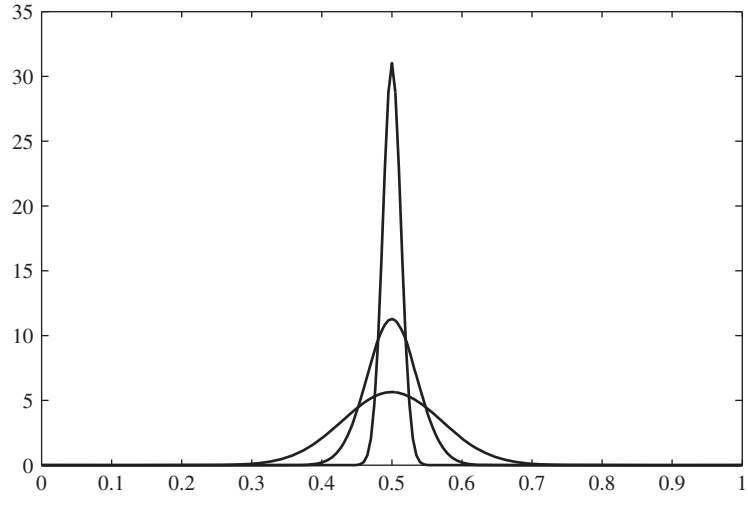


Figure 9.4 A sequence of increasingly thin-shouldered spectral functions.

we call the dirac delta function centred at $1/2$, denoted in this case by $\delta(u - 1/2)$. In more general terms we can define $\delta(u - a)$ to be the dirac delta function centred at any $a \in [0, 1]$. This so-called function has the property

$$\int_0^1 \delta(u-a)du = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} \delta(u-a)f(u)du = f(a) \quad \text{where } a \in [0,1].$$

We remark that $\delta(u - a)$ is not a function in the familiar sense, it is what is known in mathematics as a distribution or a generalized function, however we can imagine the delta function as the result of the limit of a sequence of spectral functions that are increasingly peaked at the point a.

- A representation of VaR_α.
  The reason for introducing the delta function is simply to show that if, for a given $\alpha \in (0, 1)$ we set

$$\phi_{\text{VaR}_\alpha}(u)((\mathcal{L}_t)) = \delta(u - \alpha),$$

  then we can represent VaR as

$$\begin{aligned} \text{VaR}_\alpha(\mathcal{L}_t) &= \int_0^1 \phi_{\text{VaR}_\alpha}(u)\text{VaR}_u(\mathcal{L}_t)du \\ &= \int_0^1 \delta(u-\alpha)\text{VaR}_u(\mathcal{L}_t)du. \end{aligned} \tag{9.14}$$

  We say that ϕ_{VaR_α} is the spectral risk function corresponding to VaR_α.

- A representation of TVaR_α.
  Suppose $\alpha \in (0, 1)$ and consider the function

$$\phi_{\text{TVaR}_\alpha}(u) = \begin{cases} \frac{1}{1-\alpha} & \text{if } u \geq \alpha; \\ 0 & \text{if } u < \alpha. \end{cases}$$

  The function $\phi_{\text{TVaR}_\alpha}$ is spectral as it is non-negative and

$$\int_0^1 \phi_{\text{TVaR}_\alpha}(u)du = \frac{1}{1-\alpha}\int_\alpha^1 du = 1.$$

  Furthermore, using (9.12), we see that

$$\text{TVaR}_\alpha(\mathcal{L}_t) = \int_0^1 \phi_{\text{TVaR}_\alpha}(u)\text{VaR}_u(\mathcal{L}_t)du. \tag{9.15}$$

  We say that $\phi_{\text{TVaR}_\alpha}$ is the spectral risk function corresponding to TVaR_α.
- Towards a family of coherent risk measures.
  In light of the representations (9.14) and (9.15), several researchers have investigated the possibility of defining more general risk measures that are generated by spectral functions, i.e., that have the form

$$\rho(\mathcal{L}_t) = \int_0^1 \phi(u)\text{VaR}_u(\mathcal{L}_t)du \quad \text{for some spectral function } \phi. \tag{9.16}$$

In order for this form of risk measure to prove valuable we need to answer the following puzzle:

Under what conditions is a measure of the form (9.16) a coherent risk measure?

A remarkable solution to this question is provided in Acerbi and Tasche (2002), where the following theorem is established:

Theorem 9.6. *A measure of the form (9.16) is a coherent risk measure whenever the spectral function ϕ is a non-decreasing function, i.e.,*

$$\text{if } \ 0 \leq u_1 \leq u_2 \leq 1 \ \Rightarrow \ \phi(u_1) \leq \phi(u_2).$$

This theorem opens the door to a whole new array of coherent risk measures; collectively these have come to be known as the class of spectral risk measures. There is evidence that investment banks are experimenting with spectral measures; most banks have a VaR calculation engine and so this is a relatively easy task. However, before spectral measures are truly embraced, risk managers need to understand how to select a suitable (non-decreasing) spectral function that, in some sense, is a good choice for the problem. This topic continues to receive a significant amount of academic attention; meanwhile, VaR remains the dominant risk measure in the banking industry.

Essential Mathematics for Market Risk Management
by Simon Hubbert

10

Value at Risk under a Normal Distribution

Most of the development so far has been of a theoretical nature. The practical reality is that we must calculate VaR on a daily basis for our trading portfolio, whose loss random variable is given by

$$\mathcal{L}_t = V(t) \sum_{i=1}^{n} w_i l_i(t). \tag{10.1}$$

The task of the risk manager is to investigate the way $\mathcal{L}_t$ behaves or, more scientifically, he needs to accurately model its distribution function. A glance at the formula tells us that this task is composed of two (interrelated) components:

- How is the random vector $l_t = (l_1(t), \ldots, l_n(t))^T$ of individual daily log loss random variables distributed and how can this information be used to determine the distribution of a linear combination such as $\mathcal{L}_t$?
- How, if at all, does the distribution of $\mathcal{L}_t$ change with time?

The story of modern risk management essentially begins at this point and we are going to navigate a course through the mathematical development. The subject is continually evolving; mathematical models are constructed based upon simplified assumptions of the economy, then as we gain further mathematical insight we can relax the assumptions and improve the modelling. In order to start the calculation of Value at Risk we make the simple assumption that the sequence of random loss vectors $l_t, l_{t+1}, l_{t+2}, \ldots$ possesses the same multivariate normal distribution, i.e., we assume that

$$\boldsymbol{l}_{t+\tau} \sim N(\mathbf{e}, \mathbf{V}),$$

$$\text{or equivalently} \quad \begin{pmatrix} l_1(t+\tau) \\ \vdots \\ l_n(t+\tau) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & & \vdots \\ \sigma_{n1} & \ldots & \sigma_n^2 \end{pmatrix} \right)$$

for $\tau = 0, 1, 2, \ldots$

10.1 CALCULATION OF VALUE AT RISK

A consequence of the normal assumption is that, according to Definition 3.7, the loss random variable (10.1) (which is a linear combination of the components of $\boldsymbol{l}_t$) is normally distributed; we write $\mathcal{L}_t \sim N(\mu, \sigma^2)$, where

$$\mu = V(t)\mathbf{w}^T\mathbf{e} \quad \text{and} \quad \sigma^2 = V(t)^2\mathbf{w}^T\mathbf{V}\mathbf{w}.$$

Its distribution function is given by

$$F(x) = \mathbb{P}[\mathcal{L}_t \leq x] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} \exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right) du.$$

Furthermore, since F is continuous and monotonically increasing, we can deduce that, in this framework, the VaR for the portfolio at a confidence level α is the value VaR_α that satisfies

$$F(\text{VaR}_\alpha) = \mathbb{P}[\mathcal{L}_t \leq \text{VaR}_\alpha] = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\text{VaR}_\alpha} \exp\left(-\frac{1}{2}\left(\frac{u-\mu}{\sigma}\right)^2\right) du$$

or, equivalently, if we set $\mathcal{Z}_t = (\mathcal{L}_t - \mu)/\sigma$ it is the value VaR_α that satisfies

$$\Phi\left(\frac{\text{VaR}_\alpha - \mu}{\sigma}\right) = \mathbb{P}\left[\mathcal{Z}_t \leq \frac{\text{VaR}_\alpha - \mu}{\sigma}\right] = \alpha.$$

The function $\Phi(\cdot)$ is the standard normal distribution function and its values, together with values for its inverse, are well tabulated. In view of this we can take the inverse on both sides of the equation and rearrange to show that

$$\text{VaR}_\alpha = \mu + \sigma\Phi^{-1}(\alpha). \tag{10.2}$$

We can consult the familiar statistical table for the inverse normal distribution to find the value of $\Phi^{-1}(\alpha)$; in the risk management arena we are mostly interested in 95% and 99% confidence levels, for which we have

$$\phi^{-1}(0.95) = 1.65 \quad \text{and} \quad \phi^{-1}(0.95) = 2.33.$$

We can conclude that, under the multivariate normal assumption, the VaR for the portfolio is a straightforward calculation, its value can be expressed in a neat closed form (10.2).

10.2 CALCULATION OF MARGINAL VALUE AT RISK

We can write VaR (10.2) as a function of portfolio weights, as follows:

$$\text{VaR}_\alpha(w_1, \ldots, w_n) = \sum_{i=1}^{n} w_i\mu_i + \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \sigma_{ij}}\,\Phi^{-1}(\alpha). \tag{10.3}$$

We recall that the MVaR of the kth exposure of the portfolio is equal to the partial derivative of (10.3) with respect to w_k, specifically we write

$$\text{MVaR}_\alpha^{(k)} = \frac{\partial}{\partial w_k}\text{VaR}_\alpha(w_1, \ldots, w_n)$$

$$= \frac{\partial}{\partial w_k} \left[\sum_{i=1}^{n} w_i \mu_i + \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij}} \Phi^{-1}(\alpha) \right].$$

We now compute the partial derivatives, the first one is straightforward:

$$\frac{\partial}{\partial w_k} \sum_{i=1}^{n} w_i \mu_i = \mu_k.$$

For the second term we consider the square root function

$$f(x) = \sqrt{x} \quad \text{for} \quad x > 0,$$

then the portfolio volatility can be written as

$$f\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij} \right)$$

and the kth partial derivative is given by

$$f'\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij} \right) \frac{\partial}{\partial w_k} \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij}$$

$$= \frac{\sum_{i=1}^{n} w_i \sigma_{ik}}{\sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \sigma_{ij}}} = \frac{1}{\sigma} \sum_{i=1}^{n} w_i \sigma_{ik}.$$

Hence

$$\mathrm{MVaR}_{\alpha}^{(k)} = \mu_k + \left(\frac{1}{\sigma} \sum_{i=1}^{n} w_i \sigma_{ik} \right) \Phi^{-1}(\alpha).$$

We can now explicitly verify the addition property:

$$\sum_{k=1}^{n} w_k \mathrm{MVaR}_{\alpha}^{(k)} = \sum_{k=1}^{n} w_k \left(\mu_k + \left(\frac{1}{\sigma} \sum_{i=1}^{n} w_i \sigma_{ik} \right) \Phi^{-1}(\alpha) \right)$$

$$= \sum_{k=1}^{n} w_k \mu_k + \left(\frac{1}{\sigma} \underbrace{\sum_{k=1}^{n} w_k \sum_{i=1}^{n} w_i \sigma_{ik}}_{=\sigma^2} \right) \Phi^{-1}(\alpha)$$

$$= \mu + \sigma \Phi^{-1}(\alpha) = \mathrm{VaR}_{\alpha}.$$

10.3 CALCULATION OF TAIL VALUE AT RISK

We recall that the TVaR for a confidence level α is given by (9.12). We can calculate this risk measure explicitly under the normal assumption as follows:

$$\begin{aligned}\mathrm{TVaR}_\alpha(\mathcal{L}_t) &= \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(\mathcal{L}_t)du \\ &= \frac{1}{1-\alpha}\int_\alpha^1 \left(\mu + \sigma\Phi^{-1}(u)\right)du \\ &= \mu + \frac{\sigma}{1-\alpha}\int_\alpha^1 \Phi^{-1}(u)du.\end{aligned}$$

We can transform the above integral by setting

$$x = \Phi^{-1}(u) \quad \Rightarrow \quad \Phi(x) = u.$$

With this substitution we have

$$du = d\Phi(x) = \varphi(x)dx,$$

where φ denotes the probability density function of a standard normal random variable and, from Chapter 2, this function is given explicitly as

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right).$$

The limits of integration transform as follows:

$$\alpha \mapsto \Phi^{-1}(\alpha) \quad \text{and} \quad 1 \mapsto \infty.$$

Thus, the expression for TVaR_α becomes

$$\mathrm{TVaR}_\alpha(\mathcal{L}_t) = \mu + \frac{\sigma}{1-\alpha}\frac{1}{\sqrt{2\pi}}\int_{\Phi^{-1}(\alpha)}^\infty x\exp\left(-\frac{x^2}{2}\right)dx.$$

The above integral can be computed by hand, indeed standard tables of integrals show that, in general, we have

$$\int_a^\infty xe^{-cx^2}dx = \frac{1}{2c}e^{-ca^2}.$$

Applying this result with $c = 1/2$ we find that

$$\begin{aligned}\mathrm{TVaR}_\alpha(\mathcal{L}_t) &= \mu + \frac{\sigma}{1-\alpha}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\Phi^{-1}(\alpha)^2}{2}\right) \\ &= \mu + \frac{\sigma}{1-\alpha}\varphi\left(\Phi^{-1}(\alpha)\right). \qquad (10.4)\end{aligned}$$

Thus we have discovered that, under the normal assumption, the TVaR for the portfolio possesses a neat closed-form solution (10.4).

10.4 SUB-ADDITIVITY OF NORMAL VALUE AT RISK

We recall that one of the main criticisms of VaR is that it is not a sub-additive risk measure. This criticism can be dismissed under the normal assumption, as the following result establishes.

Theorem 1.7. *Let $\mathcal{L}_t^{(1)}$ and $\mathcal{L}_t^{(2)}$ represent the daily loss random variables of two distinct portfolios. Assume that $\mathcal{L}_t^{(1)} \sim N(\mu_1, \sigma_1^2)$ and $\mathcal{L}_t^{(2)} \sim N(\mu_2, \sigma_2^2)$, then*

$$\mathrm{VaR}_\alpha(\mathcal{L}_t^{(1)} + \mathcal{L}_t^{(2)}) \leq \mathrm{VaR}_\alpha(\mathcal{L}_t^{(1)}) + \mathrm{VaR}_\alpha(\mathcal{L}_t^{(2)}).$$

Proof. The VaR for the two portfolios combined is given by

$$\mathrm{VaR}_\alpha(\mathcal{L}_t^{(1)} + \mathcal{L}_t^{(2)}) = (\mu_1 + \mu_2) + \Phi^{-1}(\alpha)\sqrt{\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2},$$

where $\rho \in [-1, 1]$ denotes the correlation coefficient between $\mathcal{L}_t^{(1)}$ and $\mathcal{L}_t^{(2)}$. The above equation is maximized when $\rho = 1$, and this observation allows us to deduce that

$$\begin{aligned}
\mathrm{VaR}_\alpha(\mathcal{L}_t^{(1)} + \mathcal{L}_t^{(2)}) &\leq \mu_1 + \mu_2 + \Phi^{-1}(\alpha)\sqrt{\sigma_1^2 + 2\sigma_1\sigma_2 + \sigma_2^2} \\
&= \mu_1 + \mu_2 + \Phi^{-1}(\alpha)\sqrt{(\sigma_1 + \sigma_2)^2} \\
&= \mu_1 + \mu_2 + \Phi^{-1}(\alpha)(\sigma_1 + \sigma_2) \\
&= \underbrace{\mu_1 + \Phi^{-1}(\alpha)\sigma_1}_{=\mathrm{VaR}_\alpha(\mathcal{L}_t^{(1)})} + \underbrace{\mu_2 + \Phi^{-1}(\alpha)\sigma_2}_{\mathrm{VaR}_\alpha(\mathcal{L}_t^{(2)})}.
\end{aligned}$$

This proves the result. □

The normal framework for portfolio risk management is clearly an appealing one; we have closed-form expressions for VaR_α and TVaR_α, the marginal contributions can easily be derived and we can establish that, in this setting, VaR_α is a coherent risk measure.

Essential Mathematics for Market Risk Management
by Simon Hubbert

11 Advanced Probability Theory for Risk Managers

In Chapter 3 we developed a rudimentary probability theory toolkit consisting of the key concepts and results which, as we have witnessed, has enabled us to set up and solve many practical, day-to-day risk management tasks. However, in addition to these daily tasks, the modern financial risk manager will also be involved in longer-term research projects and many of these will involve large-scale statistical investigations of financial data. Furthermore, it is highly likely that the portfolios under scrutiny will contain a significant number of derivative products, in which case the tools we have developed need to be enhanced. In anticipation of these more demanding tasks we compose here a collection of additional, more advanced tools from probability theory.

11.1 MOMENTS OF A RANDOM VARIABLE

We recall that a random variable X is said to be continuous if its distribution function can be written as

$$F(x) = \mathbb{P}[X \leq x] = \int_{-\infty}^{x} p(u)du,$$

where $p : \mathbb{R} \to [0, \infty)$ is the probability density function of X. The kth moment of the random variable X is defined to be the integral

$$\mu_k := \mathbb{E}[X^k] = \int_{-\infty}^{\infty} x^k p(x)dx, \quad k = 0, 1, 2, \ldots \tag{11.1}$$

We remark that these integrals can be bounded if we replace the power of x by the power of its absolute value $|x|$, i.e., we can write

$$\mu_k \leq \int_{-\infty}^{\infty} |x|^k p(x)dx.$$

The upper bounds on the moments are simply integrals of $p(x)$ multiplied by successively heavier weight functions $|x|$, $|x|^2, \ldots$ The value of any one of these bounding integrals will only be finite if the growth of the weight function, $|x|^k$ say, is sufficiently smothered by the fast decay of the density function $p(x)$. In view of this we say that

$$\text{the moment } \mu_k \text{ exists only if } \int_{-\infty}^{\infty} |x|^k p(x)dx < \infty.$$

We note that the first moment of X is precisely the mean or expectation of X, which we denote simply as μ. This quantity, as we know, serves as a measure of the central location

of the distribution of X. In fact, using μ as a centrality parameter we define the kth central moment of X to be the integral

$$m_k(X) = \mathbb{E}[(X-\mu)^k] = \int_{-\infty}^{\infty} (x-\mu)^k p(x)dx, \quad k = 0, 1, 2, \ldots \tag{11.2}$$

Using this definition it is easy to establish the following simple but useful relation:

$$\begin{aligned} m_k(\alpha X + \beta) &= \int_{-\infty}^{\infty} (\alpha x + \beta - \alpha\mu - \beta)^k p(x)dx \\ &= \int_{-\infty}^{\infty} \alpha^k (x-\mu)^k p(x)dx \\ &= \alpha^k m_k(X), \quad \text{for } \alpha \neq 0,\ \beta \in \mathbb{R} \text{ and } k = 0, 1, 2, \ldots \end{aligned} \tag{11.3}$$

The central moments of a distribution play an important role in characterizing the random variable X. The zeroth central moment is one because it corresponds to the integral of the density function. The first central moment is zero; this follows from the linearity of the expectation operator as

$$m_1(X) = \mathbb{E}[X-\mu] = \mathbb{E}[X] - \mu = 0.$$

The second central moment is precisely the variance of X, which we denote σ^2. To illustrate the explicit computation of the general moments of a distribution we have the following result for normally distributed random variables:

Lemma 11.1. *If $X \in N(\mu, \sigma^2)$, then its moments are given by the formula*

$$m_n(X) = \begin{cases} 0 & \text{if} \quad n = 2k+1,\ k = 0, 1, 2, \ldots; \\ \dfrac{(2k)!}{2^k k!}\sigma^{2k} & \text{if} \quad n = 2k,\ k = 0, 1, 2, \ldots \end{cases} \tag{11.4}$$

Proof. Let $Z = (X-\mu)/\sigma$ denote the standardization of X, then using (11.3) we can write

$$m_n(X) = \sigma^n m_n(Z) = \frac{\sigma^n}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^n e^{-x^2/2} dx. \tag{11.5}$$

Now if $n = 2k+1$ then the function $x \mapsto x^{2k+1}e^{-x^2/2}$ is odd, i.e., it takes on opposite values at x and $-x$, and thus its integral over $\mathbb{R}$ is zero. On the other hand, if $n = 2k$ we can focus on

$$m_{2k}(Z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2k} e^{-x^2/2} dx.$$

To compute this integral we let $u(x) = x^{2k-1}$ and $v(x) = -e^{-x^2/2}$, so that

$$\frac{du}{dx} = (2k-1)x^{2k-2} \quad \text{and} \quad \frac{dv}{dx} = xe^{-x^2/2}.$$

This allows us to integrate by parts, since

$$\begin{aligned} m_{2k}(Z) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u(x) \frac{dv}{dx} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{d}{dx}\Big(u(x)v(x)\Big) dx - \frac{\sigma^{2k}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{du}{dx} v(x) dx. \end{aligned}$$

The first integral appearing in the above expression is zero, since $u(x)v(x) = -x^{2k-1}e^{-x^2/2} \to 0$ as $x \to \pm\infty$, and so we are left with

$$\begin{aligned} m_{2k}(Z) &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{du}{dx} v(x) dx \\ &= (2k-1) \cdot \underbrace{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2k-2} e^{-x^2/2} dx}_{=m_{2k-2}(Z)}. \end{aligned}$$

We can deduce from this that

$$\begin{aligned} m_{2k}(Z) &= (2k-1) m_{2k-2}(Z) \\ &= (2k-1)(2k-3) m_{2k-4}(Z) \\ &= (2k-1)(2k-3)(2k-5) \cdots 3 \cdot 1 \\ &= \frac{(2k)(2k-1)(2k-2) \cdots 3 \cdot 2 \cdot 1}{(2k)(2k-2) \cdots 2} \\ &= \frac{(2k)!}{2^k k!}, \end{aligned}$$

and thus by (11.5) we have

$$m_{2k}(X) = \sigma^{2k} \frac{(2k)!}{2^k k!}.$$

□

Higher moments are often also very useful in finance, in particular the following two:

$$\begin{aligned} \text{skewness } \mathcal{S}(X) &= \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \int_{\mathbb{R}} \left(\frac{x-\mu}{\sigma}\right)^3 p(x) dx; \\ \text{kurtosis } \mathcal{K}(X) &= \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \int_{\mathbb{R}} \left(\frac{x-\mu}{\sigma}\right)^4 p(x) dx. \end{aligned} \tag{11.6}$$

Skewness is an indication of whether the density function leans to the right or to the left of its mean: if $\mathcal{S}(X) > 0$, then X is more likely to exceed μ than to be less than μ, and vice versa. Clearly, if $\mathcal{S}(X) = 0$ then the probability weight is divided equally to the left and the right of the mean, in this case we say that X has a symmetric distribution.

A large kurtosis is an indication that $|X|$ can take large values with relatively high probability. The benchmark for kurtosis calculations is given by the familiar normal random variable where, according to Lemma 11.1, we can deduce that if $X \sim N(\mu, \sigma)$ then $\mathcal{K}(X) = 3$. Thus, we say that a random variable whose kurtosis coefficient is > 3 has a fat-tailed distribution where extreme values both positive and negative are likely to occur relatively often. Analogously, a kurtosis coefficient < 3 indicates a thin-tailed distribution where extreme values are less likely to occur.

11.2 THE CHARACTERISTIC FUNCTION

The characteristic function of a random variable is a highly useful tool in probability theory. In order to define it we need to access a few facts from functional analysis; the branch of mathematics that deals with spaces of functions and how they behave under certain operations. The function space that we are most interested in is the space of integrable functions:

$$f : \mathbb{R} \to \mathbb{R} \text{ is integrable if } \int_{\mathbb{R}} |f(x)|dx < \infty. \tag{11.7}$$

To signify that a function is integrable, we write $f \in L_1(\mathbb{R})$. The operator that we are most interested in is the Fourier transform and this is defined, for all integrable functions, by the formula

$$(\mathcal{F}f)(\boldsymbol{\xi}) = \widehat{f}(\xi) = \int_{\mathbb{R}} \exp(-ix\xi) f(x)dx \quad \text{for} \quad \xi \in \mathbb{R}.$$

We recall that for any real number θ the quantity $\exp(-i\theta)$ represents the complex number $z = \cos\theta - i\sin\theta$.

It is well known that we can recover $f \in L_1(\mathbb{R})$ from its Fourier transform via the inversion formula

$$f(x) = (\mathcal{F}^{-1}\widehat{f})(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \exp(i\xi x)\widehat{f}(\xi)d\xi. \tag{11.8}$$

Another nice property of Fourier theory establishes that differentiation can be viewed as multiplication:

Proposition 11.2. *If f and its derivative f' are both integrable functions (on $\mathbb{R}$) such that $f(x) \to 0$ as $x \to \pm\infty$, then the following identities hold:*

$$\widehat{\frac{df}{dx}}(\xi) = i\xi\,\widehat{f}(\xi) \quad \text{and} \quad \frac{d}{d\xi}\widehat{f}(\xi) = -i(\widehat{xf})(\xi). \tag{11.9}$$

Proof. The first identity is established by integrating by parts:

$$\begin{aligned}
\widehat{\frac{df}{dx}}(\xi) &= \int_{\mathbb{R}} f'(x)e^{-i\xi x}dx \\
&= -\int_{\mathbb{R}} f'(x)\left(-i\xi e^{-i\xi x}\right)dx \\
&= i\xi\,\widehat{f}(\xi).
\end{aligned}$$

The second identity is established by differentiating under the integral sign:

$$\begin{aligned}\frac{d}{d\xi}\widehat{f}(\xi) &= \int_{\mathbb{R}} f(x)\frac{d}{d\xi}e^{-i\xi x}dx \\ &= \int_{\mathbb{R}} -ixf(x)e^{-i\xi x}dx \\ &= -i(\widehat{xf})(\xi).\end{aligned}$$

□

A probability density function is a typical example of an integrable function and it turns out that its Fourier transform is a highly useful tool for discovering further probabilistic properties. To fix the notation for future investigations we provide the following definition:

Definition 11.3. *The characteristic function of a continuous random variable X with density function $p(x)$ is defined to be*

$$\phi_X(u) = \mathbb{E}[\exp(iuX)] = \int_{\mathbb{R}} p(x)\exp(ixu)dx = \widehat{p}(-u). \tag{11.10}$$

Now, using the Fourier inversion formula transform (11.8), we can immediately deduce that the characteristic function uniquely determines the density function of X, since

$$p(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \phi_X(u)\exp(-iux)du, \quad \text{for } x \in \mathbb{R}.$$

In addition, the distribution function F of the X can also be extracted from its characteristic function via the famous Gil-Pelaez formula (see Gil-Pelaez (1951)),

$$F(x) = \frac{1}{2} + \frac{1}{2\pi}\int_0^\infty \frac{\exp(iux)\phi_X(-u) - \exp(-iux)\phi_X(u)}{iu}du. \tag{11.11}$$

The following intriguing property, which follows from Proposition (11.2), establishes a link between the smoothness of ϕ_X and the number and size of the moments of X:

Lemma 11.4. *Let X denote a continuous random variable for which only the first n moments exist (i.e., are finite). Then the corresponding characteristic function ϕ_X is n-times differentiable and*

$$\frac{d^k\phi_X}{du^k}(0) = i^k\mu_k, \quad \text{for} \quad k = 1, 2, \ldots, n. \tag{11.12}$$

Proof. We begin by computing the first derivative, which is given by

$$\begin{aligned}\frac{d\phi_X}{du} &= \lim_{h\to 0}\frac{\phi_X(u+h) - \phi_X(u)}{h} \\ &= \lim_{h\to 0}\mathbb{E}\left[\frac{\exp(i(u+h)X) - \exp(ihX)}{h}\right]\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\lim_{h\to 0} \frac{\exp(i(u+h)X) - \exp(ihX)}{h}\right] \\
&= \mathbb{E}\left[\frac{d}{du}\exp(iuX)\right] \\
&= i\mathbb{E}\left[X\exp(iuX)\right].
\end{aligned}
$$

We remark that we have glossed over the justification of the exchange of the expectation operator and the limit $h \to 0$ in the above development; we leave this technical point aside, however it can be shown that this is perfectly legal since $\mathbb{E}[|X|] < \infty$.

We can now set $u = 0$ to deduce that

$$
\frac{d\phi_X}{du}(0) = i\mathbb{E}[X] = i\mu_1.
$$

The other derivatives can be computed in a similar fashion:

$$
\begin{aligned}
\frac{d^2\phi_X}{du^2} &= \frac{d}{du}\mathbb{E}\left[iX\exp(iuX)\right] \\
&= \mathbb{E}\left[iX\frac{d}{du}\exp(iuX)\right] \\
&= \mathbb{E}\left[i^2X^2\exp(iuX)\right],
\end{aligned}
$$

and so

$$
\frac{d^2\phi_X}{du^2}(0) = i^2\mathbb{E}\left[X^2\right] = i^2\mu_2,
$$

and so on. □

We are now in a position to demonstrate the effectiveness of the characteristic function with the following examples.

11.2.1 Dealing with the sum of several random variables

Let us assume that we have two continuous, independent random variables, X and Y say, whose probability density functions are given by p_X and p_Y. The sum of these two variables, $Z = X + Y$ say, is itself a random variable in its own right. The characteristic function of Z has a particularly nice representation, as the following development reveals:

$$
\begin{aligned}
\phi_Z(u) &= \mathbb{E}[\exp(iu(X+Y))] \\
&= \mathbb{E}[\exp(iuX)\exp(iuY)] \\
&= \mathbb{E}[\exp(iuX)] \cdot \mathbb{E}[\exp(iuY)] \quad \text{(due to independence)} \\
&= \phi_X(u) \cdot \phi_Y(u).
\end{aligned}
$$

Of course, this result is not restricted to the sum of two random variables and we have the following generalization:

Theorem 11.5. *Let $X_1, \ldots, X_n$ denote n independent continuous random variables whose characteristic functions are denoted by $\phi_1, \ldots, \phi_n$ respectively. The characteristic function of the random variable $Z = X_1 + \cdots X_n$ is given by*

$$\phi_Z(u) = \phi_1(u) \cdots \phi_n(u).$$

11.2.2 Dealing with a scaling of a random variable

In the previous example we took a collection of independent, continuous random variables and considered their sum as a new random variable. We encountered expressions for both the density and the characteristic functions for this sum. In this example we take only one continuous random variable X and we generate a new variable by applying a scale factor, i.e., we consider αX for some $\alpha > 0$. In this situation we can demonstrate the following links:

- If p_X and $p_{\alpha X}$ denote the density functions of X and αX respectively, then

$$p_{\alpha X}(u) = \frac{1}{\alpha} p_X\left(\frac{u}{\alpha}\right). \tag{11.13}$$

  This follows since

$$\begin{aligned}\int_a^b p_{\alpha X}(u)du = \mathbb{P}[a \le \alpha X \le b] &= \mathbb{P}\left[\frac{a}{\alpha} \le X \le \frac{b}{\alpha}\right] \\ &= \int_{a/\alpha}^{b/\alpha} p_X(x)dx \quad \text{(now set } u = \alpha x\text{)} \\ &= \int_a^b \frac{1}{\alpha} p_X\left(\frac{u}{\alpha}\right) du.\end{aligned}$$

- If ϕ_X and $\phi_{\alpha X+\beta}$ denote the respective characteristic functions of X and $\alpha X + \beta$, then we have

$$\begin{aligned}\phi_{\alpha X+\beta}(u) &= \mathbb{E}[\exp(iu(\alpha X + \beta))] \\ &= \exp(iu\beta)\mathbb{E}[\exp(iu\alpha X] = \exp(iu\beta)\phi_X(\alpha u).\end{aligned} \tag{11.14}$$

11.2.3 Normally distributed random variables

Characteristic functions are extremely useful objects, and as such it is useful to be equipped with some examples. For our purposes the most commonly used example is that of a standard normal random variable $Z \sim N(0, 1)$ and we present its derivation here. As a first step we let $\alpha \in \mathbb{R}$ and compute the following more general expectation:

$$\begin{aligned}\mathbb{E}[\exp(\alpha Z)] &= \int_{\mathbb{R}} \exp(\alpha z) \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)}_{\text{pdf of } N(0,1)} dz \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\alpha)^2 - \alpha^2}{2}\right) dz\end{aligned}$$

$$= \exp\left(\frac{\alpha^2}{2}\right) \int_{\mathbb{R}} \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(z-\alpha)^2}{2}\right)}_{\text{pdf of } N(\alpha,1) \;\Rightarrow\; \text{integral} = 1} dz \tag{11.15}$$

$$= \exp\left(\frac{\alpha^2}{2}\right).$$

It can be shown, using properties of functions of a complex variable, that the identity (11.15) holds if we allow α to be a complex number. In view of this fact we can set $\alpha = iu$, where $u \in \mathbb{R}$, and observe that the left-hand side of (11.15) is the characteristic function for $Z \in N(0, 1)$ and we have

$$\phi_Z(u) = \mathbb{E}[\exp(iuZ)] = \exp\left(-\frac{u^2}{2}\right).$$

We can now use this calculation to derive the characteristic function for an arbitrary normal random variable $X \sim N(\mu, \sigma^2)$. The idea is to express X as $\mu + \sigma Z$ where $Z \sim N(0, 1)$, and then appeal to (11.14) to deduce

$$\phi_X(u) = \exp(iu\mu)\phi_Z(\sigma u) = \exp\left(iu\mu - \frac{\sigma^2 u^2}{2}\right). \tag{11.16}$$

Let us suppose we have n univariate random variables $(X_i)_{i=1}^n$ that are known to be normally distributed. An interesting problem is to investigate under what conditions we can guarantee that the sum $S_n = X_1 + \cdots + X_n$ is also a normal random variable. Obviously we cannot conclude that S_n will always be normally distributed, for instance if X is normal and we define $Y = -X$ then $X + Y = 0$, which is not normally distributed. We can however, with the help of the characteristic function (11.16), establish the following result:

Theorem 11.6. *Let $(X_k)_{k=1}^n$ denote a sequence of **independent** normal random variables, i.e.,*

$$X_k \sim N(\mu_k, \sigma_k^2) \quad \text{for} \quad k = 1, \ldots, n.$$

The sum $S_n = X_1 + \cdots + X_n$ is also a normal random variable and, in fact, we have

$$S_n = X_1 + \cdots + X_n \sim N(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2).$$

Proof. The characteristic function of X_k is given by

$$\phi_k(u) = \exp\left(iu\mu_k - \frac{\sigma_k^2 u^2}{2}\right) \quad \text{for} \quad k = 1, \ldots, n.$$

Since the random variables are independent we can evoke Theorem 11.5 to deduce that the characteristic function of S_n is given by

$$\phi_{S_n}(u) = \exp\left(iu(\mu_1 + \cdots + \mu_n) - \frac{(\sigma_1^2 + \cdots + \sigma_n^2)u^2}{2}\right).$$

Thus, we can conclude $S_n \sim N(\mu_1 + \cdots + \mu_n, \sigma_1^2 + \cdots + \sigma_n^2)$, as required. □

11.3 THE CENTRAL LIMIT THEOREM

In order to demonstrate the power of working with characteristic functions, we embark here on an intriguing investigation which will lead us to one of the most famous of all probability theory results; the central limit theorem. We begin with the following crucial theorem:

Theorem 11.7. *For every $n = 1, 2, \ldots$, let X_n be a random variable with distribution function F_n and characteristic function ϕ_n. Let X denote a random variable with distribution function F and characteristic function ϕ_X. If $\phi_n(u) \to \phi(u)$ exists for all u (as $n \to \infty$) and ϕ is continuous at $u = 0$, then we have that*

$$F_n(x) \to F(x) \text{ as } n \to \infty$$

for every value of x at which F is continuous. We say that X_n converges to X in distribution.

The proof of this result is rather technical and we do not present it here, although the interested reader can consult Section III of Berger (1992). We observe that the result itself suggests that, by examining the behaviour of a sequence of characteristic functions, we might be able to explain the limiting behaviour of the corresponding sequence of random variables. With this result at our disposal we can now present and prove the famous central limit theorem.

Central Limit Theorem 11.8. *Let $(X_n)_{n=1}^{\infty}$ denote a sequence of independent and identically distributed random variables with mean μ and finite variance σ^2. Let*

$$A_n = \frac{X_1 - \mu}{\sigma\sqrt{n}} + \cdots + \frac{X_n - \mu}{\sigma\sqrt{n}},$$

then

$$\lim_{n\to\infty} A_n \sim N(0, 1).$$

Proof. We shall assume, without loss of generality, that each random variable has zero mean and unit variance (otherwise we could just work with the standardization $Y_k = (X_k - \mu)/\sigma$) and so, as a result, we investigate

$$A_n = \frac{X_1 + \cdots + X_n}{\sqrt{n}}.$$

As all random variables have the same distribution they also share the same characteristic function, which we shall denote as ϕ. Furthermore, since the random variables are independent we can evoke Theorem 11.5 to deduce that

$$\phi_{A_n}(u) = \phi^n\left(\frac{u}{\sqrt{n}}\right).$$

Now, a local Taylor expansion of $\phi(u)$ gives

$$\phi(u) = \phi(0) + \phi'(0)u + \phi''(0)\frac{u^2}{2!} + O(u^3).$$

We know, by definition, that $\phi(0) = 1$ and, in addition, Lemma 11.4 tells us that

$$\phi'(0) = i\mu_1 = 0 \quad \text{and} \quad \phi''(0) = i^2\mu_2 = -1.$$

Thus, as a result, the Taylor approximation collapses to

$$\phi(u) = 1 - \frac{u^2}{2} + O(u^3),$$

and so

$$\phi_{A_n}(u) = \left[1 - \frac{u^2}{2n} + O\left(\left(\frac{u}{\sqrt{n}}\right)^3\right)\right]^n.$$

A standard result from calculus tells us that if $(\alpha_n)_{n=1}^{\infty}$ is a convergent sequence of real numbers such that $\alpha_n \to \alpha$ as $n \to \infty$, then

$$\left(1 + \frac{\alpha_n}{n}\right)^n \to \exp(\alpha) \quad \text{as } n \to \infty.$$

With this result in mind we notice that we can write the expression for ϕ_{A_n} in a neater form by letting

$$\alpha_n = -\frac{u^2}{2} + O\left(\frac{u^3}{\sqrt{n}}\right) \quad \text{so that} \quad \phi_{A_n}(u) = \left(1 + \frac{\alpha_n}{n}\right)^n.$$

We notice that, in our case, $\alpha_n \to -u^2/2$ and so we can conclude that

$$\lim_{n\to\infty} \phi_{A_n}(u) = \lim_{n\to\infty} \left(1 + \frac{\alpha_n}{n}\right)^n = \exp\left(-\frac{u^2}{2}\right).$$

We recognise this limit to be the characteristic function for a standard normal random variable and this observation, combined with Theorem 11.7, completes the proof. □

We remark that an analogous version of the central limit theorem holds in the multivariate setting too. We state it here for completeness.

CLT Multivariate Version 11.9. *Let $(\boldsymbol{X}_k)_{k=1}^{\infty}$ denote a sequence of independent and identically distributed d-dimensional random vectors, i.e.,*

$$\boldsymbol{X}_k = (X_{1k}, \ldots X_{dk})^T \in \mathbb{R}^d \quad \text{for} \quad k = 1, 2, \ldots$$

We let

$$\boldsymbol{e} = \mathbb{E}[\boldsymbol{X}_k] \quad \text{and} \quad \boldsymbol{V} = \mathbb{E}[(\boldsymbol{X}_k - \boldsymbol{e})(\boldsymbol{X}_k - \boldsymbol{e})^T], \quad k = 1, 2, \ldots$$

denote the common mean vector and covariance matrix respectively. We now define the nth sample mean vector as

$$\overline{\boldsymbol{X}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{X}_k, \quad \text{for} \quad n = 1, 2, \ldots,$$

then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{e}) \to \boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{V}) \quad \text{as} \quad n \to \infty.$$

We say that $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{e})$ *converges in distribution to a multivariate normal random vector* $\boldsymbol{Z}$.

11.4 THE MOMENT-GENERATING FUNCTION

A special case of the characteristic function of a random variable X occurs when we consider evaluating it across the imaginary axis; this enables us to define a new function via the following formula:

$$\begin{aligned}\psi_X(u) = \phi_X(-iu) &= \mathbb{E}[\exp(-i^2 uX)] \\ &= \mathbb{E}[\exp(uX)] \\ &= \int_{\mathbb{R}} p(x)\exp(ux)dx.\end{aligned}$$

We must take care when we use this function; we can see from the definition that

$$\psi_X(u) \text{ is well defined at } u \in \mathbb{R} \text{ provided that } \mathbb{E}[\exp(uX)] < \infty.$$

To investigate this statement further we use the series expansion of the exponential function (about the origin 0) to give the following representation:

$$\begin{aligned}\psi_X(u) &= \mathbb{E}\left[1 + uX + \frac{u^2X^2}{2!} + \frac{u^3X^3}{3!} + \cdots\right] \\ &= 1 + u\mathbb{E}[X] + \frac{u^2}{2!}\mathbb{E}[X^2] + \frac{u^3}{3!}\mathbb{E}[X^3] + \cdots \\ &= \sum_{k=0}^{\infty}\frac{u^k}{k!}\mathbb{E}[X^k] = \sum_{k=0}^{\infty}\left(\frac{u^k}{k!}\right) \times (k\text{th moment of } X).\end{aligned} \tag{11.17}$$

We can conclude from this that

$$\begin{gathered}\psi_X(u) \text{ is well defined for } u \in (-\delta, \delta) \text{ for some } \delta > 0, \\ \text{if and only if all moments of } X \text{ are finite.}\end{gathered} \tag{11.18}$$

In this case we can compare the Taylor expansion (about zero) of ψ_X,

$$\psi_X(u) = \sum_{k=0}^{\infty}\frac{u^k}{k!}\psi_X^{(k)}(0),$$

with (11.17) to reveal that the moments of X can be computed via ψ_X as

$$\mu_k = \mathbb{E}[X^k] = \psi_X^{(k)}(0), \quad \text{for} \quad k = 0, 1, \ldots$$

In view of this special property we call ψ_X the moment-generating function for the random variable X. Given that we have found the characteristic function for a normal random variable $X \sim N(\mu, \sigma^2)$ (11.16), we can now also easily access its moment-generating function, it is given by

$$\psi_X(u) = \phi_X(-iu) = \exp\left(u\mu + \frac{\sigma^2 u^2}{2}\right). \tag{11.19}$$

11.5 THE LOG-NORMAL DISTRIBUTION

A random variable Y is said to be log-normally distributed if its logarithm is normally distributed. Alternatively, if $X \sim N(\mu, \sigma)$ then the random variable Y defined by

$$Y = \exp(X)$$

is log-normally distributed.

In view of this definition we can immediately see that log-normally distributed random variables are non-negative, but what of their other properties?

We begin our investigation by considering the formal distribution of a random variable $Y = \exp(X)$, where $X \sim N(\mu, \sigma^2)$; for any $y \geq 0$ we can write

$$\begin{aligned} F(y) = \mathbb{P}[Y \leq y] &= \mathbb{P}[\exp(X) \leq y] \\ &= \mathbb{P}[X \leq \log(y)] \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\log(y)} \exp\left(\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx. \end{aligned}$$

We now let

$$u = \exp(x), \text{ i.e., } x = \log(u) \Rightarrow dx = \frac{du}{u}$$

to transform the integral to give

$$F(y) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} \exp\left(\frac{1}{2}\left(\frac{\log(u)-\mu}{\sigma}\right)^2\right) \frac{du}{u}.$$

This allows us to deduce that the density function of Y is given by

$$\begin{aligned} p(y) &= \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log(y)-\mu}{\sigma}\right)^2\right) \\ &= \frac{1}{y\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log\left(\frac{y}{\exp(\mu)}\right)}{\sigma}\right)^2\right). \end{aligned} \tag{11.20}$$

The moments of the log-normal distribution have a familiar representation,

$$\mu_k = \mathbb{E}[Y^k] = \mathbb{E}[\exp(kX)].$$

We recognize the right-hand side of this expression to be the moment-generating function $\psi(u)$ of $X \sim N(\mu, \sigma^2)$ evaluated at $u = k$. Appealing to equation (11.19), we have

$$\mu_k = \exp\left(k\mu + \frac{\sigma^2 k^2}{2}\right). \tag{11.21}$$

We can use this discovery to compute the first four key central moments of the log-normal distribution:

- The mean is given by

$$\mathbb{E}[Y] = \mu_1 = \exp\left(\mu + \frac{\sigma^2}{2}\right). \tag{11.22}$$

- The variance is given by

$$\mathbb{E}[(Y - \mu_1)^2] = \mu_2 - \mu_1^2 = \exp\left(2\mu + \sigma^2\right)\Big[\exp(\sigma^2) - 1\Big]. \tag{11.23}$$

- The third central moment is given by

$$\begin{aligned}\mathbb{E}[(Y-\mu)^3] &= \mathbb{E}[Y^3] - 3\mathbb{E}[Y^2]\mu + 2\mu^3 \\ &= \exp\left(3\mu + \frac{3\sigma^2}{2}\right)\Big[\exp(3\sigma^2) - 3\exp(\sigma^2) + 2\Big] \\ &= \exp\left(3\mu + \frac{3\sigma^2}{2}\right)\Big(\exp(\sigma^2) - 1\Big)^2\Big(\exp(\sigma^2) + 2\Big)\end{aligned}$$

  and so the skewness coefficient is

$$S(Y) = (\exp(\sigma^2) + 2)\sqrt{\exp(\sigma^2) - 1}. \tag{11.24}$$

- The fourth central moment is given by

$$\begin{aligned}&\mathbb{E}[(Y-\mu)^4] = \mathbb{E}[Y^4] - 4\mathbb{E}[Y^3]\mu + 6\mathbb{E}[Y^2]\mu^2 - 3\mu^4 \\ &= \exp\left(4\mu + 2\sigma^2\right)\Big[\exp(6\sigma^2) - 4\exp(3\sigma^2) + 6\exp(\sigma^2) - 3\Big] \\ &= \exp\left(4\mu + 2\sigma^2\right)\Big(\exp(\sigma^2) - 1\Big)^2\Big(\exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 3\Big)\end{aligned}$$

  and so the kurtosis coefficient is

$$\mathcal{K}(X) = \exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 3. \tag{11.25}$$

Essential Mathematics for Market Risk Management
by Simon Hubbert

12 A Survey of Useful Distribution Functions

In this chapter we introduce the reader to a collection of probability distributions that will prove useful in different areas of risk management, statistical techniques and financial modelling.

12.1 THE GAMMA DISTRIBUTION

In mathematics there exist a whole host of so-called special functions. These functions possess useful properties that are frequently exploited to simplify the solutions of seemingly difficult problems. In probability theory one of the most commonly used special functions is the Gamma function, denoted by $\Gamma(\cdot)$ and given by

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx, \quad \text{for} \quad \alpha > 0. \tag{12.1}$$

If we set

$$u(x) = \frac{x^\alpha}{\alpha} \quad \text{and} \quad v(x) = e^{-x},$$

then we have

$$u'(x) = x^{\alpha-1} \quad \text{and} \quad v(x) = -e^{-x}.$$

This allows us to integrate by parts to deduce that

$$\begin{aligned}
\Gamma(\alpha) &= \int_0^\infty u'(x)v(x)dx \\
&= \int_0^\infty \frac{d}{dx}\left(u(x)v(x)\right) dx - \int_0^\infty u(x)v'(x)dx \\
&= \underbrace{\Big[u(x)v(x)\Big]_{x=0}^{x\to\infty}}_{=0} + \int_0^\infty \frac{x^\alpha}{\alpha} e^{-x} dx = \frac{\Gamma(\alpha+1)}{\alpha}.
\end{aligned}$$

We have thus discovered that the Gamma function satisfies the property

$$\Gamma(\alpha+1) = \alpha\Gamma(\alpha), \quad \alpha > 0. \tag{12.2}$$

Furthermore, since $\Gamma(1) = 1$, we can deduce that, if $\alpha = n$ is a non-negative integer then, using (12.2), we have

$$\Gamma(n+1) = n\Gamma(n) = n(n-1)\Gamma(n-1) = \cdots = n!\Gamma(1) = n!$$

i.e., the Gamma function extends the familiar factorial function from non-negative integers to all positive real numbers.

The recursive property (12.2) is extremely helpful if we need to compute the Gamma function at special values. For instance, given the value of $\Gamma(1/2)$ we can easily derive the values of $\Gamma(n+1/2)$ for all non-negative integers n. We note that

$$\begin{aligned}\Gamma(1/2) &= \int_0^\infty \frac{e^{-x}}{\sqrt{x}}dt \quad \left(\text{now set} \quad x = \frac{y^2}{2} \Rightarrow dx = ydy\right) \\ &= \frac{\sqrt{2}}{2}\int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2}\right)dy = \frac{\sqrt{2}}{2}\sqrt{2\pi}\int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)}_{\text{pdf of } N(0,1)}dy \\ &= \sqrt{\pi}.\end{aligned}$$

We can derive alternative expressions for the Gamma function by manipulating the integral (12.1) that defines it. For example, taking the following change of variables:

$$x \mapsto \frac{x}{\lambda} \quad \text{for} \quad \lambda > 0,$$

the expression (12.1) becomes

$$\Gamma(\alpha) = \int_0^\infty \frac{x^{\alpha-1}e^{-x/\lambda}}{\lambda^\alpha}dx, \quad \text{for} \quad \alpha > 0.$$

Using this scaled expression we can define, for a fixed $\alpha > 0$, the following function:

$$p(x) = \begin{cases} 0 & \text{for} \quad x < 0; \\ \dfrac{x^{\alpha-1}e^{-x/\lambda}}{\Gamma(\alpha)\lambda^\alpha} & \text{for} \quad x \geq 0. \end{cases}$$

The integral of this function across $\mathbb{R}$ equals one, furthermore the function

$$F(x) = \int_{-\infty}^{x} p(u)du = \int_0^x \frac{u^{\alpha-1}e^{-u/\lambda}}{\Gamma(\alpha)\lambda^\alpha}du \tag{12.3}$$

is a continuous, strictly increasing function from 0 (at $x < 0$) to 1 (as $x \to \infty$), thus F is a distribution function and its derivative p is its associated probability density function. In view of this we provide the following definition:

Definition 12.1. *A non-negative random variable X is said to be Gamma distributed if it satisfies*

$$\mathbb{P}[X \leq x] = \int_0^x \frac{u^{\alpha-1}e^{-u/\lambda}}{\Gamma(\alpha)\lambda^\alpha}du \quad \text{for} \quad \text{all } x \geq 0.$$

The positive parameters α and λ define, respectively, the shape and scale of the distribution.

We say that F, defined by (12.3), is the Gamma distribution with shape parameter α and scale parameter λ. The characteristic function for a Gamma distributed random variable, which we write as $X \sim \Gamma(\alpha, \lambda)$, is calculated as follows:

$$\begin{aligned}\phi_X(u) &= \int_0^\infty \exp(iux)\frac{x^{\alpha-1}\exp(-x/\lambda)}{\lambda^\alpha \Gamma(\alpha)}dx \\ &= \frac{1}{\lambda^\alpha\Gamma(\alpha)}\int_0^\infty \exp\left(x\left(iu-\frac{1}{\lambda}\right)\right)x^{\alpha-1}dx.\end{aligned}$$

We now employ the substitution

$$v = -x\left(iu - \frac{1}{\lambda}\right)$$

and the integral becomes

$$\begin{aligned}\phi_X(u) &= \frac{1}{\lambda^\alpha\Gamma(\alpha)}\int_0^\infty \exp(-v)\left(\frac{v}{-\left(iu-\frac{1}{\lambda}\right)}\right)^{\alpha-1}\frac{dv}{-\left(iu-\frac{1}{\lambda}\right)} \\ &= \frac{1}{(1-i\lambda u)^\alpha}\frac{1}{\Gamma(\alpha)}\underbrace{\int_0^\infty \exp(-v)v^{\alpha-1}dv}_{=\Gamma(\alpha)\text{ by (12.1)}} \\ &= \frac{1}{(1-i\lambda u)^\alpha}.\end{aligned} \tag{12.4}$$

The moment-generating function of X is then given by

$$\psi_X(u) = \phi_X(-iu) = \frac{1}{(1-\lambda u)^\alpha}.$$

The first derivative of ψ_X is given by

$$\psi_X'(u) = \frac{\alpha\lambda}{(1-\lambda u)^{\alpha+1}}$$

and, evaluating this at zero, we can extract the mean of X, i.e.,

$$\mathbb{E}[X] = \psi_X'(0) = \alpha\lambda.$$

In general we can show that

$$\mathbb{E}[X^k] = \psi_X^{(k)}(0) = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)}\lambda^k.$$

Using these values we can now compute that the variance of $X \sim \Gamma(\alpha, \lambda)$ is given by

$$\mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2] - \mu^2 = \alpha(\alpha+1)\lambda^2 - \alpha^2\lambda^2 = \alpha\lambda^2.$$

In a similar manner, we can also show that the third and fourth central moments are given by

$$\begin{aligned}\mathbb{E}[(X-\mu)^3] &= \mathbb{E}[X^3] - 3\mathbb{E}[X^2]\mu + 2\mu^3 \\ &= 2\alpha\lambda^3\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[(X-\mu)^4] &= \mathbb{E}[X^4] - 4\mathbb{E}[X^3]\mu + 6\mathbb{E}[X^2]\mu^2 - 3\mu^4 \\ &= 3\alpha(2+\alpha)\lambda^4.\end{aligned}$$

respectively. Hence, using these two equations, the skewness and kurtosis coefficients for $X \sim \Gamma(\alpha, \lambda)$ are given by

$$\mathcal{S}(X) = \frac{2}{\sqrt{\alpha}} \quad \text{and} \quad \mathcal{K}(X) = 3 + \frac{6}{\alpha} \text{ respectively.} \tag{12.5}$$

12.2 THE CHI-SQUARED DISTRIBUTION

To motivate the discussion of this distribution we begin by considering a random vector $\mathbf{X} = (X_1, \ldots, X_n)^T \in \mathbb{R}^n$ whose components are independent zero-mean normal random variables that share the same variance, i.e., $X_k \sim N(0, \sigma^2)$ for $k = 1, \ldots, n$. We recall that we have already shown (Theorem 11.6) that the sum $S_n = \mathbf{1}^T\mathbf{X} = X_1 + \cdots + X_n$ is also normally distributed, specifically $S_n \sim N(0, n\sigma^2)$. We are now going to investigate the distribution of the sum of squares, i.e., we ask

$$\text{how is the random variable } Z_n = \mathbf{X}^T\mathbf{X} = X_1^2 + \cdots + X_n^2 \text{ distributed?}$$

We set about this question by considering the distribution of the square of a single normal random variable, i.e., we let $X \sim N(0, \sigma)$ and define $Z_1 = X^2$. We note that, by definition, Z_1 is a non-negative random variable and so, for any $z > 0$, we can formally write its distribution as

$$\begin{aligned}F(z) = \mathbb{P}[Z_1 \le z] &= \mathbb{P}[-\sqrt{z} \le X \le \sqrt{z}] \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\sqrt{z}}^{\sqrt{z}} \exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2\right) dx.\end{aligned}$$

Setting $u = x/\sigma$ we find that

$$\begin{aligned}F(z) &= \frac{1}{\sqrt{2\pi}} \int_{-\sqrt{z}/\sigma}^{\sqrt{z}/\sigma} \exp\left(-\frac{u^2}{2}\right) du \\ &= \Phi\left(\frac{\sqrt{z}}{\sigma}\right) - \Phi\left(-\frac{\sqrt{z}}{\sigma}\right).\end{aligned}$$

We can find the density function of Z_1 by differentiating its distribution with respect to z, this gives

$$\begin{aligned} p(z) &= \frac{z^{-1/2}}{2\sigma}\Phi'\left(\frac{\sqrt{z}}{\sigma}\right) + \frac{z^{-1/2}}{2\sigma}\Phi'\left(-\frac{\sqrt{z}}{\sigma}\right) \\ &= \frac{z^{-1/2}}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\sqrt{z}}{\sigma}\right)^2\right) + \frac{z^{-1/2}}{2\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(-\frac{\sqrt{z}}{\sigma}\right)^2\right) \\ &= \frac{z^{-1/2}}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\sqrt{z}}{\sigma}\right)^2\right). \end{aligned}$$

We can conclude therefore that the distribution function of Z_1 can be written as

$$F(z) = \int_0^z \frac{u^{-1/2}e^{-u/2\sigma^2}}{\sqrt{\pi}\sqrt{2\sigma^2}}du.$$

If we set $\lambda = 2\sigma^2$ and employ the Gamma function identity $\Gamma(1/2) = \sqrt{\pi}$ we can simplify the above expression to give

$$F(z) = \int_0^z \frac{u^{-1/2}e^{-u/\lambda}}{\Gamma(1/2)\lambda^{1/2}}du.$$

Comparing this expression to (12.3) we discover that Z_1 is a special case of a Gamma distributed random variable, specifically we could write $Z_1 \sim \Gamma(1/2, 2\sigma^2)$. In order to emphasize the importance of this discovery, we say that Z_1 has the chi-squared distribution and we write

$$Z_1 \sim \chi_1^2(\sigma^2) = \Gamma(1/2, 2\sigma^2).$$

Using this discovery we can immediately deduce, using equation (12.4), that the characteristic function of Z_1 is given by

$$\phi_{Z_1}(u) = \frac{1}{(1-2i\sigma^2u)^{1/2}}.$$

Armed with this we can appeal directly to Theorem 11.5 to deduce that the characteristic function of the more general random variable $Z_n = X_1^2 + \cdots + X_n^2$ is given by

$$\phi_{Z_n}(u) = \frac{1}{(1-2i\sigma^2u)^{n/2}},$$

which we recognize (see (12.4)) as the characteristic function of a random variable with the $\Gamma(n/2, 2\sigma^2)$ distribution. We can now provide a full answer to the original question:

$$\begin{gathered} \text{if } Z_n = X_1^2 + \cdots + X_n^2 \text{ where } (X_k)_{k=1}^n \text{ are independent and } \sim N(0, \sigma^2) \\ \text{then } Z_n \sim \Gamma(n/2, 2\sigma^2) = \chi_n^2(\sigma) \end{gathered}$$

and we say

$$\begin{gathered} Z_n \text{ is chi-squared distributed with } n \text{ degrees of freedom} \\ \text{and shape parameter } \sigma. \end{gathered} \tag{12.6}$$

The density function of such a random variable is given by

$$p_n(z) = \frac{z^{\frac{n}{2}-1} e^{-z/2\sigma^2}}{(2\sigma^2)^{n/2}\Gamma(n/2)}. \tag{12.7}$$

In the previous section we derived the key descriptive statistics, namely mean, variance, skewness and kurtosis, for a Gamma distributed random variable. With this information we have immediate access to the same statistics for a chi-squared random variable since, as we have demonstrated, the chi-squared distribution coincides with a special type of Gamma distribution. These results are captured in the following theorem:

Theorem 12.2. *Let $Z_n \sim \chi_n^2(\sigma)$ denote a chi-squared random variable with n degrees of freedom and with shape parameter σ. The mean, variance, skewness and kurtosis of Z_n are given by the following expressions:*

$$\begin{aligned} &\text{mean} \quad \mathbb{E}[Z_n] = n\sigma^2, \\ &\text{variance} \quad \mathbb{E}[(Z_n - n\sigma^2)^2] = 2n\sigma^4, \\ &\text{skewness} \quad S(Z_n) = \sqrt{\frac{8}{n}}, \\ &\text{kurtosis} \quad \mathcal{K}(Z_n) = 3 + \frac{12}{n}. \end{aligned}$$

We close the discussion on the chi-squared distribution by noting that, in practice, it is common to scale the normal random variables so that they have unit variance. Specifically, the following alternative definition is also employed:

Definition 12.3. *Let $(X_k)_{k=1}^n$ denote a sequence of independent random variables which are normally distributed and have zero mean, i.e., $X_k \sim N(0, \sigma_k^2)$ for $k = 1, \ldots, n$. The random variable*

$$Z_n = \left(\frac{X_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{X_n}{\sigma_n}\right)^2$$

is a chi-squared random variable with n degrees of freedom and unit shape parameter; we denote this mathematically by $Z_n \sim \chi_n^2$. The density function of Z_n is given by equation (12.7) with σ set equal to one.

To compliment the above definition we have, in Figure 12.1, an illustration of standardized chi-squared density functions for different degrees of freedom. We notice that each function resembles a hill where the climb to the peak (from 0) is steeper than the descent; i.e., it is skewed to the right. Furthermore, we observe that the density functions become flatter as the number of degrees of freedom increases.

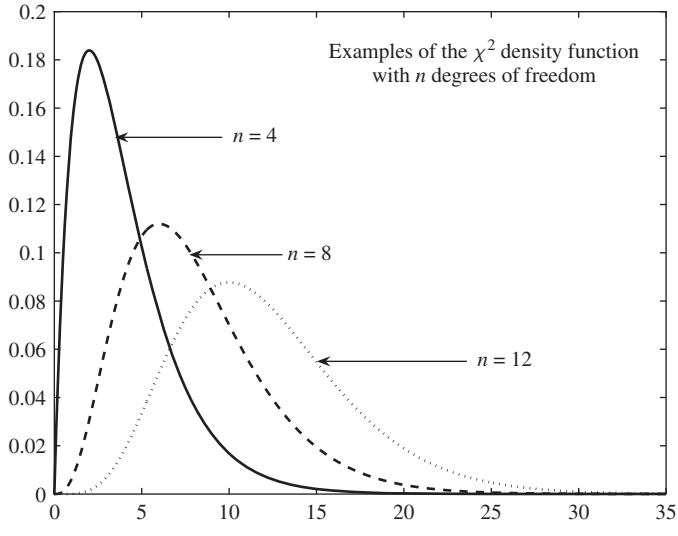


Figure 12.1 Example plots of the χ^2 density function for different degrees of freedom.

12.3 THE NON-CENTRAL CHI-SQUARED DISTRIBUTION

Suppose we have a general normal random variable $X \sim N(\mu, \sigma^2)$. We know from the previous section that

$$Z_1 = (X - \mu)^2 \sim \chi_1^2(\sigma).$$

In the current investigation we pose the following question:

How is the random variable $\widetilde{Z}_1 = X^2$ distributed?

We attack this problem with the same approach as for the chi-squared case, we begin by formally expressing the distribution function of $\widetilde{Z}_1$ as follows:

$$\begin{aligned} F(z) &= \mathbb{P}[\widetilde{Z}_1 \leq z] = \mathbb{P}[-\sqrt{z} \leq X \leq \sqrt{z}] \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\sqrt{z}}^{\sqrt{z}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx. \end{aligned}$$

Setting $u = (x - \mu)/\sigma$ we find that

$$\begin{aligned} F(z) &= \frac{1}{\sqrt{2\pi}} \int_{-(\sqrt{z}+\mu)/\sigma}^{(\sqrt{z}-\mu)/\sigma} \exp\left(-\frac{u^2}{2}\right) du \\ &= \Phi\left(\frac{\sqrt{z}-\mu}{\sigma}\right) - \Phi\left(-\left(\frac{\sqrt{z}+\mu}{\sigma}\right)\right) \end{aligned}$$

$$= \Phi\left(\frac{\sqrt{z}-\mu}{\sigma}\right) - \left(1 - \Phi\left(\frac{\sqrt{z}+\mu}{\sigma}\right)\right) \quad \text{using } \Phi(-x) = 1 - \Phi(x)$$
$$= 1 + \Phi\left(\frac{\sqrt{z}-\mu}{\sigma}\right) + \Phi\left(\frac{\sqrt{z}+\mu}{\sigma}\right).$$

We can find the density function of $\widetilde{Z}_1$ by differentiating its distribution with respect to z, this gives

$$p(z) = \frac{1}{2\sigma\sqrt{z}}\left[\Phi'\left(\frac{\sqrt{z}-\mu}{\sigma}\right) + \Phi'\left(\frac{\sqrt{z}+\mu}{\sigma}\right)\right]$$
$$= \frac{1}{2\sqrt{2\pi\sigma^2 z}}\left[\exp\left(-\frac{1}{2}\left(\frac{\sqrt{z}-\mu}{\sigma}\right)^2\right) + \exp\left(-\frac{1}{2}\left(\frac{\sqrt{z}+\mu}{\sigma}\right)^2\right)\right]$$
$$= \frac{1}{\sqrt{2\pi\sigma^2 z}}\exp\left(-\frac{1}{2}\left(\frac{z+\mu^2}{\sigma^2}\right)\right)\left[\frac{\exp(\sqrt{z}\mu/\sigma^2) + \exp(-\sqrt{z}\mu/\sigma^2)}{2}\right].$$

We can simplify this expression further by recalling that the hyperbolic cosine of any $x \in \mathbb{R}$ is defined by

$$\cosh(x) = \frac{e^x + e^{-x}}{2} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \cdots + \frac{x^{2n}}{2n!} + \cdots,$$

thus we can conclude that the density function of $\widetilde{Z}_1$ is given by

$$p(z) = \exp\left(-\frac{\mu^2}{2\sigma^2}\right)\exp\left(-\frac{z}{2\sigma^2}\right)\frac{\cosh(\sqrt{z}\mu/\sigma^2)}{\sqrt{2\pi\sigma^2 z}} \quad \text{whenever } z \geq 0$$

and so, using this expression, we can compute the characteristic function of $\widetilde{Z}_1$ as follows:

$$\phi_{\widetilde{Z}_1}(u) = \frac{\exp\left(-\frac{\mu^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}\int_0^\infty \exp\left(-\left(\frac{1}{2\sigma^2} - iu\right)x\right)\frac{\cosh(\sqrt{x}\mu/\sigma^2)}{\sqrt{x}}dx.$$

Employing the substitution $t = \sqrt{x}$, the expression becomes

$$\phi_{\widetilde{Z}_1}(u) = \sqrt{\frac{2}{\pi\sigma^2}}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\int_0^\infty \exp\left(-\left(\frac{1}{2\sigma^2} - iu\right)t^2\right)\cosh\left(\frac{\mu t}{\sigma^2}\right)dt.$$

To evaluate this integral we use the following pertinent identity:

$$\int_0^\infty \exp(-\beta t^2)\cosh(\alpha t)dt = \frac{1}{2}\sqrt{\frac{\pi}{\beta}}\exp\left(\frac{\alpha^2}{4\beta}\right) \quad \text{provided } \mathcal{R}(\beta) > 0.$$

Setting

$$\alpha = \frac{\mu}{\sigma^2} \quad \text{and} \quad \beta = \frac{1}{2\sigma^2} - iu$$

we can deduce that

$$\phi_{\widetilde{Z}_1}(u) = \exp\left(\frac{i\mu^2 u}{1-2i\sigma^2 u}\right)\frac{1}{\sqrt{1-2i\sigma^2 u}}$$
$$= \exp\left(\frac{i\mu^2 u}{1-2i\sigma^2 u}\right) \times \text{ (characteristic function of a } \chi_1^2(\sigma) \text{ r.v.).}$$

This expression allows us to view $\widetilde{Z}_1$ as a kind of shift of the plain chi-squared random variable with one degree of freedom and shape parameter σ. We say that such a variable has the non-central chi-squared distribution for which we have the following definition:

Definition 12.4. *If $X \sim N(\mu, \sigma^2)$ then the random variable X^2 is said to have the non-central chi-squared distribution with one degree of freedom, shape parameter σ and centrality parameter μ^2. To indicate this mathematically we write*

$$X \sim N(\mu, \sigma^2) \Rightarrow X^2 \sim \chi_1^2(\sigma, \mu^2).$$

We can appeal once more to Theorem 11.5 to generalize this discovery. Specifically, we can deduce that if $(X_k)_{k=1}^n$ denotes a sequence of independent random variables which are normally distributed and share the same common variance, i.e., $X_k \sim N(\mu_k, \sigma^2)$ for $k = 1, \ldots, n$, then the characteristic function of the random variable

$$\widetilde{Z}_n = X_1^2 + \cdots + X_n^2 \tag{12.8}$$

is given by

$$\phi_{\widetilde{Z}_n}(u) = \exp\left(\frac{iu\sum_{k=1}^n \mu_k^2}{1-2i\sigma^2 u}\right)\frac{1}{(1-2i\sigma^2 u)^{n/2}}$$
$$= \exp\left(\frac{iu\sum_{k=1}^n \mu_k^2}{1-2i\sigma^2 u}\right) \times \text{ (characteristic function of a } \chi_n^2(\sigma) \text{ r.v.).}$$

This observation allows us to generalize Definition 12.5 to cater for a sum of n normally distributed variables:

Definition 12.5. *If $X_k \sim N(\mu_k, \sigma^2)$ $(1 \le k \le n)$ are independent, then the random variable $\widetilde{Z}_n$ given by (12.8) is said to have the non-central chi-squared distribution with n degrees of freedom, shape parameter σ and centrality parameter $\sum_{k=1}^n \mu_k^2$. We denote this mathematically by writing*

$$\widetilde{Z}_n \sim \chi_n^2\left(\sigma, \sum_{k=1}^n \mu_k^2\right).$$

We note that the above definition relies upon the fact that the normal random variables share the same variance. If we relax the assumption of a shared variance, i.e., we consider

$$\widetilde{Z}_n = X_1^2 + \cdots + X_n^2 \quad \text{where} \quad X_k \sim N(\mu_k, \sigma_k^2) \quad k = 1, \ldots, n,$$

then the distribution of $\widetilde{Z}_n$ is much harder to pin down. However, it is always possible to scale the normal random variables so that they each have unit variance. This leads to the following alternative definition:

Definition 12.6. *Let $(X_k)_{k=1}^n$ denote a sequence of independent normally distributed random variables, $X_k \sim N(\mu_k, \sigma_k^2)$ for $k = 1, \ldots, n$. The random variable*

$$\widetilde{Z}_n = \left(\frac{X_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{X_n}{\sigma_n}\right)^2$$

is a non-central chi-squared random variable with n degrees of freedom, unit shape parameter and centrality parameter

$$\lambda = \left(\frac{\mu_1}{\sigma_1}\right)^2 + \cdots + \left(\frac{\mu_n}{\sigma_n}\right)^2.$$

We denote this mathematically by $\widetilde{Z}_n \sim \chi_n^2(\lambda)$.

We remark that when $\lambda = 0$ the variable coincides with the standard chi-squared random variable and so shares its density function. We observe from Figure 12.2 that, as the non-centrality parameter λ grows, the peak of its density function is shifted to the right and its overall shape is flattened.

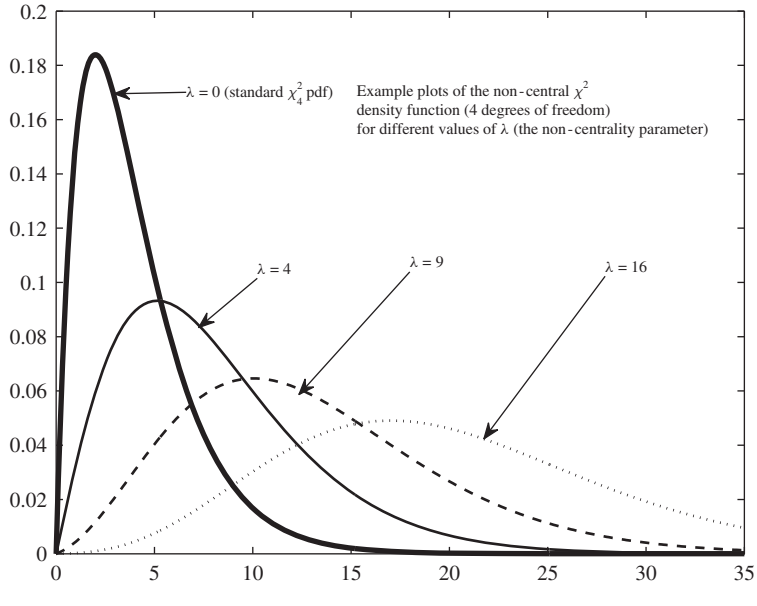


Figure 12.2 To illustrate the non-central χ^2 density function for different values of λ.

12.4 THE F-DISTRIBUTION

We open this section with a definition that pins down the class of random variables that are said to have the F-distribution:

Definition 12.7. *Let X and Y denote two independent random variables such that $X \sim \chi^2_n$ and $Y \sim \chi^2_m$. We say that the random variable*

$$U = \frac{X/n}{Y/m} \text{ is F-distributed with degrees of freedom } n \text{ and } m$$

and we write $U \sim F_{n,m}$

In order to investigate the F-distribution in more depth we require a useful theorem from calculus that deals with integration of functions. We consider two versions of this result, the familiar one-dimensional case and its generalization to two and higher dimensions.

- Integration in one dimension with a change of variable.
  Suppose we wish to integrate a function $f : \mathbb{R} \to \mathbb{R}$ over an interval $[a, b]$, i.e., we wish to compute

$$\int_a^b f(x)dx.$$

  In many applications we can make this calculation simpler if we introduce a change of variable, mathematically speaking we express x as a function ϕ say, of a new variable which we will call u. Once we have chosen the transformation such that $x = \phi(u)$, it is well known that the above integral is transformed according to

$$\begin{aligned} \int_a^b f(x)dx &= \int_{\phi(a)}^{\phi(b)} f \circ \phi(u)\phi'(u)du \\ &= \int_{\phi(a)}^{\phi(b)} \widetilde{f}(u)\phi'(u)du \quad \text{where} \quad \widetilde{f} = f \circ \phi. \end{aligned} \tag{12.9}$$

- Integration in two dimensions with a change of variables.
  Suppose now we wish to integrate a two-dimensional function $f : \mathbb{R}^2 \to \mathbb{R}$ over some region $\Omega_{xy} \subset \mathbb{R}^2$, i.e., we wish to compute

$$\int_{\Omega_{xy}} f(x, y)dxdx.$$

  Just as in the one-dimensional case we can change the variables x and y by expressing them as functions ϕ and ψ say, of two new variables which we will call u and v, i.e., we have

$$x = \phi(u, v) \quad \text{and} \quad y = \psi(u, v).$$

  Let Ω_{uv} denote the region in $\mathbb{R}^2$ that is mapped, via ϕ and ψ, to the original integration region Ω_{xy}. In direct analogy to the one-dimensional case it can be shown that the two-dimensional integral is

transformed as follows:

$$\int_{\Omega_{xy}} f(x, y)dxdx = \int_{\Omega_{uv}} \widetilde{f}(u, v)J(u, v)dudv, \tag{12.10}$$

where

$$\widetilde{f} = f(\phi(u, v), \psi(u, v)) \quad \text{and} \quad J(u, v) = \left| \frac{\partial \phi}{\partial u}\frac{\partial \psi}{\partial v} - \frac{\partial \phi}{\partial v}\frac{\partial \psi}{\partial u} \right|. \tag{12.11}$$

By comparing (12.10) with (12.9) we notice that the quantity $J(u, v)$ plays the same role in two dimensions as ϕ' does in one dimension. This quantity is called the Jacobian of the original coordinates (x, y) with respect to the transformed coordinates (u, v) and it is more commonly expressed as follows:

$$J(u, v) = \text{absolute value of det} \begin{pmatrix} \partial\phi/\partial u & \partial\psi/\partial u \\ \partial\phi/\partial v & \partial\psi/\partial v \end{pmatrix}.$$

An application of the change of variables result to probability theory is illustrated in the following investigation.

Let X and Y denote two continuous random variables whose joint probability density is given by $p(x, y)$, then if Ω_{xy} denotes a subset of $\mathbb{R}^2$ we have

$$\mathbb{P}[(X, Y) \in \Omega_{xy}] = \int_{\Omega_{xy}} p(x, y)dxdy.$$

Let us carefully create a new random variable U to be a function of X and Y, i.e., we set

$$U = f(X, Y) \quad \text{for some} \quad f : \mathbb{R}^2 \to \mathbb{R}.$$

Our aim is to find the probability density function of U. A crucial step in achieving this goal is to find a convenient function $g : \mathbb{R}^2 \to \mathbb{R}$ say, such that the random variable $V = g(X, Y)$ has the property that both X and Y are easily recoverable from U and V via

$$X = \phi(U, V) \quad \text{and} \quad Y = \psi(U, V),$$

where ϕ and ψ are smooth mappings from $\mathbb{R}^2$ to $\mathbb{R}$.

Let $q(u, v)$ denote the joint density of the pair (U, V) and let Ω_{uv} denote the region in $\mathbb{R}^2$ that is mapped, via ϕ and ψ, to Ω_{xy}, then, using (12.10), we can conclude that

$$\int_{\Omega_{xy}} p(x, y)dxdx = \int_{\Omega_{uv}} \underbrace{p(\phi(u, v), \psi(u, v))J(u, v)}_{q(u,v)} dudv.$$

Thus, the joint probability density function of (U, V) is given by

$$q(u, v) = p(\phi(u, v), \psi(u, v))J(u, v). \tag{12.12}$$

In particular, the marginal density of $U = f(X, Y)$ is thus given by

$$p_U(u) = \int_{-\infty}^{\infty} q(u, v)dv. \tag{12.13}$$

We can now employ this process to derive an expression for the density of an F-distributed random variable. To illustrate this we recall that we have two independent random variables X and Y such that

$$X \sim \chi_n^2 \quad \text{and} \quad Y \sim \chi_m^2.$$

Since these variables are assumed to be independent we can immediately deduce that the joint density function $p(x, y)$ of the pair (X, Y) is the product of their univariate densities, i.e.,

$$p(x, y) = \underbrace{\frac{e^{-x/2}x^{\frac{n}{2}-1}}{2^{n/2}\Gamma(n/2)}}_{\text{pdf of } X\sim\chi_n^2} \cdot \underbrace{\frac{e^{-y/2}y^{\frac{m}{2}-1}}{2^{m/2}\Gamma(m/2)}}_{\text{pdf of } Y\sim\chi_m^2}.$$

We can think of the F-distributed random variable as the quotient function acting on X and Y, specifically we have

$$U = f(X, Y) = \frac{X/n}{Y/m} = \frac{m}{n} \cdot \frac{X}{Y}.$$

In addition, we define

$$V = g(X, Y) = Y.$$

Now, the original random variable X can be recovered from U and V by applying the mapping

$$\phi(u, v) = \frac{n}{m}uv,$$

since

$$\phi(U, V) = \frac{n}{m}U \cdot V = \frac{n}{m}\left(\frac{m}{n} \cdot \frac{X}{Y}\right) \cdot Y = X.$$

Furthermore, it is clear that Y can also be recovered from U and V by applying the map

$$\psi(u, v) = v \quad \text{since} \quad \psi(U, V) = V = Y.$$

We can now dip into the calculus we have developed in this section, specifically equation (12.12), to discover that the joint distribution of the pair (U, V) is given by

$$q(u, v) = \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}-1}}{2^{\frac{n+m}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \exp\left(-v\left(1 + \frac{n}{m}u\right)\right) u^{\frac{n}{2}-1}v^{\frac{n+m}{2}-2}J(u, v),$$

where

$$J(u,v) = \text{absolute value of det} \begin{pmatrix} \frac{nv}{m} & 0 \\ \frac{nu}{m} & 1 \end{pmatrix} = \frac{nv}{m}.$$

Hence,

$$q(u,v) = \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{2^{\frac{n+m}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \exp\left(-v\left(1+\frac{nu}{m}\right)\right) u^{\frac{n}{2}-1} v^{\frac{n+m}{2}-1},$$

for all $u \in \mathbb{R}$ and $v \geq 0$.

We can now employ (12.13) to finally deduce that the density of the F-distributed random variable U with degrees of freedom n and m is given by

$$p_{n,m}(u) = \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}}}{2^{\frac{n+m}{2}}\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} u^{\frac{n}{2}-1} \int_0^\infty \exp\left(-v\left(1+\frac{nu}{m}\right)\right) v^{\frac{n+m}{2}-1} dv.$$

To evaluate this integral we use the following identity:

$$\int_0^\infty v^{a-1}\exp(-bv)dv = \frac{1}{b^a}\Gamma(a).$$

We set

$$a = \frac{n+m}{2} \quad \text{and} \quad b = 1+\frac{nu}{m},$$

and we find that

$$\begin{aligned} p_{n,m}(u) &= \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{n}{m} \left(\frac{nu}{m}\right)^{\frac{n}{2}-1} \left(1+\frac{nu}{m}\right)^{-\left(\frac{n+m}{2}\right)} \\ &= \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} n^{\frac{n}{2}} m^{\frac{m}{2}} u^{\frac{n}{2}-1} (m+nu)^{-\left(\frac{n+m}{2}\right)}, \quad \text{for} \quad u \in \mathbb{R}. \end{aligned} \tag{12.14}$$

12.5 THE t-DISTRIBUTION

We open this section, as before, with a definition that pins down the class of random variables that are said to have the t-distribution.

Definition 12.8. *Let X and Y denote two independent random variables such that $X \sim N(0,1)$ and $Y \sim \chi^2_n$. We say that the random variable*

$$T = \frac{X}{\sqrt{Y/n}} \quad \text{is } t\text{-distributed with } n \text{ degrees of freedom.}$$

Before we embark on a deep investigation of the nature of a t-distributed random variable we briefly pause to make two simple, but vital, observations.

- Observation 1: The symmetry of a t-distributed random variable.
  At the very basic level the t-distributed random variable is simply a quotient of two independent random variables. The fact that the random variable in the numerator is symmetric, it is $N(0, 1)$, implies that the t-distribution itself must also be symmetric.
- Observation 2: The square of a t-distributed random variable.
  The square of a t-distributed random variable T has the form

$$T^2 = \frac{X^2}{Y/n} \quad \text{where} \quad X \sim N(0,1) \quad \text{and} \quad Y \sim \chi^2_n.$$

  We know that $X^2 \sim \chi^2_1$ and so we immediately see that T^2 coincides with an F-distributed random variable with degrees of freedom 1 and n, i.e., $T^2 \sim F_{1,n}$.

Let p_T denote the probability density function for a t-distributed random variable, it is our aim to derive the closed-form representation for this function. As a starting point we can use the general properties of distribution functions to formally write

$$\int_{-t}^{t} p_T(x)dx = \mathbb{P}[-t \leq T \leq t] = \mathbb{P}[T^2 \leq t^2] = \mathbb{P}\left[\frac{X^2}{Y/n} \leq t^2\right].$$

We can use observation 2 above to conclude that

$$\int_{-t}^{t} p_T(x)dx = \int_{0}^{t^2} p_{1,n}(x)dx,$$

where $p_{1,n}(x)$ is the density function of an $F_{1,n}$ random variable. We can differentiate both sides of the above identity (with respect to t) to yield that

$$p_T(t) + p_T(-t) = 2tp_{1,n}(t^2).$$

The symmetry of the t-distribution (observation 1, above) means that $p_T(t) = p_T(-t)$ for all $t \geq 0$ and so allows us to conclude that

$$p_T(t) = tp_{1,n}(t^2).$$

We now employ expression (12.14) with $n = 1$ and $m = n$ to finally conclude that the density function for a t-distributed random variable is given by

$$p_T(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\sqrt{\pi n}} \left(1 + \frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, \quad \text{for } x \in \mathbb{R}. \tag{12.15}$$

For completeness, we let

$$t_n(x) = \int_{-\infty}^{x} p_T(u)du \tag{12.16}$$

denote the distribution function of T, and we summarize our development in the following theorem:

Theorem 12.9. *Let X and Y denote a pair of independent random variables such that $X \sim N(0,1)$ and $Y \sim \chi^2_n(1)$. The new random variable $T = \frac{X}{\sqrt{Y/n}}$ is t-distributed with n degrees of freedom. Its density function p_T is given by (12.15) and, accordingly, its distribution function t_n is given by (12.16).*

We know that the first four moments of any distribution provide useful summary information of the underlying random variable. For the t-distribution we have:

$$\begin{aligned}
\text{if } T \sim t_n \ (n > 4) \text{ then } \mathbb{E}[T] &= 0 \quad \text{(zero mean)};\\
\mathbb{E}[T^2] &= \frac{n}{n-2};\\
\mathbb{E}[T^3] &= 0 \quad (t_n \text{ is symmetric about zero});\\
\mathbb{E}[T^4] &= \frac{3n^2}{(n-2)(n-4)} \quad \left(\mathcal{K}(T) = 3\frac{n-2}{n-4} > 3\right).
\end{aligned}$$

We remark that the condition $n > 4$ above is needed to ensure that the first four moments of T exist. Indeed, it can be shown that a general t-distributed random variable (with n degrees of freedom) only possesses moments up to and including order $n-1$. In Figure 12.3 we see a plot of the standard t-density function (12.15) for different degrees of freedom; we notice that the distribution is symmetric about 0 and that, as n grows, it becomes more peaked at the origin and decays more quickly out in the tails.

We close the current discussion with the following definition which describes how the student t-distribution can be generalized to higher dimensions.

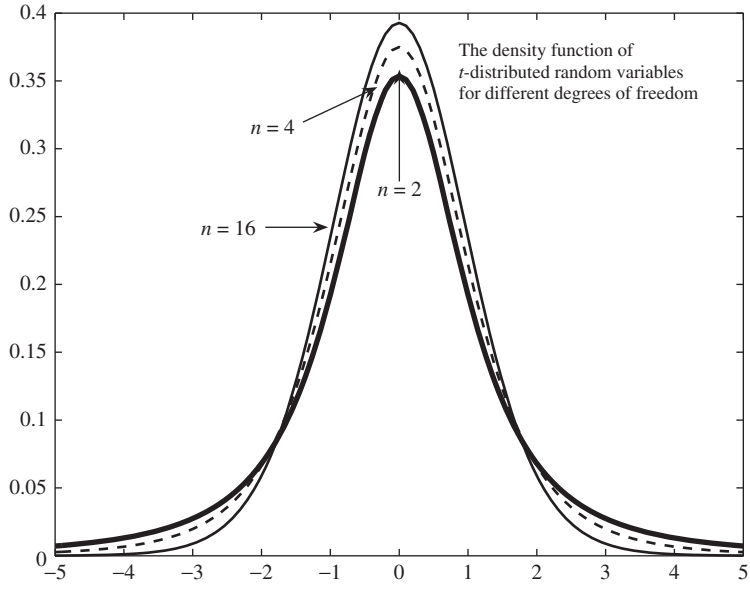


Figure 12.3 To illustrate the density function of a t-distributed random variable with different degrees of freedom.

Definition 12.10. *Let* $\boldsymbol{t} = (t_1, \ldots, t_d)^T \in \mathbb{R}^d$ *denote a* d*-dimensional random vector whose mean vector and covariance matrix are given, respectively, by*

$$\boldsymbol{e} = \mathbb{E}[\boldsymbol{t}] \quad \text{and} \quad \boldsymbol{V} = \mathbb{E}[(\boldsymbol{t} - \boldsymbol{e})(\boldsymbol{t} - \boldsymbol{e})^T].$$

This vector is said to have the multivariate t*-distribution with* $n > 2$ *degrees of freedom if its density function has the form*

$$p(\boldsymbol{t}) = \frac{\Gamma\left(\frac{n+d}{2}\right)}{(\pi d)^{n/2}\,\Gamma\left(\frac{n}{2}\right)} \frac{1}{\sqrt{|\det(\Sigma)|}} \left(1 + \frac{(\boldsymbol{t} - \boldsymbol{e})^T \Sigma^{-1} (\boldsymbol{t} - \boldsymbol{e})}{n}\right)^{\frac{n+d}{2}},$$

where

$$\Sigma = \frac{1}{n/(n-2)} \boldsymbol{V} \in \mathbb{R}^{d \times d}$$

is the so-called dispersion matrix; a scaling of the covariance matrix so that each of its diagonal entries equals one.

Essential Mathematics for Market Risk Management
by Simon Hubbert

13
A Crash Course on Financial Derivatives

So far in this book we have placed all of our attention on problems involving portfolios of fairly basic financial assets such as stocks and shares. For instance, if we have n such assets then we construct a portfolio by choosing to buy (or short sell) α_i shares in the ith asset S_i for $1 \leq i \leq n$, the resulting value of the portfolio at time t is the appropriate linear combination

$$V_{\text{lin}}(t) = \sum_{i=1}^{n} \alpha_i S_i(t).$$

In the real world however the modern-day financial risk manager will, more often than not, be faced with portfolios containing derivatives which are more complex products. Thus, in contrast to the linear portfolio above, we are now faced with an investment of the form

$$V_{\text{non}}(t) = \sum_{i=1}^{n} \alpha_i g_i \left(S_i(t)\right),$$

where each g_i is a non-linear function of the underlying asset (or in some cases several assets).

The task of managing the risk of a non-linear portfolio can only be achieved if we know, in advance, the form of the derivative pricing functions, i.e., we are equipped with $\{g_i\}_{i=1}^{n}$. In this chapter we aim to provide some mathematical insight behind the enormous challenge of derivative pricing. In order to make our development compact we shall concentrate only on plain derivative products and we sacrifice some mathematical rigour in favour of intuitive explanations. The reader who is interested in learning more is advised to consult any of the following excellent textbooks: Higham (2004), Joshi (2005), Neftci (1996), Wilmott et al (1995).

13.1 THE BLACK–SCHOLES PRICING FORMULA

In this section we shall develop the pricing framework for a plain European option. This is a contract, negotiated at time t, on an underlying $S(t)$ whereby one party, known as the writer of the option, agrees to offer another party, known as the holder of the option, the opportunity to receive a future payoff whose value is determined by the level of the underlying at some future time point T and some strike price K. We let $g(S, t)$ denote the value of the option at time t, then we can write that

$$g(S, T) = \text{payoff}(S(T), K).$$

The classic example of a European option is the plain call option which offers the holder the right, but not the obligation, to buy the asset for the fixed strike price K at a future date T known as the maturity or exercise date of the option.

If the holder buys the option at time t then he pays the market price which we denote by $g_c(S, t)$. This upfront payment can be viewed as a premium for the opportunity to receive, at time T, a potentially lucrative payoff given by

$$g_c(S, T) = \max(S(T) - K, 0).$$

We remark that an equivalent product which offers the holder the right (but not the obligation) to sell an asset for a fixed price in the future is commonly known as a put option and we denote its value at time t by $g_p(S, t)$. The payoff structure of the put option is given by

$$g_p(S, T) = \max(K - S(T), 0) = -g_c(S, T).$$

Clearly the price of both the European call and put options will fluctuate as time varies over the interval $[t, T]$. Both products can be viewed as functions of the price path of underlying risky assets.

In the early 1970s Fischer Black and Myron Scholes published their now-famous stock option pricing formula Black and Scholes (1973). The line of attack that they used to reach their formula is as follows:

1. Propose a realistic yet workable model for the evolution of a typical asset price.
2. Derive a second-order approximation for a non-linear function (representing the price of an option) whose value depends upon the price of the underlying asset.
3. Demonstrate that a simple portfolio consisting of the option and the underlying asset can be constructed in such a way that it is entirely risk free.
4. Use a no-arbitrage argument to derive the partial differential equation (PDE) which must be satisfied by the price of the option.
5. Equip the PDE with its appropriate boundary conditions and solve for the option price.
6. Use the option pricing formula to investigate how the value of the derivative reacts to changes in the underlying economic variables.

13.1.1 A model for asset returns

Before we can put the Black–Scholes machinery into action we need to establish a mathematical model that mimics the evolution of the underlying asset price. We assume that we know the price $S(t)$ at the present time t and for our very first attempt, we propose that the price at time $t + \Delta t$ is given by

$$S(t + \Delta t) = S(t) + \varepsilon,$$

where ε is a random variable; we think of it as the random shock that causes $S(t)$ to jolt up or down to $S(t + \Delta t)$. We shall assume that ε is a standard normal random variable and write

$$\Delta S = S(t + \Delta t) - S(t) = \varepsilon \sim N(0, 1). \tag{13.1}$$

We certainly cannot expect (13.1) to be a realistic model for all assets. In reality, we know that:

- some assets are more volatile than others;
- most asset prices tend to drift, i.e., despite the random fluctuations there is usually a long-term trend;
- as the time period Δt becomes larger we become more uncertain about the range of likely values for the future asset price.

We address each of these points in turn and, in response, we refine our model in the following way:

$$\begin{aligned}\Delta S = \varepsilon &\Rightarrow \Delta S = \sigma_S \varepsilon \quad \text{(allow for asset volatility)}\\ &\Rightarrow \Delta S = \mu_S \Delta t + \sigma_S \varepsilon \quad \text{(allow for deterministic drift)}\\ &\Rightarrow \Delta S = \mu_S \Delta t + \sigma_S \sqrt{\Delta t}\varepsilon \quad \text{(allow for long-term uncertainty).}\end{aligned}$$

We note that we have chosen to scale the volatility by the factor $\sqrt{\Delta t}$ so that the variance of the asset grows linearly in time.

The main problem with the modified model is that, as it stands, there is a danger that the values it delivers can imply a negative (or zero) asset price. In order to overcome this we decide instead to model returns rather than pure prices, in which case we propose that

$$\begin{gathered}\Delta S = S(t)\,(\mu \Delta t + \sigma \Delta W)\\ \text{where} \quad \Delta W = \sqrt{\Delta t}\varepsilon \quad \text{and} \quad \varepsilon \sim N(0,1).\end{gathered} \tag{13.2}$$

Note that in the above model we consider μ and σ as the drift rate and volatility of the asset returns and not the asset price.

We can now view (13.2) as our first prototype model for asset returns, it consists of a deterministic drift component $\mu \Delta t$ and a random or stochastic component driven by the random variable ΔW. The mean and variance of ΔW are given by

$$\mathbb{E}[\Delta W] = \mathbb{E}[\sqrt{\Delta t}\varepsilon] = 0 \quad \text{and} \quad \text{var}(\Delta W) = \mathbb{E}[\Delta t \varepsilon^2] = \Delta t.$$

The same calculations for $(\Delta W)^2$ give

$$\mathbb{E}[(\Delta W)^2] = \text{var}(\Delta W) = \Delta t$$

and

$$\text{var}\left((\Delta W)^2\right) = \mathbb{E}[\Delta t^2(\varepsilon^2 - 1)^2] = \Delta t^2 \mathbb{E}[(\varepsilon^2 - 1)^2] = 2(\Delta t)^2.$$

This information tells us that as the time shift Δt becomes smaller, the variance of the random variable $(\Delta W)^2$ rapidly approaches zero. To see how fast, we consider the following time intervals:

$$\Delta t = 1 \text{ day} = \frac{1}{365} \Rightarrow \text{var}\left((\Delta W)^2\right) = 1.5 \times 10^{-5};$$

$$\Delta t = 1 \text{ hour} = \frac{1}{24 \times 365} \Rightarrow \text{var}\left((\Delta W)^2\right) = 2.6 \times 10^{-8};$$

$$\Delta t = 1 \text{ second} = \frac{1}{60 \times 24 \times 365} \Rightarrow \text{var}\left((\Delta W)^2\right) = 7.2 \times 10^{-12}.$$

What we have discovered is that $(\Delta W)^2$ behaves more and more like a constant as Δt shrinks. In fact, for infinitesimally small intervals we conclude that

$$dW_t = \varepsilon\sqrt{dt} \tag{13.3}$$

and

$$(dW_t)^2 = 2dt. \tag{13.4}$$

For completeness we observe that the continuous version of this model is given by

$$dS = S(t)\left(\mu dt + \sigma dW_t\right) \text{ where } dW_t = \varepsilon\sqrt{dt}. \tag{13.5}$$

13.1.2 A second-order approximation

We know, from a mathematical perspective, that a derivative product is nothing more than a non-linear function, g say, of the underlying asset $S(t)$. In this section we shall assume that the underlying asset evolves according to our model (13.5) and our aim is to derive a useful approximation that expresses the change in the value of the derivative as a function of the corresponding change in asset price.

As a starting point we shall assume that the $g(S, t)$ is a function whose first and second derivatives are continuous functions of S and t. If we naively ignore the fact that S is a random process then a perfectly acceptable second-order approximation for g is given by

$$dg(S,t) = \frac{\partial g}{\partial t}dt + \frac{\partial g}{\partial S}dS + \frac{1}{2}\left(\frac{\partial^2 g}{\partial t^2}(dt)^2 + 2\frac{\partial^2 g}{\partial S \partial t}dSdt + \frac{\partial^2 g}{\partial S^2}(dS)^2\right).$$

We ignore all terms that decay faster than dt: in which case, we have

$$dg(S,t) = \frac{\partial g}{\partial t}dt + \frac{\partial g}{\partial S}dS + \frac{\partial^2 g}{\partial t \partial S}dtdS + \frac{1}{2}\frac{\partial^2 g}{\partial S^2}(dS)^2.$$

We now address the stochastic terms; starting with the mixed second derivative, where we find

$$\begin{aligned} dSdt &= \left(\mu Sdt + \sigma SdW_t\right)dt \\ &= \mu S(dt)^{3/2} + \sigma SdW_t dt \end{aligned}$$

$$= \mu S(dt)^{3/2} + \sigma S\varepsilon\sqrt{dt}dt \quad \text{(using (13.3))}$$
$$= S(\mu + \sigma\varepsilon)(dt)^{3/2}.$$

We can ignore this term in the expansion because it decays faster than dt.
Now, for the pure second derivative term, we find

$$\begin{aligned}(dS)^2 &= (\mu S dt + \sigma S dW_t)^2 \\ &= \mu^2 S^2 (dt)^2 + 2\mu\sigma S^2 dt dW_t + \sigma^2 S^2 (dW_t)^2 \\ &= \underbrace{\mu^2 S^2 (dt)^2 + 2\mu\sigma S^2 \varepsilon (dt)^{3/2}}_{\text{ignore: decays faster than } dt} + \sigma^2 S^2 dt \quad \text{(using (13.3) and (13.4))}.\end{aligned}$$

Piecing this together, our second-order approximation becomes

$$dg(S,t) = \left(\frac{\partial g}{\partial t} + S\mu\frac{\partial g}{\partial S} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 g}{\partial S^2}\right)dt + S\sigma\frac{\partial g}{\partial S}dW_t. \tag{13.6}$$

The above formula, which we have casually derived, is more commonly known as Itô's formula. Itô's formula is a fundamental result and is of great importance because it defines the rule for how differentials of random variables need to be manipulated; it is the stochastic analogue of the familiar chain rule.

A particularly revealing example of Itô's formula arises when we take $g(S,t) = \log(S)$; i.e., we consider the logarithm of the stock price. In this case we have

$$\frac{\partial g}{\partial t} = 0, \quad \frac{\partial g}{\partial S} = \frac{1}{S} \quad \text{and} \quad \frac{\partial^2 g}{\partial S^2} = -\frac{1}{S^2}.$$

Then (13.6) tells us that

$$d\log(S,t) = \left(\mu - \frac{\sigma^2}{2}\right)dt + \sigma dW_t.$$

This expression can be understood much better in integral form and so, letting u denote the dummy variable for time integration, we have

$$\int_t^{t+\tau} d\log(S,u) = \int_t^{t+\tau}\left(\mu - \frac{\sigma^2}{2}\right)du + \sigma\int_t^{t+\tau} dW_u,$$

or equivalently

$$\begin{aligned}\log\left(\frac{S(t+\tau)}{S(t)}\right) &= \left(\mu - \frac{\sigma^2}{2}\right)\tau + \sigma \underbrace{\int_t^{t+\tau} dW_u}_{=\Delta W_t = W_{t+\tau} - W_t} \\ &= \left(\mu - \frac{\sigma^2}{2}\right)\tau + \sigma\sqrt{\tau}\varepsilon \quad \text{where} \quad \varepsilon \sim N(0,1).\end{aligned}$$

Taking the exponential of this equation and multiplying both sides by $S(t)$ we find that

$$S(t+\tau) = S(t)\exp\left[\left(\mu - \frac{\sigma^2}{2}\right)\tau + \sigma\sqrt{\tau}\varepsilon\right]. \tag{13.7}$$

We note that since the term in square brackets (the argument of the exponential function) is a normal random variable, we say that the future value of the asset $S(t+\tau)$ possesses the log-normal distribution. We were acquainted with the log-normal distribution in the previous chapter and we can call upon the properties developed there to deduce that

$$\mathbb{E}[S(t+\tau)] = S(t)\exp\left[\left(\mu + \frac{\sigma^2}{2}\right)\tau\right] \tag{13.8}$$

and

$$\text{var}(S(t+\tau)) = (S(t))^2\exp(2\mu)\left(\exp(\sigma^2\tau) - 1\right). \tag{13.9}$$

13.1.3 The Black–Scholes formula

The aim of this section is to provide a quick derivation of the famous Black–Scholes formula for the fair price of a European call option. We begin by fixing our notation. We assume that a firm or individual trader enters into the option contract at time $t = t_0$. As we know, the contract provides the right (but not the obligation) to buy an asset, whose price evolves according to (13.5), at a future time T for a strike price K. The prevailing risk-free interest rate, assumed to be constant, is denoted by r. We kick off the analysis by considering an investment portfolio whose value at any time $t \in [t_0, T]$ is given by

$$\Pi(S,t) = g_c(S,t) - \delta_t S_t,$$

i.e., the portfolio consists of holding

- a long position in a call option on the underlying asset whose value is denoted by $g_c(S,t)$;
- plus a short position of δ_t shares (a figure that continuously changes through time) in the underlying asset.

Using Itô's formula (13.6) we can deduce that the infinitesimal change in value of this portfolio is given by

$$\begin{aligned} d\Pi(S,t) &= dg_c(S,t) - \delta_t dS_t \\ &= \left(\frac{\partial g_c}{\partial t} + S\mu\frac{\partial g_c}{\partial S} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 g_c}{\partial S^2}\right)dt + S\sigma\frac{\partial c}{\partial S}dW_t - S\delta_t(\mu dt + \sigma dW_t) \\ &= \left(\frac{\partial g_c}{\partial t} + \mu S\left(\frac{\partial g_c}{\partial S} - \delta_t\right) + \frac{\sigma^2 S^2}{2}\frac{\partial^2 g_c}{\partial S^2}\right)dt + S\sigma\left(\frac{\partial g_c}{\partial S} - \delta_t\right)dW_t. \end{aligned}$$

We can see from this development that the portfolio can be made completely risk-free by ensuring that

$$\delta_t = \frac{\partial g_c}{\partial S} \tag{13.10}$$

at all times, in which case we have a completely deterministic portfolio whose price is the solution to the differential equation

$$\frac{d\Pi}{dt}(S,t) = \frac{\partial g_c}{\partial t} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 g_c}{\partial S^2}. \tag{13.11}$$

We assume that there are no arbitrage opportunities in the financial markets and so, the price of this risk-free portfolio should mimic the risk-free bond with maturity date T, i.e., it should satisfy

$$\frac{d\Pi}{dt}(S,t) = r\Pi(S,t), \tag{13.12}$$

where r denotes the prevailing risk-free rate assumed to be constant. As a direct result we can deduce, by comparing (13.11) and (13.12), that

$$r\left(g_c(S,t) - \frac{\partial g_c}{\partial S}S(t)\right) = \frac{\partial g_c}{\partial t} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 g_c}{\partial S^2}$$

and so, upon rearranging, we have the famous Black–Scholes partial differential equation for the European call option:

$$\frac{\partial g_c}{\partial t} + \frac{\sigma^2 S^2}{2}\frac{\partial^2 g_c}{\partial S^2} + rS\frac{\partial g_c}{\partial S} - rg_c = 0. \tag{13.13}$$

In order to find the unique solution to this PDE we need to specify its boundary conditions.

- Boundary conditions in time.
  Theoretically the European call option is alive and actively traded up to its maturity date T, where its value is determined by its final payoff function

$$g_c(S,T) = \max\left(S(T) - K, 0\right). \tag{13.14}$$

- Boundary conditions in asset value.
  In theory the underlying asset can drop as low as zero; in which case the stock, and the value of the option, both become worthless. On the other hand there is, in theory, no upper bound on how high the stock price can rise. In the hypothetical situation where $S(t)$ was to grow without bound then the corresponding option price would rise with it. In mathematical terms we summarize these extreme situations by writing

$$g_c(0,t) = 0 \quad \text{and} \quad g_c(S,t) \to S(t) \quad \text{as} \quad S(t) \to \infty. \tag{13.15}$$

The price of the call option can be completely specified by solving (13.13) subject to the boundary conditions (13.14) and (13.15). A derivation can be found in any good derivatives textbook and the final formula is given by

$$g_c(S,t) = S(t)\Phi(d_1) - K\exp(-r(T-t))\Phi(d_2), \tag{13.16}$$

where Φ denotes the standard normal cumulative distribution function and the quantities d_1 and d_2 are given by

$$d_1 = \frac{\log\left(\frac{S(t)}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} \tag{13.17}$$

and

$$d_2 = \frac{\log\left(\frac{S(t)}{K}\right) + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} = d_1 - \sigma\sqrt{T-t}. \tag{13.18}$$

We remark that, using the same methodology, the price of a European put option can be shown to be

$$g_p(S,t) = K\exp(-r(T-t))\Phi(-d_2) - S(t)\Phi(-d_1). \tag{13.19}$$

13.2 RISK-NEUTRAL PRICING

One of the most surprising aspects of the derivation of the Black–Scholes European option PDE is that it does not involve the drift rate μ of the underlying, and hence the price of the option is independent of the drift rate too. In simple terms the reason why the drift parameter does not appear in the final pricing formulae is because the Black–Scholes price is calculated within the so-called risk-neutral framework. To illuminate what this means, we think in the following terms:

- The real-world likelihood of future events is quantified by the true probability measure $\mathbb{P}$. In our case the key random variable is the future value of the asset which, according to our model (13.7), evolves according to

$$S(t+\tau) = S(t)\exp(X(\tau)) \quad \text{where} \quad X(\tau) \sim N(\mu_*, \sigma_*^2)$$

  where

$$\mu_* = \left(\mu - \frac{\sigma^2}{2}\right)\tau \quad \text{and} \quad \sigma_*^2 = \sigma^2\tau. \tag{13.20}$$

  The probability density function of $X(\tau)$ is given by

$$p(x) = \frac{1}{\sigma_*\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{x-\mu_*}{\sigma_*}\right)^2\right].$$

Now, using (11.19), the moment-generating function of $X(\tau)$ is given by

$$\mathbb{E}_{\mathbb{P}}[\exp(uX(\tau))] = \int_{\mathbb{R}} \exp(ux)\, p(x)dx = \exp\left(\mu_* u + \frac{\sigma_*^2 u^2}{2}\right),$$

and we can rearrange this expression to show that

$$\int_{\mathbb{R}} \underbrace{\exp\left[u\,(x-\mu_*) - \frac{\sigma_*^2 u^2}{2}\right] p(x)}_{=q(x)}\, dx = 1.$$

- The above development has led to the construction of a new probability density function q which, after a little algebraic manipulation, can be shown to have the form

$$q(x) = \frac{1}{\sqrt{2\pi}\sigma_*} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_* - u\sigma_*^2}{\sigma_*}\right)^2\right],$$

We recognize this as the density function of the random variable $X(\tau) + u\sigma_*^2$.
In summary, we have demonstrated a useful technique of changing the original probability measure $\mathbb{P}$ to a new one which we denote as $\mathbb{Q}$. The effect of this change is that the mean of the random variable under $\mathbb{P}$ is shifted when viewed under the new measure $\mathbb{Q}$; specifically, we have shown that:

$$\text{under the original measure } \mathbb{P} \text{ we have } X(\tau) \sim N\left(\left(\mu - \frac{\sigma^2}{2}\right)\tau,\ \sigma^2\tau\right)$$

$$\text{and under } \mathbb{Q} \text{ we have } X(\tau) \sim N\left(\left(\mu - \frac{\sigma^2}{2} + u\sigma^2\right)\tau, \sigma^2\tau\right).$$

We remark that by choosing

$$u = \frac{r-\mu}{\sigma^2},$$

we find that

$$\text{under } \mathbb{Q} \text{ we have } X(\tau) \sim N\left(\left(r - \frac{\sigma^2}{2}\right)\tau, \sigma^2\tau\right),$$

i.e., we have eliminated the drift term; in this case we say that $\mathbb{Q}$ is the risk-neutral probability measure.

- It can be shown that the value of a European call option at time t can be computed in two steps:

1. Compute the expected value of the payoff of the option under the risk-neutral measure $\mathbb{Q}$.
2. Discount the risk-neutral expected payoff to provide the present value of the option.

In our case the risk-neutral probability density is given by

$$q(x) = \frac{1}{\sqrt{2\pi(T-t)}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-(r-\sigma^2/2)(T-t)}{\sigma\sqrt{T-t}}\right)^2\right).$$

The payoff function at maturity is given by

$$\max(S(T)-K, 0) = S(t)\max\left(\exp(X(T-t)) - \frac{K}{S(t)}, 0\right),$$

and its expected value under $\mathbb{Q}$ is given by

$$\begin{aligned}
&S(t)\mathbb{E}_{\mathbb{Q}}\left[\max\left(\exp(X(T-t))-\frac{K}{S(t)},0\right)\right]\\
&=S(t)\int_{\mathbb{R}}\max\left(\exp(x)-\frac{K}{S(t)},0\right)q(x)dx\\
&=\frac{S(t)}{\sqrt{2\pi(T-t)}\sigma}\int_{-\log\frac{S(t)}{K}}^{\infty}\exp(x)\exp\left(-\frac{1}{2}\left(\frac{x-(r-\frac{\sigma^2}{2}(T-t)}{\sigma\sqrt{T-t}}\right)^2\right)dx\\
&-\frac{K}{\sqrt{2\pi(T-t)}\sigma}\int_{-\log\frac{S(t)}{K}}^{\infty}\exp\left(-\frac{1}{2}\left(\frac{x-(r-\frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}}\right)^2\right)dx.
\end{aligned}$$

It is easily verified that

$$\begin{aligned}
&\exp(x)\exp\left(-\frac{1}{2}\left(\frac{x-(r-\frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}}\right)^2\right)\\
&\qquad=\exp(r(T-t))\exp\left(-\frac{1}{2}\left(\frac{x-(r+\frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}}\right)^2\right).
\end{aligned}$$

We can use this fact, together with the substitutions

$$z=\frac{x-(r\pm\frac{\sigma^2}{2})(T-t)}{\sigma\sqrt{T-t}},$$

to show that the expected payoff at maturity, under the risk-neutral measure, is given by

$$S(t)\exp(r(T-t))\underbrace{\frac{1}{\sqrt{2\pi}}\int_{-d_1}^{\infty}\exp\left(-\frac{z^2}{2}\right)dz}_{=1-\Phi(-d_1)}-K\underbrace{\frac{1}{\sqrt{2\pi}}\int_{-d_2}^{\infty}\exp\left(-\frac{z^2}{2}\right)dz}_{=1-\Phi(-d_2)},$$

where d_1 and d_2 are given by (13.17) and (13.18) respectively and Φ is the standard normal distribution. We now use the fact that $\Phi(x)=1-\Phi(-x)$ to conclude that the fair, risk-neutral value of the European option is given by

$$\begin{aligned}
g_c(S,t)&=\exp(-r(T-t))\Big[S(t)\exp(r(T-t))\Phi(d_1)-K\Phi(d_2)\Big]\\
&=S(t)\Phi(d_1)-K\exp(-r(T-t))\Phi(d_2),
\end{aligned}$$

which is exactly the same formula as (13.16) and, using the same approach, the risk-neutral price for a European put option can be obtained in a similar fashion.

The option pricing results that we have derived are almost always referred to as the Black–Scholes formulae, named naturally after the two academics who discovered the result. However, it is widely accepted that these results may not have had such an immediate and dramatic impact had it not been for the pioneering work of Robert Merton. Merton,

working independently from Black and Scholes, published his own ground-breaking paper on option pricing in the same year (Merton, 1973). Merton took a slightly different approach to the same problem and demonstrated that the price of an option could always be replicated by keeping a careful balance of the underlying stock and a risk-free bond. Merton's work effectively verified the Black–Scholes formula and over the years many academics and practitioners have used his approach to price other, more exotic derivative products. The framework created by these three talented academics has since revolutionized finance; extensions, modification, improvements and new applications of their original ideas have enabled new and innovative mathematical solutions to seemingly complex real-world financial problems.

13.3 A SENSITIVITY ANALYSIS

In the early 1970s, when Black–Scholes and Merton published their celebrated papers, the practice of trading in options was in its infancy. The market for derivatives today is enormous and the European-style option remains one of the most popular products. However, as time has evolved a whole host of different varieties of derivatives have emerged and, for most of these products, the Black–Scholes–Merton framework can, in theory, be used to deliver the fair price of the product. Unfortunately, it is not possible to derive neat closed-form solutions for all of these products and hence the explosion of academic interest in derivative pricing, triggered by the need for fast and efficient numerical pricing algorithms.

At the most basic level we can view a derivative as a real-valued function of several variables. The Black–Scholes formula states that the price of a European option can be calculated based upon two variables (time to maturity $T - t$ and asset price $S(t)$) together with three additional input parameters (strike price K, volatility of the asset σ and prevailing risk-free rate r), each of which is assumed to remain constant over time. Thus, the value of a European call option can be written as the output of some function $g_c : \mathbb{R}^5 \to [0, \infty)$:

$$\text{Euro call price} = g_c(S, K, \sigma, r, T - t).$$

In order to get a feel for how sensitive the price is to changes in the underlying variables/parameters, we employ a simple Taylor approximation. We present here an investigation for the European call option.

13.3.1 Asset price sensitivity: The delta and gamma measures

We begin by analysing how a small change in the underlying asset can affect the price of a European option. To achieve this we ensure that all variables are fixed and then make a small perturbation ΔS to the asset price S. Clearly this disturbance will move the value of the option and we can quantify its reaction by considering a first-order Taylor approximation, i.e.,

$$g_c(S + \Delta S) - g_c(S) \approx \frac{\partial g_c}{\partial S} \Delta S.$$

Thus, the size of the partial derivative of $\partial g_c / \partial S$ dictates how the option price reacts to a small change in the underlying. This quantity is commonly called the delta of the option and, when it is computed at time t, we denote its value by $\delta_t^{(\text{call})}$.

We have already encountered this quantity in the derivation of the Black–Scholes formula (13.10). We recall that we have shown there that $\delta_t^{(\text{call})}$ is precisely the number of shares in the underlying that causes the portfolio

$$\Pi_t = g_c(S,t) - \delta_t^{(\text{call})} S(t)$$

to be risk-free and hence comparable to a risk-free bond. This discovery tells us that the holder of a European call option can offset his risk exposure by short selling $\delta_t^{(\text{call})}$ shares of the underlying asset. In financial terminology the strategy is called the delta hedge and the resulting position is said to be delta-neutral. In practice, the delta hedge should be monitored and adjusted regularly over the lifetime of the option.

Given that we have the explicit Black–Scholes expression (13.16) for the option price we can simply differentiate it with respect to S to derive the value of $\delta_t^{(\text{call})}$. We note that part of the differentiation involves computing

$$\begin{aligned}
&\frac{\partial}{\partial S}\Phi(d_1(S)) = \frac{1}{\sqrt{2\pi}}\frac{\partial d_1}{\partial S}\exp\left(-\frac{d_1^2}{2}\right)\\
\text{and}\quad &\frac{\partial}{\partial S}\Phi(d_2(S)) = \frac{\partial}{\partial S}\Phi(d_1(S) - \sigma\sqrt{T-t})\\
&\qquad = \frac{1}{\sqrt{2\pi}}\frac{\partial d_1}{\partial S}\exp\left(-\frac{(d_1 - \sigma\sqrt{T-t})^2}{2}\right),
\end{aligned}$$

where d_1 is given by (13.17) and $\Phi(\cdot)$ is the standard normal distribution function. Now using (13.17) we can show that

$$\frac{\partial d_1}{\partial S} = \frac{1}{S\sigma\sqrt{T-t}} \tag{13.21}$$

and

$$\exp\left(-\frac{(d_1 - \sigma\sqrt{T-t})^2}{2}\right) = \frac{S}{K}\exp(r(T-t))\exp\left(-\frac{d_1^2}{2}\right). \tag{13.22}$$

These two identities allow us to deduce that

$$\frac{\partial}{\partial S}\Phi(d_1(S)) = \frac{1}{S\sigma\sqrt{T-t}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{d_1^2}{2}\right) \tag{13.23}$$

and

$$\frac{\partial}{\partial S}\Phi(d_2(S)) = \frac{\exp(r(T-t))}{K\sigma\sqrt{T-t}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{d_1^2}{2}\right). \tag{13.24}$$

With this preparation the delta of a call option can be computed in a straightforward manner, using (13.16) we have

$$\begin{aligned}\delta_t^{(\text{call})} &= \frac{\partial}{\partial S}\left[S\Phi(d_1(S)) - K\exp(-r(T-t))\Phi(d_2(S))\right] \\ &= \Phi(d_1(S)) + \underbrace{S\frac{\partial}{\partial S}\Phi(d_1(S)) - K\exp(-r(T-t))\frac{\partial}{\partial S}\Phi(d_2(S))}_{=0 \text{ by substituting (13.23) and (13.24)}}.\end{aligned}$$

Hence we can conclude that

$$\delta_t^{(\text{call})} = \Phi(d_1) = \Phi\left(\frac{\log\left(\frac{S(t)}{K}\right) + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}\right). \tag{13.25}$$

We remark that the same calculation for a European put option yields

$$\begin{aligned}\delta_t^{(\text{put})} &= -\Phi(-d_1) = \Phi(d_1) - 1 \\ &= \delta_t^{(\text{call})} - 1.\end{aligned} \tag{13.26}$$

Based upon these calculations we can deduce that

$$\delta_t^{(\text{call})} \in (0,1) \quad \text{and} \quad \delta_t^{(\text{put})} \in (-1,0).$$

Using the fact that Φ is a strictly increasing function we can see, from (13.25), that the delta of a call option approaches 1 as the option itself becomes deeper and deeper in the money, i.e., as $S(t)$ climbs higher than K. Similarly, the delta approaches zero if the option drops more and more out of the money, i.e., as $S(t)$ drops further and further below K. Similar observations may be made regarding the delta of a European put.

To illuminate the real-world use of the delta of an option we consider the following:

- **Situation**. We hold a European call option to buy 100 shares of an underlying asset at some fixed maturity date T.
- **Action**. Under the assumptions of the Black–Scholes framework we can eliminate the risk from holding the option by short selling $100\delta_t^{(\text{call})}$ shares in the underlying. In this case we have a delta-neutral position.
- **Moving forward**. In order to remain delta-neutral we should maintain the above strategy throughout the lifetime of the option. In practice the updates (rebalancing the number of shares to short sell) will be performed at regular intervals.

We can gain more insight into the effect that a change in asset price has on the value of a European option by employing a second-order Taylor approximation, i.e., using a call

option as our example, we consider

$$g_c(S + \Delta S) - g_c(S) \approx \delta_t^{(\text{call})} \Delta S + \frac{1}{2} \frac{\partial^2 g_c}{\partial S^2} (\Delta S)^2.$$

The inclusion of the second derivative term brings curvature information of the option to the approximation; we call this the gamma of the option and write

$$\gamma_t^{(\text{call})} = \frac{\partial^2 g_c}{\partial S^2} = \frac{\partial \delta_t}{\partial S}.$$

We can see from the final equality that the gamma itself measures the rate of change of the delta of the option, or, put another way, the rate at which the delta-neutral position of the option must be rebalanced. To illustrate we consider the following:

- If the gamma of the option is large then this implies that a small movement in the underlying asset will require a potentially significant shift in the delta. As a result we can expect a delta-neutral position to be significantly disrupted and so the holder should be prepared to rebalance the position more frequently to maintain delta-neutrality.
- If the gamma of the option is small then the delta of the option is less sensitive to changes in the underlying asset. In this case it is less likely for a delta-neutral position to be significantly disrupted and the need for rebalancing is less urgent.

A closed formula for the gamma of a European call option is readily available: using (13.25), it is given by

$$\gamma_t^{(\text{call})} = \frac{\partial \delta_t}{\partial S} = \frac{\partial}{\partial S} \Phi\left(d_1(S)\right),$$

and we note that this quantity has already been calculated, see (13.23), and so we conclude that

$$\gamma_t^{(\text{call})} = \frac{1}{S\sigma\sqrt{T-t}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{d_1^2}{2}\right). \tag{13.27}$$

It is also clear from (13.26) that the gamma of a put option is the same as the gamma for the equivalent call option (i.e., same strike and maturity date), and so we write

$$\gamma_t = \gamma_t^{(\text{call})} = \gamma_t^{(\text{put})}.$$

13.3.2 Time decay sensitivity: The theta measure

We now move on to analyse how the price of a European call option changes over small time intervals. The approach is the same as before, i.e., we fix all parameters and then consider a small time shift Δt and examine the resulting first-order Taylor approximation

$$g_c(t + \Delta t) - g_c(t) \approx \frac{\partial g_c}{\partial t} \Delta t.$$

We can compute the time derivative of the call option as follows:

$$\frac{\partial g_c}{\partial t} = \frac{\partial}{\partial t} \left[S\Phi(d_1(t)) - K \exp(-r(T-t)) \Phi(d_2(t)) \right]$$

$$= S\exp\left(-\frac{d_1^2}{2}\right)\frac{\partial d_1}{\partial t} - K\exp(-r(T-t))\exp\left(-\frac{d_2^2}{2}\right)\frac{\partial d_1}{\partial t}$$
$$- Kr\exp(-r(T-t))\Phi(d_2(t)).$$

We can now use (13.22) to show that this simplifies to give

$$S\exp\left(-\frac{d_1^2}{2}\right)\left[\frac{\partial d_1}{\partial t} - \frac{\partial d_2}{\partial t}\right] - Kr\exp(-r(T-t))\Phi(d_2(t)).$$

Finally, we can use (13.18) to deduce that

$$\frac{\partial}{\partial t}(d_1 - d_2) = \frac{\partial}{\partial t}(\sigma\sqrt{T-t}) = -\frac{\sigma}{2\sqrt{T-t}},$$

and so it follows that

$$\frac{\partial g_c}{\partial t} = -\left[\frac{S\sigma}{2\sqrt{T-t}}\exp\left(-\frac{d_1^2}{2}\right) + Kr\exp(-r(T-t))\Phi(d_2(t))\right].$$

We note that the time derivative for a European call option is always negative, this indicates that if all input parameters remain unchanged then the value of the call option will naturally depreciate over time. It is customary to define the theta of an option as the value of its time derivative, and we write

$$\theta_t^{(\text{call})} = \frac{\partial g_c}{\partial t}. \tag{13.28}$$

The theta for a put option can be calculated in precisely the same way, and we find that

$$\theta_t^{(\text{put})} = \frac{\partial g_p}{\partial t} = \theta_t^{(\text{call})} + rK\exp(-r(T-t)). \tag{13.29}$$

13.3.3 The remaining sensitivity measures

Our sensitivity analysis so far has only focused on quantifying how the price of a European call option reacts to small changes in time and in asset price. The remaining quantities which are needed to price the option are:

- the fixed strike price K;
- the volatility σ of the asset;
- the risk-free interest rate r.

Now, in the Black–Scholes framework, it is assumed that both the volatility and the risk-free interest rate are fixed constants. At short intervals the assumption of a constant interest rate is reasonable, however the assumption of constant volatility is controversial and is an issue which we shall address in more detail later in this book. Despite these assumptions it is helpful to investigate how the option price would react to small changes in these parameters. As before, we achieve this by considering first-order Taylor approximations.

- For volatility we have

$$g_c(\sigma + \Delta\sigma) - g_c(\sigma) \approx \frac{\partial g_c}{\partial \sigma}\Delta\sigma$$

and we define

$$\upsilon^{(\text{call})} = \frac{\partial g_c}{\partial \sigma} \quad \text{to be the vega of the option.}$$

Using the Black–Scholes formulae it can be shown that the vega of a call is the same as the vega of a put, and it is given by

$$\upsilon = \upsilon^{(\text{call})} = \upsilon^{(\text{put})} = S\sqrt{T-t}\exp\left(-\frac{d_1^2}{2}\right).$$

- For the risk-free rate we have

$$g_c(r + \Delta r) - g_c(r) \approx \frac{\partial g_c}{\partial r}\Delta r,$$

and we define

$$\rho^{(\text{call})} = \frac{\partial g_c}{\partial \sigma} \quad \text{to be the rho of the option.}$$

Using the Black–Scholes formula (13.16) it can be shown that the rho of a European call is given by

$$\rho^{(\text{call})} = K(T-t)\exp(-r(T-t))\Phi(d_2)$$

and, equivalently, the rho for a European put is given by

$$\rho^{(\text{put})} = -K(T-t)\exp(-r(T-t))\Phi(-d_2).$$

We note that a European call option always has a positive rho, whereas the rho for a European put option is always negative. This indicates that a hike in the risk-free rate causes call option prices to rise, whereas put option prices fall. This makes economic sense as higher interest rates tend to increase future expectations of asset prices; thus making call options more appealing and put options less so.

Essential Mathematics for Market Risk Management
by Simon Hubbert

14
Non-linear Value at Risk

In this chapter we will show how analytic and semi-analytic expressions for VaR can be developed for portfolios consisting entirely of derivative products. The material is based upon the influential research paper Britten-Jones and Schaeffer (1999) and it is organized in the following stages:

1. We fix our notation by briefly reviewing the calculation of VaR for a linear portfolio of the form

$$V_{\text{lin}}(t) = \sum_{i=1}^{n} \alpha_i S_i(t). \tag{14.1}$$

2. In order to reduce the dimensionality of the VaR calculation we propose a factor model structure for the underlying assets of the form

$$S_i = a_i + \sum_{r=1}^{k} b_{ir} f_r, \quad i = 1, \ldots, n.$$

3. Using the factor model structure we propose linear and quadratic approximations for a non-linear portfolio of the form

$$V_{\text{non}}(t) = \sum_{i=1}^{n} \alpha_i g(f_1, \ldots, f_k, t). \tag{14.2}$$

We then demonstrate how these approximations can be used to derive closed-form solutions to the VaR calculation.

14.1 LINEAR VALUE AT RISK REVISITED

Let us assume that the portfolio under scrutiny is linear, i.e., it has the form (14.1). We shall assume that the risk manager monitors the portfolio over a time period of $t \mapsto t + \Delta t$, where the values of the underlying assets change from $S_i(t)$ to $S_i(t + \Delta t)$ for $i = 1, \ldots, n$. We now let

$$\Delta L_i = -(S_i(t + \Delta t) - S_i(t)), \quad i = 1, \ldots, n$$

denote the potential loss random variables for each of the underlying assets and, with this in place, we define

$$\Delta \mathcal{L}_t = \sum_{i=1}^{n} \alpha_i \Delta L_i$$

to be the corresponding potential loss random variable for the portfolio.

If we assume that the distribution function of $\Delta\mathcal{L}_t$ is continuous and monotonically increasing (i.e., no jumps or plateaus) then, for a given confidence level α, the value at risk for the portfolio is the number VaR_α that satisfies

$$\mathbb{P}[\Delta\mathcal{L}_t \leq \text{VaR}_\alpha] = \alpha.$$

Furthermore, if we assume that the vector

$$\mathbf{\Delta L} = (\Delta L_1, \ldots, \Delta L_n)^T \sim N(\mathbf{e}, \mathbf{V})$$

then we can deduce that

$$\Delta\mathcal{L}_t \sim N\left(\boldsymbol{\alpha}^T\mathbf{e}, \boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha}\right),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^T$ denotes the composition vector of the portfolio. We can now appeal to the normal VaR framework we developed in Chapter 10, to conclude that

$$\text{VaR}_\alpha(\Delta\mathcal{L}_t) = \boldsymbol{\alpha}^T\mathbf{e} + \Phi^{-1}(\alpha)\sqrt{\boldsymbol{\alpha}^T\mathbf{V}\boldsymbol{\alpha}}.$$

14.2 APPROXIMATIONS FOR NON-LINEAR PORTFOLIOS

We now turn attention to non-linear products and assume that we hold an investment portfolio that consists entirely of derivatives, i.e., it has the form (14.2). As before we choose to monitor the portfolio over a time period $t \mapsto t + \Delta t$, where the value of the derivatives changes from $g_i(S_i(t))$ to $g_i(S_i(t + \Delta t))$. We denote this change in value by

$$\Delta g_i = g_i\big(S_i(t + \Delta t)\big) - g_i\big(S_i(t)\big), \quad i = i, \ldots, n \tag{14.3}$$

and consequently the random change in the portfolio value is captured by

$$\Delta V_{\text{non}}(t) = \sum_{i=1}^{n} \alpha_i \Delta g_i.$$

In order to derive VaR estimates for this non-linear portfolio we need some insight into the distributional properties of $\Delta V_{\text{non}}(t)$. We recall that in the linear case we simply assumed that the price changes in the underlying asset values were normally distributed. We are unable to make the same assumption in this setting since, in general, if $X \sim N(\mu, \sigma^2)$ and g is a non-linear function, then we cannot conclude that $g(X)$ is a normally distributed random variable.

In order to make progress we decide to approximate $\Delta V_{\text{non}}(t)$ and we do this by employing two fairly simple Taylor approximations.

- First-order approximation.
  Here our approximation uses only first derivative information and is given by

$$\Delta g_i \approx \Delta^\delta g_i = \frac{\partial g_i}{\partial t}\Delta t + \frac{\partial g_i}{\partial S_i}\Delta S_i, \quad i = 1, \ldots, n. \tag{14.4}$$

  This approximation is often referred to as the delta approximation (hence the notation Δ^δ) because the term $\partial g_i/\partial S_i$ is more commonly known as the delta of derivative g_i.

- Second-order approximation.
  Here we enhance (14.4) by including second derivative information. The approximation is given by

$$\Delta g_i \approx \Delta^{\gamma} g_i = \frac{\partial g_i}{\partial t}\Delta t + \frac{\partial g_i}{\partial S_i}\Delta S_i + \frac{1}{2}\frac{\partial^2 g_i}{\partial S_i^2}(\Delta S_i)^2$$

$$= \Delta^{\delta} g_i + \frac{1}{2}\frac{\partial^2 g_i}{\partial S_i^2}(\Delta S_i)^2. \tag{14.5}$$

  This approximation is often referred to as the gamma approximation (hence the notation Δ^{γ}) because the term $\partial^2 g_i/\partial S_i^2$ is more commonly known as the gamma of derivative g_i.

Before we consider how to employ these approximations in practice we shall now devote some attention to the computational aspects of the problem. One of the biggest hurdles we face when calculating VaR is the huge computational cost; an extremely large number of parameters is required and the matrices needed to store these numbers can become ill-conditioned, i.e., almost singular. For this reason it is always wise to employ a factor model to ease the computational burden. For the current problem we aim to select a small number of observable risk factors based upon their ability to explain asset price movements. If k risk factors $\{f_1(t), \ldots, f_k(t)\}$ are selected then we propose the following linear model:

$$S_i(t) = a_i + \sum_{r=1}^{k} b_{ir} f_r(t), \quad i = 1, \ldots, n$$

and so, for the change in value of the underlying assets, we have

$$\Delta S_i = \sum_{r=1}^{k} b_{ir}\Delta f_r, \quad i = 1, \ldots, n, \tag{14.6}$$

where

$$\Delta f_r = f_r(t + \Delta t) - f_r(t), \quad r = 1, \ldots, k.$$

Under this model we can view each of our derivatives $g_i (1 \le i \le n)$ as functions of the same underlying risk factors, i.e., we consider

$$g(f_1, \ldots, f_k, t) = g\left(a_i + \sum_{r=1}^{k} b_{ir} f_r(t), t\right).$$

Now we can show, with the help of the chain rule, that the delta approximation to the change in value of the derivative function g_i can be written as

$$\Delta^{\delta} g_i = \frac{\partial g_i}{\partial t}\Delta t + \sum_{r=1}^{k} b_{ir}\frac{\partial g_i}{\partial f_r}\Delta f_r, \quad i = 1, \ldots, n. \tag{14.7}$$

Similarly, we can show that the gamma approximation is given by

$$\Delta^{\gamma} g_i = \Delta^{\delta} g_i + \frac{1}{2}\sum_{r=1}^{k}\sum_{s=1}^{k} b_{ir} b_{is}\frac{\partial^2 g_i}{\partial f_r \partial f_s}\Delta f_r \Delta f_s, \quad i = 1, \ldots, n. \tag{14.8}$$

We can now use these two approximations to investigate the change in the portfolio value. We shall do this in two stages so as not to be too overwhelmed by the mathematical notation.

14.2.1 Delta approximation for the portfolio

Using (14.7) we can immediately deduce that the delta approximation to the change in the portfolio value is given by

$$\begin{aligned}\Delta V_{\text{non}}(t) \approx \Delta V_{\text{non}}^{\delta}(t) &= \sum_{i=1}^{n} \alpha_i \left(\frac{\partial g_i}{\partial t} \Delta t + \sum_{r=1}^{k} b_{ir} \frac{\partial g_i}{\partial f_r} \Delta f_r \right) \\ &= \left(\sum_{i=1}^{n} \alpha_i \frac{\partial g_i}{\partial t} \right) \Delta t + \sum_{r=1}^{k} \left(\sum_{i=1}^{n} \alpha_i b_{ir} \frac{\partial g_i}{\partial f_r} \right) \Delta f_r.\end{aligned}$$

To simplify this expression we use the fact that the gradient of a financial derivative with respect to time is usually referred to as the theta of the derivative; its value serves as a measure of the sensitivity of the derivative product to the passage of time. In view of this we let

$$\theta_t = \sum_{i=1}^{n} \alpha_i \frac{\partial g_i}{\partial t} \tag{14.9}$$

denote the aggregate theta of the portfolio. By the same argument we also let

$$\delta_r = \sum_{i=1}^{n} \alpha_i b_{ir} \frac{\partial g_i}{\partial f_r}, \quad r = 1, \ldots, k \tag{14.10}$$

represent the aggregate delta of the portfolio with respect to risk factor f_r. The expression for the delta approximation for the change in portfolio value can then be expressed more succinctly as

$$\Delta V_{\text{non}}^{\delta}(t) = \theta_t \Delta t + \sum_{r=1}^{k} \delta_r \Delta f_r.$$

We notice that the sum in the above expression is simply the dot product between the two vectors

$$\boldsymbol{\delta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_k \end{pmatrix} \quad \text{and} \quad \mathbf{\Delta f} = \begin{pmatrix} \Delta f_1 \\ \vdots \\ \Delta f_k \end{pmatrix} \tag{14.11}$$

and so, in vector notation, we can write

$$\Delta V_{\text{non}}^{\delta}(t) = \theta_t \Delta t + \boldsymbol{\delta}^T \mathbf{\Delta f}. \tag{14.12}$$

14.2.2 Gamma approximation for the portfolio

Using precisely the same approach as for the delta approximation we can use (14.8) to show that the gamma approximation for the change in portfolio value is given by

$$\begin{aligned}\Delta V_{\text{non}}(t) \approx \Delta V_{\text{non}}^{\gamma}(t) &= \sum_{i=1}^{n} \alpha_i \left(\Delta^{\delta} g_i + \frac{1}{2} \sum_{r=1}^{k} \sum_{s=1}^{k} b_{ir} b_{is} \frac{\partial^2 g_i}{\partial f_r \partial f_s} \Delta f_r \Delta f_s \right) \\ &= \theta_t \Delta t + \sum_{r=1}^{k} \delta_r \Delta f_r + \frac{1}{2} \sum_{r=1}^{k} \sum_{s=1}^{k} \Delta f_r \left(\sum_{i=1}^{n} \alpha_i b_{ir} b_{is} \frac{\partial^2 g_i}{\partial f_r \partial f_s} \right) \Delta f_s.\end{aligned}$$

In analogy with the development for the delta approximation, we let

$$\Gamma_{rs} = \sum_{i=1}^{n} \alpha_i b_{ir} b_{is} \frac{\partial^2 g_i}{\partial f_r \partial f_s} (1 \leq r, s \leq k) \tag{14.13}$$

denote the aggregate mixed gamma of the portfolio with respect to risk factors f_r and f_s. The approximation can then be expressed as

$$\Delta V_{\text{non}}^{\gamma}(t) = \theta_t \Delta t + \sum_{r=1}^{k} \delta_r \Delta f_r + \frac{1}{2} \sum_{r=1}^{k} \sum_{s=1}^{k} \Delta f_r \Gamma_{rs} \Delta f_s.$$

We make one final simplification by collecting and storing the aggregate gamma terms in a $k \times k$ symmetric matrix $\mathbf{\Gamma}$, the approximation can then be written as the following quadratic form:

$$\Delta V_{\text{non}}^{\gamma}(t) = \theta_t \Delta t + \boldsymbol{\delta}^T \mathbf{\Delta f} + \frac{1}{2} \mathbf{\Delta f}^T \mathbf{\Gamma} \mathbf{\Delta f}. \tag{14.14}$$

We shall assume that the matrix $\mathbf{\Gamma}$ is invertible, in which case we can complete the square, i.e., employ (4.13) to give

$$\Delta V_{\text{non}}^{\gamma}(t) = \theta_t \Delta t - \frac{1}{2} \boldsymbol{\delta}^T \mathbf{\Gamma}^{-1} \boldsymbol{\delta} + \frac{1}{2} \left(\mathbf{\Delta f} + \mathbf{\Gamma}^{-1} \boldsymbol{\delta} \right)^T \mathbf{\Gamma} \left(\mathbf{\Delta f} + \mathbf{\Gamma}^{-1} \boldsymbol{\delta} \right). \tag{14.15}$$

We close the current discussion by highlighting the fact that (14.14) simplifies considerably if a single-factor model is employed, i.e., if we assume

$$S_i(t) = a_i + b_i f(t), \quad i = 1, \ldots, n. \tag{14.16}$$

In this case, the gamma approximation for the change in portfolio value is given by the straightforward quadratic equation

$$\begin{aligned}\Delta V_{\text{non}}^{\gamma}(t) &= \theta_t \Delta t + \underbrace{\left(\sum_{i=1}^{n} \alpha_i b_i \frac{\partial g_i}{\partial f} \right)}_{=\delta} \Delta f + \frac{1}{2} \underbrace{\left(\sum_{i=1}^{n} \alpha_i b_i^2 \frac{\partial^2 g_i}{\partial f^2} \right)}_{=\gamma} (\Delta f)^2 \\ &= \theta_t \Delta t + \delta \Delta f + \frac{1}{2} \gamma (\Delta f)^2,\end{aligned}$$

or equivalently, by completing the square, we have

$$\begin{aligned}\Delta V^{\gamma}_{\text{non}}(t) &= \theta_t \Delta t - \frac{\delta^2}{2\gamma} + \frac{\gamma}{2}\left(\Delta f + \frac{\delta}{\gamma}\right)^2 \\ &= \theta_t \Delta t - \frac{1}{2}\delta\gamma^{-1}\delta + \frac{1}{2}(\Delta f + \gamma^{-1}\delta)\gamma(\Delta f + \gamma^{-1}\delta). \end{aligned} \tag{14.17}$$

We remark that the final line of (14.17) has been provided to demonstrate the analogy between this expression and its multivariate generalization (14.15). We shall find both of these expressions useful in the next section, where we develop VaR estimates for non-linear portfolios.

14.3 VALUE AT RISK FOR DERIVATIVE PORTFOLIOS

In the previous section we developed simple approximations for the change in value $\Delta V_{\text{non}}(t)$ of a non-linear portfolio (14.2) over a time period Δt. We shall now demonstrate how these approximations can be used to provide VaR estimates for the portfolio. We note that the potential loss for the portfolio is given by the random variable

$$\Delta \mathcal{L}_{\text{non}}(t) = -\Delta V_{\text{non}}(t)$$

and so, for a given confidence level α, the Value at Risk for the portfolio is the number VaR_α that satisfies

$$\mathbb{P}[\Delta \mathcal{L}_{\text{non}}(t) \leq \text{VaR}_\alpha] = \alpha.$$

We now present a mathematical investigation comprised of the following three stages:

1. Value at Risk with a multi-factor delta model.
2. Value at Risk with a single-factor gamma model.
3. Value at Risk with a multi-factor gamma model.

In each stage we shall assume, as in the linear case, that the changes in the underlying driving factors are normally distributed.

14.3.1 Multi-factor delta approximation

Under the multi-factor model (14.6) we can employ (14.12) to approximate the potential loss of the non-linear portfolio as

$$\Delta \mathcal{L}_{\text{non}}(t) \approx \Delta \mathcal{L}^{\delta}_{\text{non}}(t) = -\theta_t \Delta t - \boldsymbol{\delta}^T \mathbf{\Delta f}. \tag{14.18}$$

Under this model we say that the delta-VaR for some confidence level α is the value $\text{VaR}^{\delta}_{\alpha}$ that satisfies

$$\mathbb{P}[\Delta \mathcal{L}^{\delta}_{\text{non}}(t) \leq \text{VaR}^{\delta}_{\alpha}] = \alpha.$$

Now, if we assume that the vector $\mathbf{\Delta f}$ is joint normally distributed with zero mean and covariance matrix $\mathbf{\Omega}_{\Delta f}$, then we can conclude from (14.18) that

$$\mathcal{L}^{\delta}_{\text{non}}(t) \sim N\left(-\theta_t \Delta t, \boldsymbol{\delta}^T \mathbf{\Omega}_{\Delta f} \boldsymbol{\delta}\right)$$

and, consequently, we can use the analytic formula for normal-VaR to deduce that

$$\mathrm{VaR}_\alpha^\delta = -\theta_t \Delta t + \Phi^{-1}(\alpha)\sqrt{\boldsymbol{\delta}^T \boldsymbol{\Omega}_{\Delta f} \boldsymbol{\delta}}. \tag{14.19}$$

We say that $\mathrm{VaR}_\alpha^\delta$ is the delta-normal approximation to the VaR for the non-linear portfolio (14.2).

14.3.2 Single-factor gamma approximation

Under the single-factor model (14.16) we can employ (14.17) to approximate the potential loss of the non-linear portfolio as

$$\Delta \mathcal{L}_{\mathrm{non}}(t) \approx \Delta \mathcal{L}_{\mathrm{non}}^\gamma(t) = \frac{\delta^2}{2\gamma} - \theta_t \Delta t - \frac{\gamma}{2}\left(\Delta f + \frac{\delta}{\gamma}\right)^2. \tag{14.20}$$

Under this model we say that the delta-VaR for some confidence level α is the value $\mathrm{VaR}_\alpha^\gamma$ that satisfies

$$\mathbb{P}[\Delta \mathcal{L}_{\mathrm{non}}^\gamma(t) \leq \mathrm{VaR}_\alpha^\gamma] = \alpha.$$

Now, if we assume that $\Delta f \sim N(0, \sigma^2)$ then it follows that

$$\Delta f + \frac{\delta}{\gamma} \sim N\left(\frac{\delta}{\gamma}, \sigma^2\right).$$

Now we recall Definition 12.5, which tells us that

$$\left(\Delta f + \frac{\delta}{\gamma}\right)^2 \sim \chi_1^2\left(\sigma, \left(\frac{\delta}{\gamma}\right)^2\right).$$

We can use this information, together with (14.20), to deduce that

$$\left(\frac{\Delta f + \frac{\delta}{\gamma}}{\sigma}\right)^2 = \frac{\frac{\delta^2}{2\gamma} - \theta_t \Delta t - \Delta \mathcal{L}_{\mathrm{non}}^\gamma(t)}{\frac{1}{2}\gamma\sigma^2} \sim \chi_1^2\left(\left(\frac{\delta}{\sigma\gamma}\right)^2\right).$$

The density function for this random variable is given by

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(x + \frac{\delta^2}{\gamma^2\sigma^2}\right)\right) \frac{\cosh(\delta\sqrt{x}/\gamma\sigma)}{\sqrt{x}}$$

and, for a given confidence level α, numerical algorithms can be used to find the $(1-\alpha)$-quantile $x_{1-\alpha}$ that satisfies

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_{1-\alpha}} \exp\left(-\frac{1}{2}\left(x + \frac{\delta^2}{\gamma^2\sigma^2}\right)\right) \frac{\cosh(\delta\sqrt{x}/\gamma\sigma)}{\sqrt{x}} dx = 1 - \alpha.$$

Once we have this value we can write

$$\mathbb{P}\left[\frac{\frac{\delta^2}{2\gamma} - \theta_t \Delta t - \Delta\mathcal{L}^{\gamma}_{\text{non}}(t)}{\frac{1}{2}\gamma\sigma^2} \leq x_{1-\alpha}\right] = 1-\alpha,$$

or equivalently,

$$\begin{aligned}
&\mathbb{P}\left[-\Delta\mathcal{L}^{\gamma}_{\text{non}}(t) \leq \theta_t \Delta t - \frac{\delta^2}{2\gamma} + \frac{1}{2}\gamma\sigma^2 x_{1-\alpha}\right] = 1-\alpha \\
\Rightarrow \quad &\mathbb{P}\left[\Delta\mathcal{L}^{\gamma}_{\text{non}}(t) > -\theta_t \Delta t + \frac{\delta^2}{2\gamma} - \frac{1}{2}\gamma\sigma^2 x_{1-\alpha}\right] = 1-\alpha \\
\Rightarrow \quad &1 - \mathbb{P}\left[\Delta\mathcal{L}^{\gamma}_{\text{non}}(t) \leq -\theta_t \Delta t + \frac{\delta^2}{2\gamma} - \frac{1}{2}\gamma\sigma^2 x_{1-\alpha}\right] = 1-\alpha.
\end{aligned}$$

The final equation allows us to deduce that

$$\mathbb{P}\left[\Delta\mathcal{L}^{\gamma}_{\text{non}}(t) \leq -\theta_t \Delta t + \frac{\delta^2}{2\gamma} - \frac{1}{2}\gamma\sigma^2 x_{1-\alpha}\right] = \alpha$$

and this, in turn, allows us to read off the gamma VaR estimate, it is given by

$$\text{VaR}^{\gamma}_{\alpha} = -\theta_t \Delta t + \frac{\delta^2}{2\gamma} - \frac{1}{2}\gamma\sigma^2 x_{1-\alpha}.$$

14.3.3 Multi-factor gamma approximation

We now turn to the more realistic setting where a multi-factor model (14.6) of k risk factors is assumed to drive the fluctuations in the n underlying asset prices. We have shown (14.15) that, in this setting, the gamma approximation to the change in value of the portfolio over a time period Δt is given by

$$\begin{aligned}
\Delta V^{\gamma}_{\text{non}}(t) = \theta_t \Delta t &- \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta} \\
&+ \frac{1}{2}\left(\boldsymbol{\Delta f} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta}\right)^T \boldsymbol{\Gamma}\left(\boldsymbol{\Delta f} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta}\right).
\end{aligned} \tag{14.21}$$

Let us now assume that $\boldsymbol{\Delta f}$, the vector of risk factor changes, has the multivariate normal distribution with zero mean vector and covariance matrix $\boldsymbol{\Omega}_{\boldsymbol{\Delta f}}$. We shall assume that $\boldsymbol{\Omega}_{\boldsymbol{\Delta f}}$ is positive definite, in which case it possesses a square-root decomposition (see Theorem 2.6):

$$\boldsymbol{\Omega}_{\boldsymbol{\Delta f}} = \left(\boldsymbol{\Omega}^{1/2}_{\boldsymbol{\Delta f}}\right)\left(\boldsymbol{\Omega}^{1/2}_{\boldsymbol{\Delta f}}\right)^T, \tag{14.22}$$

where

$\boldsymbol{\Omega}^{1/2}_{\boldsymbol{\Delta f}}$ is a lower triangular matrix

whose inverse

$$\boldsymbol{\Omega}_{\mathbf{\Delta f}}^{-1/2} = \left(\boldsymbol{\Omega}_{\mathbf{\Delta f}}^{1/2}\right)^{-1} \text{ is also lower triangular.}$$

Equipped with these facts we can now make the following deduction:

$$\begin{aligned}\mathbf{\Delta f} \sim N(\mathbf{0}, \boldsymbol{\Omega}_{\mathbf{\Delta f}}) &\Rightarrow \mathbf{\Delta f} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta} \sim N\left(\boldsymbol{\Gamma}^{-1}\boldsymbol{\delta}, \boldsymbol{\Omega}_{\mathbf{\Delta f}}\right) \\ &\Rightarrow \boldsymbol{\Omega}_{\mathbf{\Delta f}}^{-1/2}\left(\mathbf{\Delta f} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta}\right) \sim N\left(\boldsymbol{\Omega}_{\mathbf{\Delta f}}^{-1/2}\boldsymbol{\Gamma}^{-1}\boldsymbol{\delta}, \mathbf{I}_k\right).\end{aligned}$$

Thus we can see that the above shift and rescaling has the effect of transforming $\mathbf{\Delta f}$ to a new random vector

$$\mathbf{\Delta y} := \boldsymbol{\Omega}_{\mathbf{\Delta f}}^{-1/2}\left(\mathbf{\Delta f} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta}\right) \tag{14.23}$$

whose components are independent normal random variables. Furthermore, using (14.23) we see that

$$\mathbf{\Delta f} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta} = \boldsymbol{\Omega}_{\mathbf{\Delta f}}^{1/2}\mathbf{\Delta y}$$

and we can substitute this into (14.21) to yield

$$\begin{aligned}\Delta V_{\text{non}}^{\gamma}(t) = \theta_t \Delta t &- \frac{1}{2}\boldsymbol{\delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\delta} \\ &+ \frac{1}{2}\Delta\mathbf{y}^T\left(\boldsymbol{\Omega}_{\mathbf{\Delta f}}^{1/2}\right)^T\boldsymbol{\Gamma}\boldsymbol{\Omega}_{\mathbf{\Delta f}}^{1/2}\Delta\mathbf{y}.\end{aligned}$$

To make the presentation neater we let

$$\mathbf{A} = \left(\boldsymbol{\Omega}_{\mathbf{\Delta f}}^{1/2}\right)^T\boldsymbol{\Gamma}\boldsymbol{\Omega}_{\mathbf{\Delta f}}^{1/2} \tag{14.24}$$

and so we have

$$\Delta V_{\text{non}}^{\gamma}(t) = \theta_t \Delta t - \frac{1}{2}\boldsymbol{\delta}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\delta} + \frac{1}{2}\mathbf{\Delta y}^T\mathbf{A}\Delta\mathbf{y}. \tag{14.25}$$

The final and crucial step of the analysis is to notice that $\mathbf{A}$ (14.24) is a real symmetric $k \times k$ matrix, and so we can employ the spectral theorem (Theorem 2.4) to write

$$\mathbf{A} = \mathbf{P}\,\text{diag}(\lambda_1, \ldots, \lambda_k)\,\mathbf{P}^T,$$

where $\{\lambda_1, \ldots, \lambda_k\}$ denote the k real eigenvalues of $\mathbf{A}$ and where $\mathbf{P}$ is the $k \times k$ matrix whose jth column is the eigenvector of $\mathbf{A}$ corresponding to the eigenvalue λ_j, for $j = 1, \ldots, k$.

We now make one final variable transformation by setting

$$\mathbf{\Delta z} = \mathbf{P}^T\mathbf{\Delta y} \sim N\left(\mathbf{P}^T\boldsymbol{\Omega}_{\mathbf{\Delta f}}^{-1/2}\boldsymbol{\Gamma}^{-1}\boldsymbol{\delta},\ \underbrace{\mathbf{P}^T\mathbf{P}}_{=\mathbf{I}_k}\right).$$

Substituting this into (14.25) we find that

$$\begin{aligned}\Delta V^{\gamma}_{\text{non}}(t) &= \theta_t \Delta t - \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\Delta}\mathbf{z}^T \ \text{ diag } (\lambda_1, \ldots, \lambda_k)\ \boldsymbol{\Delta}\mathbf{z} \\ &= \theta_t \Delta t - \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta} + \frac{1}{2}\sum_{j=1}^{k}\lambda_j(\Delta z_j)^2. \end{aligned} \tag{14.26}$$

Equivalently, we can conclude that, under the multi-factor gamma model, the approximate potential loss of the non-linear portfolio (14.2) is given by

$$\Delta \mathcal{L}^{\gamma}_{\text{non}}(t) = -\theta_t \Delta t + \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta} - \frac{1}{2}\sum_{j=1}^{k}\lambda_j(\Delta z_j)^2. \tag{14.27}$$

Thus we have reduced the approximate loss random variable $\mathcal{L}^{\gamma}_{\text{non}}(t)$ to a constant plus a linear combination of squares of k unit variance normal random variables or, in other words, a linear combination of k non-central chi-squared random variables with one degree of freedom. Thus, in order to estimate the VaR under this approximation, we need access to the distribution of the random variable

$$\mathcal{Z} = \sum_{j=1}^{k}\lambda_j(\Delta z_j)^2.$$

To illustrate, let us suppose we know the distribution of $\mathcal{Z}$ and so we can find the $(1-\alpha)$-quantile, i.e., the number $x_{1-\alpha}$ that satisfies

$$\mathbb{P}[\mathcal{Z} \leq x_{1-\alpha}] = 1 - \alpha.$$

We can then follow the same steps as in the single-factor case to finally conclude that the multi-factor gamma estimate of VaR is given by

$$\text{VaR}^{\Gamma}_{\alpha} = -\theta_t \Delta t + \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{\Gamma}^{-1}\boldsymbol{\delta} - \frac{1}{2}x_{1-\alpha}. \tag{14.28}$$

Unfortunately, unlike the single-factor case, the distribution of a random variable $\mathcal{Z}$ is not known. However, all is not lost as we can gain fresh insight by focusing upon the characteristic function of $\mathcal{Z}$. Specifically, we recall from Chapter 11 (Section 11.3) that the characteristic function of the non-central chi-squared random variable $(\Delta z_j)^2$, where $\Delta z_j \sim N(\mu_j, 1)$, is given by

$$\phi_{(\Delta z_j)^2}(u) = \mathbb{E}[\exp(u(\Delta z_j)^2)] = \exp\left(\frac{i\mu_j^2 u}{1-2iu}\right)\frac{1}{\sqrt{1-2iu}}.$$

We assume that $\{(\Delta z_j)^2 : j = 1, \ldots, k\}$ are independent random variables and that the coefficients $(\lambda_j)_{j=1}^{k}$ are distinct, in which case we can deduce that

$$\phi_{\mathcal{Z}}(u) = \mathbb{E}[\exp(u\mathcal{Z})] = \mathbb{E}\left[\exp\left(u\sum_{j=1}^{k}\lambda_j(\Delta z_j)^2\right)\right]$$

$$
\begin{aligned}
&= \mathbb{E}[\exp(u\lambda_1(\Delta z_1)^2)] \cdots \mathbb{E}[\exp(u\lambda_k(\Delta z_k)^2)] \\
&= \prod_{j=1}^{k} \exp\left(\frac{i\mu_j^2\lambda_j u}{1-2i\lambda_j u}\right) \frac{1}{\sqrt{1-2i\lambda_j u}} \\
&= \exp\left(i\sum_{j=1}^{k} \frac{\mu_j^2\lambda_j u}{1-2i\lambda_j u}\right) \prod_{j=1}^{k} \frac{1}{\sqrt{1-2i\lambda_j u}}.
\end{aligned}
$$

Given this closed-form representation for the characteristic function we appeal to the Gil-Pelaez formula (11.11) and immediately write that the distribution of $\mathcal{Z}$ is given by

$$
F(x) = \frac{1}{2} + \frac{1}{2\pi}\int_0^\infty \frac{\exp(iux)\phi_{\mathcal{Z}}(-u) - \exp(-iux)\phi_{\mathcal{Z}}(u)}{iu} du. \tag{14.29}
$$

Unfortunately, there is no known closed-form expression for this integral, however it can be computed using numerical methods; see Imhof (1961) for a popular approach.

We close this chapter by emphasizing that the problem of calculating VaR for a large portfolio of derivatives products is a difficult one. In practice, the most popular approach is to employ a numerical simulation (see Chapter 20), however the drawback of this route is that the calculation process can take a very long time. In view of this there is clearly a market for accurate and stable approximations and, in this chapter, we have developed a useful mathematical framework where such estimates can be derived; the interested reader is encouraged to consult Britten-Jones and Schaeffer (1999) for the finer details and also Mina and Ulmer (1999) where the authors compare four alternative approaches, including the one we have developed here.

Essential Mathematics for Market Risk Management
by Simon Hubbert

15 Time Series Analysis

A classic image that is frequently used to capture the essence of modern finance is that of the analyst whose face is fixed firmly in front of a computer screen displaying feeds of ever-changing financial data; colourful graphs, tables and charts that flicker in response to changing financial markets. In mathematical terms these data feeds are called time series and they represent evolution of a random process at discrete time intervals (or ticks). The successful risk manager must possess the quantitative skills to deal with financial time series data, e.g., to build realistic mathematical models, to test a given hypothesis and to forecast future values. From a theoretical perspective we assume that a given time series has an infinite past and will stretch out into the infinite future and we capture this via a bi-infinite sequence

$$\{X_t : t \in \mathbb{Z}\} = \{\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots\}. \tag{15.1}$$

Clearly a bi-infinite time series is an idealization, yet it covers the case where the observations can be regarded as outcomes of an experiment (e.g., tossing a coin) that can, theoretically, be repeated infinitely often. The purpose of this chapter is to develop the theory behind some of the most popular time series models.

15.1 STATIONARY PROCESSES

There are many ways in which a time series can behave, it may exhibit some long-term trend, it may reveal some seasonal patterns, we may see some abrupt changes as time evolves, and so on. One important feature occurs if we discover that the distribution of the process remains unchanged from one realization to the next; if this is the case we say that X_t is a stationary process and its definition is as follows:

Definition 15.1. *A random process* $\{X_t : t \in \mathbb{Z}\}$ *is said to be* strictly stationary *if its probabilistic laws remain unchanged through time. Specifically, if, for every* $n \geq 1$ *and every selection* $t_1, \ldots, t_n$*, of distinct integers, we have that*

$$\text{the joint distribution of } \begin{pmatrix} X_{t_1} \\ \vdots \\ X_{t_n} \end{pmatrix} = \text{the joint distribution of } \begin{pmatrix} X_{t_1+\tau} \\ \vdots \\ X_{t_n+\tau} \end{pmatrix}$$

for every $\tau \in \mathbb{Z}$.

In practice it is difficult to verify whether or not a random process is strictly stationary since the requirements are too strong. In order to circumvent this an alternative definition is used which focuses only on the mean and the covariance structure of the process and

thus avoids imposing heavy demands on the whole distribution. This weaker definition is as follows:

Definition 15.2. *A random process* $\{X_t : t \in \mathbb{Z}\}$ *is said to be* weakly stationary *if*

- *the mean of the process remains constant through time, i.e.,* $\mathbb{E}[X_t] = \mu$ *for all* $t \in \mathbb{Z}$*;*
- *the auto-covariance of the process is a bounded function of the time lag* τ *alone, i.e., there exists* $\gamma : \mathbb{Z} \to \mathbb{R}$ *such that*

$$cov(X_t, X_{t-\tau}) = \gamma(\tau) < \infty \text{ for all } \tau \in \mathbb{Z}.$$

It is helpful to point out some of the differences between the strict and weak definitions of stationarity and also to emphasize why weak stationarity is a useful concept.

1. A process that is strictly stationary is automatically weakly stationary provided its auto-covariance function is bounded.
2. A process that is weakly stationary is not necessarily strictly stationary.
3. An important special case where weak stationarity implies strict stationarity arises if we assume that the process X_t is normally distributed, since, in this case, the distribution is completely determined by its constant mean μ and its constant variance $\sigma^2 = \gamma(0)$. More accurately, we assume that for every $n \geq 1$ and every selection $t_1, \ldots, t_n$ of distinct integers, the joint distribution of the vector $(X_{t_1}, \ldots, X_{t_n})^T$ is multivariate normal.

For practical applications the notion of weak stationarity provides the practitioner with a nice manageable description of random processes that do not exhibit a deterministic trend or a seasonal cycle. For this reason we shall, from now on, refer to weakly stationary processes simply as stationary processes.

Examples

15.1.1 Purely random processes

A purely random process is a simple but important example of a stationary process and arises when the sequence $\{X_t : t \in \mathbb{Z}\}$ consists of uncorrelated random variables having constant mean μ and constant variance σ^2. Clearly the auto-correlation function for such a process is given by

$$\rho(\tau) = \text{correl } (X_t, X_{t-\tau}) = \begin{cases} 1 & \text{when} \quad \tau = 0, \\ 0 & \text{when} \quad \tau \neq 0, \end{cases} \quad \text{for all } t \in \mathbb{Z}. \tag{15.2}$$

15.1.2 White noise processes

If the mean of a purely random process is zero then we have what is known as a white noise process; terminology taken from engineering applications. We usually reserve the notation $\{\varepsilon_t : t \in \mathbb{Z}\}$ for white noise.

We know from Chapter 3 that if X and Y are independent then they are uncorrelated, however, if X and Y are uncorrelated then they are **not** necessarily independent. In view

of this we can strengthen the definition of white noise, replacing the pairwise uncorrelated condition with independence. This leads to the following definition:

Definition 15.3. *A process $\{\varepsilon_t : t \in \mathbb{Z}\}$ is said to constitute a strict white noise process if the random variables in the sequence are independent and have zero mean and constant variance.*

A special example of strict white noise arises when the random variables of the sequence are normally distributed; in this case we say that $\{\varepsilon_t : t \in \mathbb{Z}\}$ is a Gaussian white noise process.

15.1.3 Random walk processes

A simple but important example of a non-stationary process can be constructed from a purely random process $\{Z_t : t \in \mathbb{Z}\}$. The construction is as follows:

- Define a time-dependent process, starting at time $t = 0$, by setting

$$Y_0 = 0 \text{ and then } Y_t = Y_{t-1} + Z_t \quad \text{for} \quad t = 1, 2 \ldots \tag{15.3}$$

- The definition of the process $(Y_t)_{t=0}^{\infty}$ allows us to write

$$Y_t = Z_1 + \cdots + Z_t \text{ for } t = 1, 2 \ldots$$

  and, with this representation, it is clear that the process is not stationary since both its mean and variance are time-dependent functions given by

$$\mathbb{E}[Y_t] = t\mu \text{ and } \mathbb{E}[(Y_t - \mu)^2] = t\sigma^2.$$

- Any process $(Y_t)_{t=0}^{\infty}$ that can be expressed as (15.3) is said to be a random walk. In other words, the process $(Y_t)_{t=0}^{\infty}$ is a random walk if

$$\{\nabla Y_t = Y_t - Y_{t-1} : \ t = 1, 2, \ldots\} \text{ is a purely random process.}$$

We call ∇ the difference operator and we see that it has the effect of transforming the non-stationary (random walk) process into a stationary (purely random) process.

15.2 MOVING AVERAGE PROCESSES

The defining property of the random walk process is that it is initiated at a fixed time point ($t = 0$) where its value is set equal to zero ($Y_0 = 0$), the process is then allowed to evolve so that its value at any future time t is the sum of t realizations of a purely random process, i.e.,

$$Y_t = \sum_{\tau=0}^{t-1} Z_{t-\tau} \text{ where } (Z_t)_{t=0}^{\infty} \text{ is purely random.}$$

Suppose that we start with a white noise process $(\varepsilon_t)_{t\in\mathbb{Z}}$ say, then we can construct a zero-mean random walk by setting

$$Y_0 = 0 \text{ and } Y_t = \sum_{\tau=0}^{t-1} \varepsilon_{t-\tau} \text{ for } t = 1, 2, \ldots$$

We note that the full history of the process drives its next future value.

A moving average process can be constructed in a similar way with the important exception that we do not impose a starting point. In contrast, we choose a positive integer q, a set of real coefficients $\{\beta_0, \ldots, \beta_q\}$ and we define a new process $(Y_t)_{t\in\mathbb{Z}}$ by setting

$$Y_t = \beta_0\varepsilon_t + \beta_1\varepsilon_{t-1} + \cdots + \beta_q\varepsilon_{t-q} \text{ for all } t \in \mathbb{Z}. \tag{15.4}$$

We see immediately that Y_t depends, in a linear fashion, on the recent history of the white noise process; the length of the historical period is determined by the value of q. We say that $(Y_t)_{t\in\mathbb{Z}}$ is a moving average process of order q, or more succinctly $(Y_t)_{t\in\mathbb{Z}}$ is MA(q).

The process has zero mean and its variance is given by

$$\begin{aligned}\text{variance } (Y_t) = \mathbb{E}[Y_t^2] &= \mathbb{E}\left[\left(\sum_{i=0}^{q}\beta_i\varepsilon_{t-i}\right)\left(\sum_{j=0}^{q}\beta_j\varepsilon_{t-j}\right)\right] \\ &= \sum_{i=0}^{q}\sum_{j=0}^{q}\beta_i\beta_j\mathbb{E}[\varepsilon_{t-i}\varepsilon_{t-j}]. \end{aligned} \tag{15.5}$$

Using the independence property of the white noise process we know that

$$\mathbb{E}[\varepsilon_{t-i}\varepsilon_{t-j}] = \text{cov}(\varepsilon_{t-i}, \varepsilon_{t-j}) = \begin{cases} 0 & i \neq j, \\ \sigma_\varepsilon^2 & i = j. \end{cases}$$

and so (15.5) becomes

$$\text{variance}(Y_t) = \sigma_\varepsilon^2 \sum_{i=0}^{q} \beta_i^2.$$

In a similar fashion we can compute the auto-covariance

$$\begin{aligned}\text{cov}(Y_t, Y_{t-\tau}) &= \mathbb{E}\left[\left(\sum_{i=0}^{q}\beta_i\varepsilon_{t-i}\right)\left(\sum_{j=0}^{q}\beta_j\varepsilon_{t-j-\tau}\right)\right] \\ &= \sum_{i=0}^{q}\sum_{j=0}^{q}\beta_i\beta_j\mathbb{E}[\varepsilon_{t-i}\varepsilon_{t-j-\tau}]\end{aligned}$$

$$= \sum_{i=0}^{q} \sum_{k=\tau}^{q+\tau} \beta_i \beta_k \mathbb{E}[\varepsilon_{t-i}\varepsilon_{t-k}] \quad (\text{setting } k = j + \tau)$$

$$= \begin{cases} 0 & \text{if } \tau > q; \\ \sigma_\varepsilon^2 \sum_{i=\tau}^{q} \beta_i \beta_{i+\tau} & \text{if } \tau = 0, 1, \ldots, q. \end{cases}$$

We note that the auto-covariance of this process does not depend upon the time index t. This observation, together with the fact that process mean and variance are finite, allows us to deduce that a moving average process is yet another example of a stationary process.

15.3 AUTO-REGRESSIVE PROCESSES

In real-world situations we often find evidence which suggests that a random process is influenced by its own recent history. In view of this we select a time frame of length p (a positive integer) and then, having observed p realizations of the process, we define

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + \cdots + \alpha_p Y_{t-p} + \varepsilon_t, \tag{15.6}$$

where $(\varepsilon_t)_{t\in\mathbb{Z}}$ represents a white noise process whose variance we denote as σ_ε^2. The coefficients $(\alpha_k)_{k=0}^{p}$ are obtained from a linear regression; a technique we originally encountered in Chapter 4. Given that our new process (15.6) can be viewed as a regression on its past p values, we say that Y_t is an auto-regressive process of order p or, more succinctly, $(Y_t)_{t\in\mathbb{Z}}$ is AR(p).

We now investigate this process for the case $p = 1$, where the process is assumed to be influenced by its previous value. The model, in this case, can be developed as follows:

$$\begin{aligned} Y_t &= \omega + \alpha Y_{t-1} + \varepsilon_t \\ &= \omega + \alpha \left[\omega + \alpha Y_{t-2} + \varepsilon_{t-1}\right] + \varepsilon_t \\ &= \omega(1+\alpha) + \alpha^2 Y_{t-2} + \alpha_1 \varepsilon_{t-1} + \varepsilon_t. \end{aligned}$$

One further re-substitution gives

$$\begin{aligned} Y_t &= \omega(1+\alpha) + \alpha^2 \left[\omega + \alpha Y_{t-3} + \varepsilon_{t-1}\right] + \alpha\varepsilon_{t-1} + \varepsilon_t \\ &= \omega(1+\alpha+\alpha^2) + \alpha^3 Y_{t-3} + \alpha^2\varepsilon_{t-2} + \alpha\varepsilon_{t-1} + \varepsilon_t. \end{aligned}$$

The pattern persists if we continue re-substituting and we find that, in the limit, the model can be written alternatively as

$$Y_t = \omega \sum_{k=0}^{\infty} \alpha^k + \sum_{k=0}^{\infty} \alpha^k \varepsilon_{t-k}. \tag{15.7}$$

We note that the process is only stable if $|\alpha| < 1$, in which case we can use the geometric series identity

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \quad \text{provided} \quad |x| < 1 \tag{15.8}$$

to deduce that

$$Y_t = \frac{\omega}{1-\alpha} + \sum_{k=0}^{\infty} \alpha^k \varepsilon_{t-k}. \tag{15.9}$$

Comparing this representation to (15.4) we see that an AR(1) process can be viewed as a shifted moving average process of infinite order whose coefficients are given by $(\alpha^k)_{k=0}^{\infty}$. This alternative view is very helpful, for instance we can immediately deduce that

$$\mathbb{E}[Y_t] = \frac{\omega}{1-\alpha} \tag{15.10}$$

and that

$$\begin{aligned}
\text{variance}(Y_t) &= \text{variance}\left(\sum_{k=0}^{\infty} \alpha^k \varepsilon_{t-k}\right) \\
&= \sum_{k=0}^{\infty} \text{variance}(\alpha^k \varepsilon_{t-k}) \quad \text{(using independence)} \\
&= \sigma_\varepsilon^2 (1 + \alpha^2 + \alpha^4 + \cdots) \\
&= \frac{\sigma_\varepsilon^2}{1-\alpha^2}
\end{aligned} \tag{15.11}$$

The auto-covariance of this process (against a version at time lag $\tau > 0$) is calculated in a similar fashion as

$$\begin{aligned}
\text{cov}(Y_t, Y_{t-\tau}) &= \mathbb{E}\left[\left(\sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i}\right)\left(\sum_{j=0}^{\infty} \alpha^j \varepsilon_{t-j-\tau}\right)\right] \\
&= \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \alpha^i \alpha^j \mathbb{E}[\varepsilon_{t-i}\varepsilon_{t-j-\tau}] \\
&= \sum_{i=0}^{\infty}\sum_{k=\tau}^{\infty} \alpha^i \alpha^k \mathbb{E}[\varepsilon_{t-i}\varepsilon_{t-k}] \quad \text{(setting } k = j+\tau\text{)} \\
&= \sigma_\varepsilon^2 \sum_{i=0}^{\infty} \alpha^i \alpha^{i+\tau} = \frac{\alpha^\tau \sigma_\varepsilon^2}{1-\alpha^2} = \alpha^\tau \ \text{variance}(Y_t).
\end{aligned} \tag{15.12}$$

Using this calculation we can deduce that the auto-correlation function of an AR(1) process depends only upon the time lag τ and is given by

$$\rho(\tau) = \alpha^{\tau} \text{ for } \tau = 0, 1, 2, \ldots \tag{15.13}$$

In summary we have discovered that an AR(1) process with regression coefficient $\alpha \in (-1, 1)$ is a stationary process.

15.4 AUTO-REGRESSIVE MOVING AVERAGE PROCESSES

To complete the survey of the most commonly used time series models we consider the so-called auto-regressive moving average processes which, as their name suggests, represent an augmentation of an AR(p) process and an MA(q) process. Formally,

$$Y_t = \underbrace{\alpha_0 + \sum_{k=1}^{p} \alpha_k Y_{t-k} + \varepsilon_t}_{\text{AR}(p)} + \underbrace{\sum_{k=1}^{q} \beta_k \varepsilon_{t-k}}_{\text{MA}(q)}, \tag{15.14}$$

and, mathematically speaking, we say Y_t described by (15.14) is an ARMA(p, q) process. At first glance, it appears that by introducing even more unknown parameters we are increasing the complexity of the model. However, when this model is employed in practice it is common to find that relatively low values of p and q are needed to ensure a faithful fit to real data. Indeed, one of the most commonly used models is the simplest one; namely ARMA(1,1) which is given by

$$Y_t = \alpha Y_{t-1} + \varepsilon_t + \beta \varepsilon_{t-1}. \tag{15.15}$$

In order to unearth some of the key statistical properties of this process we need to introduce the notion of a lag operator L, which is defined by its action on a time series entry as

$$L(Y_t) = Y_{t-1} \text{ and more generally } L^k(Y_t) = Y_{t-k}.$$

In addition, we let I denote the identity operator whose action on a time series element leaves it unchanged, i.e., $I(Y_t) = Y_t$. These operators are helpful as they allow us to express (15.15) as

$$(I - \alpha L)(Y_t) = (I + \beta L)\varepsilon_t. \tag{15.16}$$

Now, if we assume that $|\alpha| < 1$ then we can deduce that

$$\begin{aligned}
&(I + \alpha L + \alpha^2 L^2 + \alpha^3 L^3 + \cdots)(I - \alpha L) \\
&= I \underbrace{-\alpha L + \alpha L}_{=0} \underbrace{-\alpha^2 L^2 + \alpha^2 L^2}_{=0} \underbrace{-\alpha^3 L^3 + \alpha^3 L^3}_{=0} - \alpha^4 L^4 + \cdots \\
&= I.
\end{aligned}$$

In mathematical terms we say that the operator $\sum_{k=0}^{\infty} \alpha^k L^k$ is the inverse of $I - \alpha L$ and write

$$\sum_{k=0}^{\infty} \alpha^k L^k := (I - \alpha L)^{-1}.$$

Thus, applying this inverse to (15.16) we find

$$\begin{aligned} Y_t &= (I + \alpha + \alpha^2 L^2 + \alpha^3 L^3 + \cdots)(1 + \beta L)(\varepsilon_t) \\ &= \left(I + (\alpha + \beta)L + \alpha(\alpha + \beta)L^2 + \alpha^2(\alpha + \beta)L^3 + \cdots\right)(\varepsilon_t) \\ &= \varepsilon_t + (\alpha + \beta)\left[\varepsilon_{t-1} + \alpha\varepsilon_{t-2} + \alpha^2\varepsilon_{t-3} + \ldots\right]. \end{aligned} \tag{15.17}$$

Using this expression we can immediately deduce that the process has zero mean,

$$\mathbb{E}[Y_t] = 0, \tag{15.18}$$

and that its variance is given by

$$\begin{aligned} \text{variance}(Y_t) &= \left[1 + \frac{(\alpha + \beta)^2}{1 - \alpha^2}\right]\sigma_\varepsilon^2 \\ &= \left[\frac{1 + 2\alpha\beta + \beta^2}{1 - \alpha^2}\right]\sigma_\varepsilon^2. \end{aligned} \tag{15.19}$$

It is also straightforward to compute the auto-covariance of the process. Specifically, we find that

$$\begin{aligned} \text{cov}(Y_t, Y_{t-1}) &= \left[(\alpha + \beta) + \frac{\alpha(\alpha + \beta)^2}{1 - \alpha^2}\right]\sigma_\varepsilon^2 \\ &= \left[\frac{(\alpha + \beta)(1 + \alpha\beta)}{1 - \alpha^2}\right]\sigma_\varepsilon^2. \end{aligned} \tag{15.20}$$

This allows us to deduce that the auto-correlation function at the one-day time lag is given by

$$\rho(1) = \frac{\text{cov}(Y_t, Y_{t-1})}{\text{variance}(Y_t)} = \frac{(\alpha + \beta)(1 + \alpha\beta)}{1 + 2\alpha\beta + \beta^2}. \tag{15.21}$$

To determine the complete auto-correlation function we note that it satisfies

$$\rho(\tau) = \alpha\rho(\tau - 1) \text{ for } \tau \geq 2,$$

and thus we can conclude that

$$\rho(\tau) = \frac{\text{cov}(Y_t, Y_{t-1})}{\text{var}(Y_t)} = \frac{\alpha^{\tau-1}(\alpha+\beta)(1+\alpha\beta)}{1+2\alpha\beta+\beta^2} \text{ for } \tau \geq 1. \tag{15.22}$$

We have discovered that the ARMA(1,1) process with coefficient $\alpha \in (-1, 1)$ and $\beta \geq 0$ is a stationary process.

Essential Mathematics for Market Risk Management
by Simon Hubbert

16
Maximum Likelihood Estimation

In almost all financial institutions there are IT systems in place which collect financial data from a huge array of sources and then deliver them in real time to the computer screen of the end user. A successful practitioner must therefore possess the relevant quantitative skills to make sense of these data. The maximum likelihood method is an extremely popular statistical tool that is used for deriving estimates of unknown distributional parameters of a random process or some statistical experiment where, in theory, infinitely many observations may be taken. The method provides estimates based upon a finite sample of observations. To set the scene we shall assume that a statistical experiment can be performed which reveals the outcome of the random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$. If we let $\mathcal{X}$ denote the sample space of $\mathbf{X}$, i.e., the full range of all possible outcomes of $\mathbf{X}$, then the statistical experiment effectively delivers a sample of size n, i.e., a point $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathcal{X}$. We shall assume that the experiment can be performed repeatedly and that for each run we observe a new sample vector $\mathbf{x}$. We note that the outcome vector itself will vary randomly with each new run of the experiment. In view of this we reserve upper case letters for random variables and random vectors (i.e., the statistical experiment) and lower case letters to denote observed values (i.e., the outcome of the experiment).

The maximum likelihood method itself is based upon the assumption that we know, in advance, the general mathematical form of the joint density function of the random vector $\mathbf{X}$. The density function itself will typically depend upon a set of parameters, $\theta_1, \ldots, \theta_k$ say, and if these are known, say $\theta_i = \theta_i^*$ for $i = 1, \ldots, k$, then the density function is fixed and given by

$$p_{\text{joint}}(\mathbf{x}|\boldsymbol{\theta}^*) \text{ where } \boldsymbol{\theta}^* = (\theta_1^*, \ldots, \theta_k^*)^T \text{ and } \mathbf{x} \in \mathcal{X}.$$

In general, the parameters that define the density function are free to vary and we let Θ denote the parameter space of all possible values that the k-dimensional vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ can feasibly take.

In practice the investigator will run an experiment to produce a sample vector $\mathbf{x} = (x_1, \ldots, x_n)^T$. The task is then to search the parameter space Θ to find the vector $\widehat{\boldsymbol{\theta}} = (\widehat{\theta_1}, \ldots, \widehat{\theta_k})^T$ that, in some sense, provides the best explanation of the outcome vector $\mathbf{x}$. To describe this mathematically we define the so-called likelihood function (associated with our sample data) as

$$L(\boldsymbol{\theta}) = L(\theta_1, \ldots, \theta_k) = p_{\text{joint}}(x_1, \ldots, x_n|\theta_1, \ldots, \theta_k). \tag{16.1}$$

Furthermore, if the components $(X_i)_{i=1}^n$ of $\mathbf{X}$ are independent and identically distributed random variables (as they very often are), then (16.1) can be written in factorized form:

$$L(\boldsymbol{\theta}) = L(\theta_1, \ldots, \theta_k) = \prod_{j=1}^{n} p(x_j|\theta_1, \ldots, \theta_k). \tag{16.2}$$

The likelihood function quantifies how well the parameter vector $\boldsymbol{\theta}$ explains the observed sample and so our aim is to find the parameters that maximize its value. In summary, we must solve the maximum likelihood problem which is set up as follows:

$$\begin{gathered}\text{find the vector } \widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)^T \in \Theta \text{ such that} \\ L(\widehat{\theta}_1, \ldots, \widehat{\theta}_k) \geq L(\theta_1, \ldots, \theta_k) \\ \text{for all parameter vectors } \boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Theta.\end{gathered} \tag{16.3}$$

In many cases it is often easier to maximize the logarithm of the likelihood function, which we write as

$$LL(\theta_1, \ldots, \theta_k) = \log \circ L(\theta_1, \ldots, \theta_k) = \sum_{j=1}^{n} p(x_j|\theta_1, \ldots, \theta_k).$$

This is justified because the log function is strictly increasing and so does not shift the position of the maximum of L. The problem itself is typically solved by employing a suitable mathematical optimization algorithm within a computer package, the resulting solution $\widehat{\boldsymbol{\theta}}$ will be called the maximum likelihood estimate for the true parameter $\boldsymbol{\theta}^*$. We alert the user to the following important remarks:

- It may be the case that a maximum likelihood estimate cannot be found, i.e., for a given sample there may be no vector $\widehat{\boldsymbol{\theta}}$ in the parameter space Θ that maximizes the log-likelihood function. However, in practice such situations are rare.
- For a given sample we may find that there is more than one vector $\widehat{\boldsymbol{\theta}}$ that maximizes the likelihood, i.e., the maximizer is not necessarily unique. Again, this would represent an exceptional situation.

In the case where the parameter space is unrestricted, i.e., $\Theta = \mathbb{R}^k$, we can attack the problem analytically, provided the log-likelihood function possesses continuous first and second derivatives. In such a case we know, from Chapter 4, that $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ is a turning point of LL if it satisfies the first-order conditions, i.e., if

$$\frac{\partial LL}{\partial \theta_i}(\widehat{\theta}_1, \ldots, \widehat{\theta}_k) = \frac{\partial LL}{\partial \theta_i}(\widehat{\boldsymbol{\theta}}) = 0 \quad \text{for} \quad i = 1, \ldots, k. \tag{16.4}$$

We note again that the solution for (16.4) may not be unique and it may not correspond to a maximum. However, in most of the standard cases the solution will be unique.

To investigate the nature of the turning point we compute the second derivative matrix $\mathbf{A}(\boldsymbol{\theta}) \in \mathbb{R}^{k \times k}$ of LL, whose entries are given by

$$\mathbf{A}(\boldsymbol{\theta})_{ij} = \frac{\partial^2 LL}{\partial \theta_i \theta_j}(\theta_1, \ldots, \theta_k) \quad \text{for} \quad (1 \leq i, j \leq k). \tag{16.5}$$

If we can then show that the matrix $\mathbf{A}(\widehat{\boldsymbol{\theta}})$ is negative definite, then we can conclude that the vector $\widehat{\boldsymbol{\theta}}$ is at least a local maximum of the log-likelihood function. In most of the cases that we shall encounter we will find that $\widehat{\boldsymbol{\theta}}$ is in fact a global maximizer and hence we can conclude that $\widehat{\theta}_i$ is the maximum likelihood estimator of the true parameter θ_i^* for $i = 1, \ldots, k$.

16.1 SAMPLE MEAN AND VARIANCE

A standard starting point of any statistical investigation is to set about deriving estimates for the mean and variance of the random variable under scrutiny. A natural way of doing this is to take the following steps:

- Run a statistical experiment to access a sample vector $(x_1, \ldots, x_n)^T \in \mathcal{X}$ of n observed values or outcomes.
- To estimate the mean we take a simple average of the values, i.e., we define

$$\widehat{\mu}_n = \sum_{j=1}^{n} \frac{1}{n} x_j. \tag{16.6}$$

- To estimate the variance we follow the same approach and take the simple average of the squares of the mean-adjusted observations, i.e., we define

$$\widehat{\sigma}_n^2 = \sum_{j=1}^{n} \frac{1}{n} \left(x_j - \widehat{\mu}_n\right)^2. \tag{16.7}$$

We know that every sample statistic is itself a random variable, (depending upon the sample) and so we reserve the tilde notation to indicate when we consider it in this way, i.e., we have

$$\widetilde{\mu}_n^2 = \sum_{j=1}^{n} \frac{1}{n} X_j \text{ random variable,}$$

$$\widetilde{\sigma}_n^2 = \sum_{j=1}^{n} \frac{1}{n} \left(X_j - \widetilde{\mu}_n\right)^2 \text{ random variable.} \tag{16.8}$$

The decision to use a simple average to define these two sample statistics is partly guided by our intuition (the simple average is a natural candidate), however, there is also an important theoretical argument which also supports this definition. This is captured in the following result:

Theorem 16.1. *Let $(x_1, \ldots, x_n)^T \in \mathcal{X}$ denote a sample vector whose values are known to be independently drawn from a normal distribution $N(\mu, \sigma^2)$. The maximum likelihood estimates of μ and σ^2 (based upon this sample) are precisely the simple averages (16.6) and (16.7) respectively.*

Proof. The likelihood function for the sample vector is defined as

$$L(\mu, \sigma^2) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x_j - \mu}{\sigma}\right)^2\right)$$

$$= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2}\sum_{j=1}^{n} \frac{(x_j - \mu)^2}{\sigma^2}\right).$$

In addition, the log-likelihood function is given by

$$LL(\mu, \sigma^2) = -\frac{n}{2}\left(\log(2\pi) + \log\sigma^2\right) - \frac{1}{2}\sum_{j=1}^{n}\frac{(x_j - \mu)^2}{\sigma^2}.$$

To find the maximum of this function we set up and solve the first-order conditions. Thus, differentiating this function with respect to μ and σ^2 yields

$$\begin{aligned}\frac{\partial}{\partial\mu}LL(\mu, \sigma^2) &= \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j - \mu)\\ &= \frac{1}{\sigma^2}\left(\sum_{j=1}^{n}x_j - n\mu\right)\end{aligned}$$

and

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}LL(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^{n}(x_j - \mu)^2\\ &= -\frac{1}{2\sigma^4}\left(n\sigma^2 - \sum_{j=1}^{n}(x_j - \mu)^2\right).\end{aligned}$$

Setting these derivatives to zero we find that the function has a turning point at the sample mean and sample variance, i.e.,

$$\frac{\partial}{\partial\mu}LL(\widehat{\mu}, \widehat{\sigma}^2) = 0 \Rightarrow \widehat{\mu} = \sum_{j=1}^{n}\frac{1}{n}x_j$$

and

$$\frac{\partial}{\partial\sigma^2}LL(\widehat{\mu}, \widehat{\sigma}^2) = 0 \Rightarrow \widehat{\sigma}^2 = \sum_{j=1}^{n}\frac{1}{n}(x_j - \widehat{\mu})^2.$$

To verify that the solution $(\widehat{\mu}, \widehat{\sigma}^2)$ is indeed a maximum of LL we check the definiteness of its 2×2 Hessian matrix. In this statistical framework the Hessian, the second derivative matrix of the log-likelihood function, is given by

$$\begin{aligned}A(\mu, \sigma^2) &= \begin{pmatrix}\dfrac{\partial LL^2}{\partial\mu^2} & \dfrac{\partial LL^2}{\partial\mu\partial\sigma^2}\\ \dfrac{\partial LL^2}{\partial\sigma^2\partial\mu} & \dfrac{\partial LL^2}{\partial\sigma^4}\end{pmatrix}\\ &= \begin{pmatrix}-\frac{n}{\sigma^2} & \dfrac{n}{\sigma^4}\left(n\mu - \sum_{j=1}^{n}x_j\right)\\ \frac{1}{\sigma^4}\left(n\mu - \sum_{j=1}^{n}x_j\right) & \frac{1}{2\sigma^6}\left(n\sigma^2 - 2\sum_{j=1}^{n}(x_j - \mu)^2\right)\end{pmatrix}\end{aligned}$$

$$= \begin{pmatrix} -\frac{n}{\sigma^2} & \frac{1}{\sigma^4}(\mu - \widehat{\mu}) \\ \frac{1}{\sigma^4}(\mu - \widehat{\mu}) & \frac{n}{2\sigma^6}\left(\sigma^2 - 2\widehat{\sigma}^2\right) \end{pmatrix}.$$

Setting $(\mu, \sigma^2) = (\widehat{\mu}, \widehat{\sigma}^2)$ we find that

$$A(\widehat{\mu}, \widehat{\sigma}^2) = \begin{pmatrix} -n/\widehat{\sigma}^2 & 0 \\ 0 & -n/2\widehat{\sigma}^4 \end{pmatrix}$$

which is negative definite, and so we can conclude that the sample estimates $\widehat{\mu}$ and $\widehat{\sigma}^2$ given by (16.6) and (16.7) are precisely the maximum likelihood estimates under the normal assumption. □

16.2 ON THE ACCURACY OF STATISTICAL ESTIMATORS

In order to assess the suitability of any estimation method we need a framework for measuring its accuracy. To achieve this we shall assume that our aim is to either estimate a single distributional parameter θ or, more generally, a real-valued function $f(\boldsymbol{\theta})$ of the vector of parameters $\boldsymbol{\theta}$. We will assume that $\widehat{f}(\mathbf{X})$ is some estimate (e.g., a maximum likelihood estimate MLE) of $f(\boldsymbol{\theta})$ that has been derived from a finite sample of observations, i.e., a realization of a random vector $\mathbf{X}$. We now provide two crucial error measurements:

1. The bias of an estimator $\widehat{f}$ to $f(\boldsymbol{\theta})$ is the function

$$\mathcal{B}(\boldsymbol{\theta}) = \mathbb{E}[\widehat{f}(\mathbf{X})|\boldsymbol{\theta}] - f(\boldsymbol{\theta}) \quad \boldsymbol{\theta} \in \Theta. \tag{16.9}$$

2. The mean square error (MSE) of an estimator $\widehat{f}$ is the function

$$\text{MSE}(\widehat{f}|\boldsymbol{\theta}) = \mathbb{E}\left[\left(\widehat{f}(\mathbf{X}) - f(\boldsymbol{\theta})\right)^2 |\boldsymbol{\theta}\right] \quad \boldsymbol{\theta} \in \Theta. \tag{16.10}$$

We observe that if $\mathcal{B}(\boldsymbol{\theta}) = 0$ then we can conclude that the expected value of the statistical estimate matches its true value and we say that $\widehat{f}$ is an unbiased estimate of $f(\boldsymbol{\theta})$. The mean square error provides us with a measure of how widely the statistical estimate varies from its true value, clearly the most successful estimates are those that are unbiased and have minimal MSE. We note that if $\widehat{f}$ is unbiased then its MSE coincides with its variance, in fact one can easily establish that

$$\text{MSE}(\widehat{f}|\boldsymbol{\theta}) = \text{variance}(\widehat{f}|\boldsymbol{\theta}) + \mathcal{B}(\boldsymbol{\theta})^2.$$

16.2.1 Sample mean example

The sample mean (16.6) is an unbiased estimate of μ, as the following calculation shows:

$$\mathcal{B}(\widetilde{\mu}_n) = \mathbb{E}[\widetilde{\mu}_n] - \mu = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[X_j] - \mu = \frac{1}{n}\underbrace{\sum_{j=1}^{n}\mu}_{=n\mu} - \mu = 0. \tag{16.11}$$

In addition, the corresponding variance of (16.6) is

$$\begin{aligned}\text{variance}(\widetilde{\mu}_n) &= \text{variance}\left(\sum_{j=1}^{n}\frac{1}{n}X_j\right)\\ &= \frac{1}{n^2}\sum_{j=1}^{n}\text{variance}\,(X_j) \quad \text{(by independence assumption)}\\ &= \frac{1}{n^2}\underbrace{\sum_{j=1}^{n}\sigma^2}_{=n\sigma^2} = \frac{\sigma^2}{n}.\end{aligned}$$

16.2.2 Sample variance example

The bias of the sample variance estimate (16.7) of σ^2 is given by

$$\mathcal{B}(\widetilde{\sigma}_n) = \mathbb{E}[\widetilde{\sigma}_n^2] - \sigma^2,$$

where

$$\mathbb{E}[\widetilde{\sigma}_n^2] = \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}[X_j^2] - \mathbb{E}[\widetilde{\mu}_n^2]. \tag{16.12}$$

Now, given that μ and σ^2 denote the true mean and variance of the underlying random variable, we can infer that

$$\sigma^2 = \mathbb{E}[X_j^2] - \mu^2 \Rightarrow \mathbb{E}[X_j^2] = \sigma^2 + \mu^2.$$

We have already shown that the mean and variance of the sample mean $\widetilde{\mu}_n$ are given by μ (the true mean) and σ^2/n respectively, thus

$$\text{variance}(\widetilde{\mu}_n) = \frac{\sigma^2}{n} = \mathbb{E}[\widetilde{\mu}_n^2] - \mu^2 \Rightarrow \mathbb{E}[\widetilde{\mu}_n^2] = \frac{\sigma^2}{n} + \mu^2.$$

Using these two observations in (16.12) we have

$$\mathbb{E}[\widetilde{\sigma}_n^2] = \sigma^2 + \mu^2 - \left(\frac{\sigma^2}{n} + \mu^2\right) = \sigma^2\left(1 - \frac{1}{n}\right), \tag{16.13}$$

and so

$$\mathcal{B}(\widetilde{\sigma}_n) = -\frac{\sigma^2}{n}.$$

Thus, we conclude that the sample variance given in (16.7) is biased as it tends to underestimate the true variance. However, the magnitude of the bias becomes smaller and smaller as more and more observations are included, in fact, in the limit as $n \to \infty$ we say that $\widetilde{\sigma}_n^2$ becomes asymptotically unbiased.

In order to have a consistently unbiased variance estimate we can alter the observation weight from $1/n$ to $1/(n-1)$ and we consider

$$\widehat{\sigma}_n^2 = \sum_{j=1}^{n} \frac{1}{n-1} \left(x_j - \widehat{\mu}\right)^2 \text{ observed value,}$$

$$\widetilde{\sigma}_n^2 = \sum_{j=1}^{n} \frac{1}{n-1} \left(X_j - \widetilde{\mu}\right)^2 \text{ random variable.} \tag{16.14}$$

It can be shown that the variance of the unbiased estimator is given by

$$\text{variance}(\widetilde{\sigma}_n^2) = \mathbb{E}\left[\left(\widetilde{\sigma}_n^2 - \sigma^2\right)^2\right] = \frac{1}{n}\left[m_4 - \frac{n-3}{n-1}\sigma^4\right], \tag{16.15}$$

where m_4 denotes the fourth central moment of the underlying random variable. Clearly it is desirable to employ statistical estimators that are unbiased and which also have small variance. In view of this a natural question to ask is:

How small can the variance of an unbiased estimator be?

We shall address this question by attempting to derive a lower bound for the variance. To begin with we will consider the case where our aim is to estimate a function of a single parameter θ, i.e., we let $\widehat{f}(\mathbf{X})$ denote an unbiased estimator of $f(\theta)$. We can write this function in the following way:

$$f(\theta) = \mathbb{E}[\widehat{f}(\mathbf{X})|\theta] = \int_{\mathcal{X}} \widehat{f}(\mathbf{x}) p(\mathbf{x}|\theta) d\mathbf{x}. \tag{16.16}$$

We note that since $p(\mathbf{x}|\theta)$ is a density function we have

$$\int_{\mathcal{X}} p(\mathbf{x}|\theta) d\mathbf{x} = 1 \Rightarrow \frac{\partial}{\partial\theta} \int_{\mathcal{X}} p(\mathbf{x}|\theta) d\mathbf{x} = 0.$$

We shall assume the partial derivative above can be taken inside the integral sign without any effect, i.e., we shall assume that

$$\frac{\partial}{\partial\theta} \int_{\mathcal{X}} p(\mathbf{x}|\theta) dx = \int_{\mathcal{X}} \frac{\partial p(\mathbf{x}|\theta)}{\partial\theta} d\mathbf{x} = 0. \tag{16.17}$$

Now differentiating f, given by (16.16), with respect to θ, and using the regularity condition (16.17), we can deduce that

$$\begin{aligned}\frac{\partial f}{\partial \theta} = f'(\theta) &= \int_{\mathcal{X}} \widehat{f}(\mathbf{x}) \frac{\partial p(\mathbf{x}|\theta)}{\partial \theta} d\mathbf{x} \\ &= \int_{\mathcal{X}} \widehat{f}(\mathbf{x}) \left(\frac{1}{p(\mathbf{x}|\theta)} \frac{\partial p(\mathbf{x}|\theta)}{\partial \theta} \right) p(\mathbf{x}|\theta) d\mathbf{x} \\ &= \int_{\mathcal{X}} \widehat{f}(\mathbf{x}) \frac{\partial \log(p(\mathbf{x}|\theta))}{\partial \theta} p(\mathbf{x}|\theta) d\mathbf{x}.\end{aligned}$$

Now using (16.17) once more we can also deduce that

$$\begin{aligned}0 = f(\theta) \int_{\mathcal{X}} \frac{\partial p(\mathbf{x}|\theta)}{\partial \theta} d\mathbf{x} &= \int_{\mathcal{X}} f(\theta) \frac{\partial p(\mathbf{x}|\theta)}{\partial \theta} d\mathbf{x} \\ &= \int_{\mathcal{X}} f(\theta) \frac{\partial \log(p(\mathbf{x}|\theta))}{\partial \theta} p(\mathbf{x}|\theta) d\mathbf{x}.\end{aligned}$$

This allows us to write

$$f'(\theta) = \int_{\mathcal{X}} (\widehat{f}(\mathbf{x}) - f(\theta)) \frac{\partial \log(p(\mathbf{x}|\theta))}{\partial \theta} p(\mathbf{x}|\theta) d\mathbf{x}.$$

We now apply the Cauchy–Schwarz inequality to deduce that

$$\begin{aligned}\left(\frac{\partial f}{\partial \theta} \right)^2 &= \left(\int_{\mathcal{X}} (\widehat{f}(\mathbf{x}) - f(\theta)) \frac{\partial \log(p(\mathbf{x}|\theta))}{\partial \theta} p(\mathbf{x}|\theta) d\mathbf{x} \right)^2 \\ &\leq \left(\int_{\mathcal{X}} \left(\widehat{f}(\mathbf{x}) - f(\theta) \right)^2 p(\mathbf{x}|\theta) d\mathbf{x} \right) \left(\int_{\mathcal{X}} \left(\frac{\partial \log(p(\mathbf{x}|\theta))}{\partial \theta} \right)^2 p(\mathbf{x}|\theta) d\mathbf{x} \right) \\ &= \text{variance}(\widehat{f}(\mathbf{X}|\theta)) \cdot \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log(p(\mathbf{x}|\theta)) \right)^2 \Big| \theta \right].\end{aligned}$$

Rearranging this inequality we find the famous Cramer–Rao inequality

$$\text{variance}(\hat{f}(\mathbf{X}|\theta)) \geq \frac{\left(f'(\theta) \right)^2}{I(\theta)}, \tag{16.18}$$

where

$$I(\theta) = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} LL(\theta) \right)^2 \Big| \theta \right] = \mathbb{E} \left[\left(\frac{\partial}{\partial \theta} \log(p(\mathbf{x}|\theta)) \right)^2 \Big| \theta \right].$$

The bounding quantity $I(\theta)$ is commonly called the Fisher information and, under mild regularity assumptions, it can be shown that

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} LL(\theta)\Big|\theta\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log(p(\mathbf{x}|\theta))\Big|\theta\right].$$

For the case $f(\theta) = \theta$, the bound states that

$$\text{variance}(\widetilde{\theta}|\theta) \geq \frac{1}{I(\theta)}.$$

This is often referred to as the minimum variance bound and any unbiased estimator which attains this lower bound is said to be the most efficient unbiased estimator of the true parameter θ.

The Cramer–Rao bound (16.18) can be generalized to a higher dimensional setting where we use random sample data to consider unbiased estimates $\widehat{\boldsymbol{\theta}}(\mathbf{X}) = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)^T$ of the true parameter vector $\boldsymbol{\theta}$. In this setting we define the $k \times k$ Fisher information matrix

$$\mathbf{I}_n(\boldsymbol{\theta}) = -\mathbb{E}[\mathbf{A}(\boldsymbol{\theta})|\boldsymbol{\theta}],$$

where $\mathbf{A}(\boldsymbol{\theta})$ is the second derivative matrix of the log-likelihood function for the observed sample data of size n, see (16.5); thus the information matrix is defined componentwise by

$$(\mathbf{I}_n(\boldsymbol{\theta}))_{ij} = -\mathbb{E}\left[\frac{\partial^2 LL}{\partial\theta_i\theta_j}(\theta_1, \ldots, \theta_k)\right] \quad \text{for} \quad (1 \leq i, j \leq k).$$

The higher-dimensional version of the Cramer–Rao bound tells us that

$$\text{variance}(\widetilde{\theta}_i|\boldsymbol{\theta}) \geq [\mathbf{I}_n(\boldsymbol{\theta})^{-1}]_{ii}, \quad \text{for} \quad i = 1, \ldots, k,$$

that is, the variance of any unbiased estimate for parameter i is bounded from below by the ith diagonal entry of the inverse of the Fisher information matrix.

16.3 THE APPEAL OF THE MAXIMUM LIKELIHOOD METHOD

In view of our theoretical development it is clear that the most desirable sample estimates are those which can be shown to be unbiased and most efficient i.e., when the Cramer–Rao lower bound is attained. Suppose we now turn our attention to the class of maximum likelihood estimates. Unfortunately, there exist counter-examples to show that these estimates are not necessarily unbiased and nor are they most efficient. For example, we have shown that the maximum likelihood estimate of variance, based upon a sample of independent observations from $N(\mu, \sigma^2)$, coincides with the simple average (16.7) and, as shown in the previous section, this is a biased estimator. In fact, the real appeal of the maximum likelihood approach is revealed when we allow the sample size to grow without bound, i.e., we let $n \to \infty$, where their properties become very nice indeed. In particular, if $\mathbf{X} = (X_1, \ldots, X_n)^T$ denotes a random vector of size n whose joint density function $p(\mathbf{x})$ depends upon a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Theta \subset \mathbb{R}^k$. If we let $\widehat{\boldsymbol{\theta}}(\mathbf{X}) = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ denote

the maximum likelihood estimate of these parameters, based on the sample vector $\mathbf{x}$ of realized values of $\mathbf{X}$, then more often than not (provided the density function fulfils certain regularity conditions) we can conclude that:

1. $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ is asymptotically unbiased, i.e.,

$$|\widehat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}| \to 0 \quad \text{as} \quad n \to \infty.$$

2. $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ is asymptotically efficient, i.e.,

$$\text{variance}\left((\widehat{\boldsymbol{\theta}}(\mathbf{X}))_i\right) \to \left(\mathbf{I}(\boldsymbol{\theta})^{-1}\right)_{ii} \quad \text{as} \quad n \to \infty, \quad \text{for} \quad i = 1, \ldots, k;$$

where

$$\mathbf{I}(\boldsymbol{\theta}) = \lim_{n\to\infty}\left(\frac{1}{n}\mathbf{I}_n(\boldsymbol{\theta})\right)$$

is the asymptotic information matrix, assumed to be finite and invertible.

3. $\widehat{\boldsymbol{\theta}}(\mathbf{X})$ is asymptotically normal, i.e.,

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{X}) - \boldsymbol{\theta}\right) \to \mathbf{Z} \sim N\left(\mathbf{0}, \mathbf{I}(\boldsymbol{\theta})^{-1}\right) \quad \text{as} \quad n \to \infty. \tag{16.19}$$

It is these appealing asymptotical properties that have made the maximum likelihood method so popular in practice. To take advantage of the result in practice one must be able to estimate the asymptotic information matrix $\mathbf{I}(\boldsymbol{\theta})$; this can be a significant challenge, however a common approach is to use the empirical second derivative matrix (provided its form is not too complicated), i.e., we would set

$$\left(\widehat{\mathbf{I}}(\boldsymbol{\theta})\right)_{ij} = -\frac{1}{n}\frac{\partial^2 LL}{\partial\theta_i\partial\theta_j}(\widehat{\theta}_1, \ldots, \widehat{\theta}_k) \quad 1 \le i, j \le k. \tag{16.20}$$

Thus, if our maximum likelihood estimates are based upon a large enough sample then it can be argued that they are almost unbiased/most efficient and approximately normally distributed. In other words, for practical purposes, the result is translated as

$$\text{for large } n \quad \widehat{\boldsymbol{\theta}}(\mathbf{X}) \approx \boldsymbol{\theta} + \mathbf{Z}_n \quad \text{where} \quad \mathbf{Z}_n \sim N\left(\mathbf{0}, \widehat{\mathbf{I}}(\boldsymbol{\theta})^{-1}\right).$$

Obviously we must be cautious when translating a sharp asymptotic result into an approximate real-world version. A result that can be shown, theoretically, to hold in the limit as $n \to \infty$ is of little use in practice if we only have access to a small sample.

Essential Mathematics for Market Risk Management
by Simon Hubbert

17 The Delta Method for Statistical Estimates

In the previous chapter we showed how statistical estimates arising from the maximum likelihood method enjoy some very appealing asymptotical properties. In this chapter we shall use the central limit theorem to develop the so-called delta method, a useful technique which can be used to deliver the asymptotic properties of a wide range of point estimates. We will illustrate the method by showing how it can be used to provide the asymptotic properties of the sample variance, the sample skewness and the sample kurtosis.

17.1 THEORETICAL FRAMEWORK

We kick-start our development by recalling that the multivariate version of the CLT considers a sequence of independent and identically distributed random vectors,

$$\mathbf{X}_k = (X_{1k}, \ldots, X_{dk})^T \in \mathbb{R}^d \text{ for } k = 1, 2, \ldots,$$

whose common mean vector and covariance matrix are denoted by

$$\mathbf{e} = \mathbb{E}[\mathbf{X}_k] \text{ and } \mathbf{V} = \mathbb{E}[(\mathbf{X}_k - \mathbf{e})(\mathbf{X}_k - \mathbf{e})^T]$$

respectively. The theorem then focuses upon the sequence of mean vectors

$$\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k \quad \text{for} \quad n = 1, 2, \ldots$$

and its conclusion reveals that the transformed sequence

$$\sqrt{n}(\overline{\mathbf{X}}_n - \mathbf{e}) \text{ converges to a random vector } \mathbf{Z} \sim N(\mathbf{0}, \mathbf{V}).$$

Suppose now that we have a function $f : \mathbb{R}^d \to \mathbb{R}$ of d variables and we are interested in the sequence $(f(\overline{\mathbf{X}}_n))_{n=1}^{\infty}$. A natural question to ask is:

Is it true that the transformed sequence $\sqrt{n}(f(\overline{\mathbf{X}}_n) - f(\mathbf{e}))$ *converges to a normal random variable?*

The answer to this question is positive provided the function f is differentiable at $\mathbf{e}$ and the following theorem (more commonly known as the delta method) captures this result.

Theorem 17.1. *Let $(\mathbf{X}_k)_{k=1}^{\infty}$ denote a sequence of independent and identically distributed d-dimensional random vectors, i.e.,*

$$\mathbf{X}_k = (X_{1k}, \dots X_{dk})^T \in \mathbb{R}^d \quad \text{for} \quad k = 1, 2, \dots$$

We let

$$\mathbf{e} = \mathbb{E}[\mathbf{X}_k] \quad \text{and} \quad \mathbf{V} = \mathbb{E}[(\mathbf{X}_k - \mathbf{e})(\mathbf{X}_k - \mathbf{e})^T], \quad k = 1, 2, \dots$$

denote the common mean vector and covariance matrix respectively. Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and let $\nabla f(\mathbf{e}) \in \mathbb{R}^d$ denote its gradient evaluated at $\mathbf{e}$, then

$$\sqrt{n}(f(\overline{\mathbf{X}}_n) - f(\mathbf{e})) \to Z \sim N(\mathbf{0}, a^2) \quad \text{as} \quad n \to \infty,$$

where

$$a^2 = \nabla f(\boldsymbol{e})^T \boldsymbol{V} \nabla f(\boldsymbol{e}). \tag{17.1}$$

A rigorous proof of this result is beyond the scope of this book. We can, however, gain an intuitive feel by arguing that, when n is very large, we can use the linear part of the Taylor expansion of f as a local approximation, i.e., we write

$$\sqrt{n}(f(\overline{\mathbf{X}}_n) - f(\mathbf{e})) \approx \nabla f(\mathbf{e}) \left[\sqrt{n}(\overline{\mathbf{X}}_n - \mathbf{e})\right].$$

Next, we argue that as $n \to \infty$ we have convergence, i.e.,

$$\lim_{n\to\infty} \sqrt{n}(f(\overline{\mathbf{X}}_n) - f(\mathbf{e})) = \nabla f(\mathbf{e}) \underbrace{\lim_{n\to\infty} \left[\sqrt{n}(\overline{\mathbf{X}}_n - \mathbf{e})\right]}_{\sim N(\mathbf{0}, \mathbf{V})}$$

and so we can conclude that $\sqrt{n}(f(\overline{\mathbf{X}}_n) - f(\mathbf{e}))$ is asymptotically normal $N(0, a^2)$, where the variance a^2 is given by (17.1).

In order to set the scene for the application of the delta method we shall assume that we have a sequence $(X_k)_{k\geq 1}$ of independent and identically distributed random variables. We then use this to define a sequence of d-dimensional random vectors by setting

$$\mathbf{X}_k = \left(X_k, X_k^2, \dots, X_k^d\right)^T \in \mathbb{R}^d.$$

We note that

$$\mathbf{e} = \mathbb{E}[\mathbf{X}_k] = (\mu_1, \mu_2, \dots, \mu_d)^T \in \mathbb{R}^d,$$

i.e., the mean vector of our sequence coincides with the d-dimensional moment vector of the underlying distribution. Furthermore, we let $\mathbf{V} \in \mathbb{R}^{d\times d}$ denote the covariance matrix of $\mathbf{X}_k$ whose ij^{th} entry is given by

$$\mathbf{V}_{ij} = \mathbb{E}\left[(X_k^i - \mu_i)(X_k^j - \mu_j)\right]$$

$$= \mathbb{E}\left[X_k^i X_k^j\right] - \mu_i \mu_j$$

$$= \mu_{i+j} - \mu_i \mu_j \quad (1 \le i, j \le d).$$

The multivariate CLT in this setting is concerned with the sequence

$$\overline{\mathbf{X}}_n = \frac{1}{n}\sum_{k=1}^{n} \mathbf{X}_k = \frac{1}{n}\begin{pmatrix} \sum_{k=1}^n X_k \\ \sum_{k=1}^n X_k^2 \\ \vdots \\ \sum_{k=1}^n X_k^d \end{pmatrix} = \begin{pmatrix} \overline{X} \\ \overline{X^2} \\ \vdots \\ \overline{X^d} \end{pmatrix}$$

and it tells us that

$$\sqrt{n}\left[\begin{pmatrix} \overline{X} \\ \overline{X^2} \\ \vdots \\ \overline{X^d} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_d \end{pmatrix}\right] \to \mathbf{Z} \sim N(\mathbf{0}, \mathbf{V}).$$

It turns out, as we shall see, that several important statistical estimators can be expressed as a function of the sample moments $\overline{X}, \overline{X^2}, \overline{X^3} \ldots$ In such cases we can employ the delta method to deduce their asymptotic behaviour. The rest of this chapter is devoted to examples of the delta method in action.

17.2 SAMPLE VARIANCE

For our first demonstration of the delta method we revisit the sample variance estimator (16.7). We observe that this can be written as

$$\widetilde{\sigma}_n^2(X) = \frac{1}{n}\sum_{j=1}^{n}\left(X_j - \overline{X}\right)^2$$

$$= \underbrace{\frac{1}{n}\sum_{j=1}^{n} X_j^2}_{=\overline{X^2}} - 2\overline{X}\underbrace{\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)}_{=\overline{X}} + \overline{X}^2$$

$$= \overline{X^2} - \overline{X},$$

i.e., we have that

$$\widetilde{\sigma}_n^2(X) = f(\overline{X}, \overline{X^2}) \text{ where } f(x, y) = y - x^2.$$

We note that f is differentiable and its gradient vector is given by

$$\nabla_{x,y} f(x, y) = \begin{pmatrix} -2x \\ 1 \end{pmatrix}.$$

If we now assume that the first four moments of the underlying distribution are well defined then, using the multivariate CLT, we can deduce that

$$\sqrt{n}\left[\begin{pmatrix} \overline{X} \\ \overline{X^2} \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}\right] \to \mathbf{Z} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mu_2 - \mu_1^2 & \mu_3 - \mu_1\mu_2 \\ \mu_3 - \mu_1\mu_2 & \mu_4 - \mu_2^2 \end{pmatrix}\right].$$

An application of the delta method now delivers the required asymptotic result:

$$\begin{aligned} \lim_{n\to\infty}\left[\sqrt{n}\left(\widetilde{\sigma}^2(X) - \sigma^2\right)\right] &= \lim_{n\to\infty}\left[\sqrt{n}\left(f(\overline{X}, \overline{X^2}) - f(\mu_1, \mu_2)\right)\right] \\ &= Z \sim N\left(0, a^2\right), \end{aligned}$$

where

$$\begin{aligned} a^2 &= (-2\mu_1, 0)\begin{pmatrix} \mu_2 - \mu_1^2 & \mu_3 - \mu_1\mu_2 \\ \mu_3 - \mu_1\mu_2 & \mu_4 - \mu_2^2 \end{pmatrix}\begin{pmatrix} -2\mu_1 \\ 1 \end{pmatrix} \\ &= 4[\mu_1^2(2\mu_2 - 1) - \mu_1\mu_3] + \mu_4 - \mu_2^2. \end{aligned}$$

It is possible to dramatically simplify this result. All we need is the simple observation that the sample variance estimator is shift-invariant, i.e., for any constant α we have that

$$\widetilde{\sigma}_n^2(X - \alpha) = \widetilde{\sigma}_n^2(X).$$

This follows from the fact that

$$\begin{aligned} X_j - \alpha - \overline{(X - \alpha)} &= X_j - \alpha - \frac{1}{n}\sum_{i=1}^{n}(X_i - \alpha) \\ &= X_j - \frac{1}{n}\sum_{i=1}^{n} X_i \quad (= X_j - \overline{X}). \end{aligned} \tag{17.2}$$

In view of this we can set $\alpha = \mu$, the mean of the distribution, and apply the delta method again. Firstly, the CLT tells us that

$$\sqrt{n}\left[\begin{pmatrix} \overline{X - \mu} \\ \overline{(X - \mu)^2} \end{pmatrix} - \begin{pmatrix} 0 \\ \sigma^2 \end{pmatrix}\right] \to \mathbf{Z} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & m_3 \\ m_3 & m_4 - \sigma^2 \end{pmatrix}\right],$$

where m_3 and m_4 denote the third and fourth central moments (11.2) of the distribution. Now, applying the delta method to mean-centred variables we find that

$$\lim_{n\to\infty}\left[\sqrt{n}\left(\widetilde{\sigma}_n^2(X)-\sigma^2\right)\right]=\lim_{n\to\infty}\left[\sqrt{n}\left(f\left(\begin{array}{c}\overline{X-\mu}\\ \overline{(X-\mu)^2}\end{array}\right)-f\left(\begin{array}{c}0\\ \sigma^2\end{array}\right)\right)\right]$$

$$=Z\sim N\left(0,m_4-\sigma^4\right).$$

We close this example by observing that if random variables are drawn from a normal distribution, i.e., each $X_j\sim N(\mu,\sigma^2)$, then we know from Chapter 11 that $m_4=3\sigma^4$ and so the above asymptotic result simplifies to

$$\lim_{n\to\infty}\left[\sqrt{n}\left(\widetilde{\sigma}_n^2(X)-\sigma^2\right)\right]=Z\sim N\left(0,2\sigma^4\right).$$

17.3 SAMPLE SKEWNESS AND KURTOSIS

For our second illustration of the delta method we investigate the asymptotic properties of the common sample estimates of the coefficients of skewness and kurtosis of the underlying distribution. As before we shall make the theoretical assumption that we have a sequence $(X_k)_{k\geq 1}$ of independent and identically distributed random variables. In practice we have access to a finite sample which we can use to build estimates. The skewness estimate, based upon a sample of size n, is given by

$$\mathcal{S}_n(X)=\frac{\frac{1}{n}\sum_{j=1}^n(X_j-\overline{X})^3}{\left(\frac{1}{n}\sum_{j=1}^n(X_j-\overline{X})^2\right)^{3/2}} \tag{17.3}$$

and, analogously, the kurtosis estimate is given by

$$\mathcal{K}_n(X)=\frac{\frac{1}{n}\sum_{j=1}^n(X_j-\overline{X})^4}{\left(\frac{1}{n}\sum_{j=1}^n(X_j-\overline{X})^2\right)^2}. \tag{17.4}$$

Expanding the numerator of (17.3) we find that

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^n\left(X_j-\overline{X}\right)^3&=\frac{1}{n}\sum_{j=1}^n\left(X_j^3-3\overline{X}X_j^2+3\overline{X}^2X_j-\overline{X}^3\right)\\&=\overline{X^3}-3\overline{X}\underbrace{\left(\frac{1}{n}\sum_{j=1}^nX_j^2\right)}_{\overline{X^2}}+3\overline{X}^2\underbrace{\left(\frac{1}{n}\sum_{j=1}^nX_j\right)}_{=\overline{X}}-\overline{X}^3\\&=\overline{X^3}-3\overline{X}\cdot\overline{X^2}+2\overline{X}^3.\end{aligned}$$

This observation allows us to write the sample skewness as a function of its first three sample moments. Specifically, we define

$$f_{\text{skew}}(x, y, z) = \frac{z - 3xy + 2x^3}{(y - x^2)^{3/2}}$$

and it follows that

$$\mathcal{S}_n(X) = f_{\text{skew}}\left(\overline{X}, \overline{X^2}, \overline{X^3}\right).$$

A similar but more lengthy calculation can be completed for the kurtosis coefficient. In this case we find that the function

$$f_{\text{kurt}}(s, t, u, v) = \frac{v^4 - 4su + 6s^2t - 3s^4}{(t - s^2)^2}$$

has the property that

$$\mathcal{K}_n(X) = f_{\text{kurt}}\left(\overline{X}, \overline{X^2}, \overline{X^3}, \overline{X^4}\right).$$

Before we employ the delta method we observe that (17.3) and (17.4) are both shift- and scale-invariant, i.e., for any $\alpha \in \mathbb{R}$ and $\beta > 0$ then

$$\mathcal{S}_n(X) = \mathcal{S}_n\left(\frac{X - \alpha}{\beta}\right) \quad \text{and} \quad \mathcal{K}_n(X) = \mathcal{K}_n\left(\frac{X - \alpha}{\beta}\right).$$

In view of this we can, theoretically, let $\alpha = \mu$ and $\beta = \sigma$ (the mean and standard deviation of the distribution respectively) and so consider the standardized distribution which, by definition, has zero mean and unit variance. Furthermore, the third and fourth moments of the standardized distribution coincide with $\mathcal{S}$ and $\mathcal{K}$, the skewness and kurtosis coefficients we are aiming to estimate.

17.3.1 Analysis of skewness

In order to investigate the asymptotic properties of the skewness estimator we need to assume that the first six moments of the distribution are finite. In this case the CLT tells us that

$$\sqrt{n}\left[\begin{pmatrix} \overline{(X - \mu)/\sigma} \\ \overline{((X - \mu)/\sigma)^2} \\ \overline{((X - \mu)/\sigma)^3} \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \\ \mathcal{S} \end{pmatrix}\right] \to Z,$$

where

$$Z \sim N\left[\begin{pmatrix}0\\0\\0\end{pmatrix}, \begin{pmatrix}1 & \mathcal{S} & \mathcal{K}\\ \mathcal{S} & \mathcal{K}-1 & \frac{m_5}{\sigma^5}-\mathcal{S}\\ \mathcal{K} & \frac{m_5}{\sigma^5}-\mathcal{S} & \frac{m_6}{\sigma^6}-\mathcal{S}^2\end{pmatrix}\right].$$

Given that the sample skewness is shift- and scale-invariant, we can employ the delta method to deduce that

$$\begin{aligned}
&\lim_{n\to\infty}\left[\sqrt{n}\left(\widetilde{\mathcal{S}}_n(X)-\mathcal{S}\right)\right]\\
&=\lim_{n\to\infty}\left[\sqrt{n}\left(f_{\text{skew}}\begin{pmatrix}\overline{(X-\mu)/\sigma}\\ \overline{((X-\mu)/\sigma)^2}\\ \overline{((X-\mu)/\sigma)^3}\end{pmatrix}-f_{\text{skew}}\begin{pmatrix}0\\1\\\mathcal{S}\end{pmatrix}\right)\right]\\
&=Z\sim N\left(0,a^2\right),
\end{aligned}$$

where

$$\begin{aligned}
a^2&=\nabla f_{\text{skew}}\begin{pmatrix}0\\1\\\mathcal{S}\end{pmatrix}^T\begin{pmatrix}1 & \mathcal{S} & \mathcal{K}\\ \mathcal{S} & \mathcal{K}-1 & \frac{m_5}{\sigma^5}-\mathcal{S}\\ \mathcal{K} & \frac{m_5}{\sigma^5}-\mathcal{S} & \frac{m_6}{\sigma^6}-\mathcal{S}^2\end{pmatrix}\nabla f_{\text{skew}}\begin{pmatrix}0\\1\\\mathcal{S}\end{pmatrix}\\
&=\begin{pmatrix}-3 & -3\mathcal{S}/2 & 1\end{pmatrix}\begin{pmatrix}1 & \mathcal{S} & \mathcal{K}\\ \mathcal{S} & \mathcal{K}-1 & \frac{m_5}{\sigma^5}-\mathcal{S}\\ \mathcal{K} & \frac{m_5}{\sigma^5}-\mathcal{S} & \frac{m_6}{\sigma^6}-\mathcal{S}^2\end{pmatrix}\begin{pmatrix}-3\\-3\mathcal{S}/2\\1\end{pmatrix} \qquad (17.5)\\
&=9-6\mathcal{K}+\frac{\mathcal{S}^2}{4}(35+9\mathcal{K})-3\mathcal{S}\frac{m_5}{\sigma^5}+\frac{m_6}{\sigma^6}.
\end{aligned}$$

17.3.2 Analysis of kurtosis

We can also apply the delta method to deliver the asymptotic properties of the sample kurtosis estimator (17.4). We require the additional stronger assumption that the first eight central moments of the underlying distribution are finite and then we follow the same procedure as for the skewness case. Firstly, we evoke the CLT to deduce that

$$\sqrt{n}\left[\begin{pmatrix}\overline{(X-\mu)/\sigma}\\ \overline{((X-\mu)/\sigma)^2}\\ \overline{((X-\mu)/\sigma)^3}\\ \overline{((X-\mu)/\sigma)^4}\end{pmatrix}-\begin{pmatrix}0\\1\\\mathcal{S}\\\mathcal{K}\end{pmatrix}\right]\to Z,$$

where

$$Z \sim N\left[\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \underbrace{\begin{pmatrix}1 & \mathcal{S} & \mathcal{K} & \frac{m_5}{\sigma^5}\\ \mathcal{S} & \mathcal{K}-1 & \frac{m_5}{\sigma^5}-\mathcal{S} & \frac{m_6}{\sigma^6}-\mathcal{K}\\ \mathcal{K} & \frac{m_5}{\sigma^5}-\mathcal{S} & \frac{m_6}{\sigma^6}-\mathcal{S}^2 & \frac{m_7}{\sigma^7}-\mathcal{K}\mathcal{S}\\ \frac{m_5}{\sigma^5} & \frac{m_6}{\sigma^6}-\mathcal{K} & \frac{m_7}{\sigma^7}-\mathcal{K}\mathcal{S} & \frac{m_8}{\sigma^8}-\mathcal{K}^2\end{pmatrix}}_{\mathbf{V}_{\text{kurt}}}\right].$$

The delta method can now be employed as before to reveal that

$$\begin{aligned}
&\lim_{n\to\infty}\left[\sqrt{n}\left(\widetilde{\mathcal{K}}_n(X)-\mathcal{K}\right)\right]\\
&=\lim_{n\to\infty}\left[\sqrt{n}\left(f_{\text{kurt}}\begin{pmatrix}\overline{(X-\mu)/\sigma}\\ \overline{((X-\mu)/\sigma)^2}\\ \overline{((X-\mu)/\sigma)^3}\\ \overline{((X-\mu)/\sigma)^4}\end{pmatrix}-f_{\text{kurt}}\begin{pmatrix}0\\1\\\mathcal{S}\\\mathcal{K}\end{pmatrix}\right)\right]\\
&=Z\sim N\left(0,a^2\right),
\end{aligned}$$

where

$$a^2=\nabla f_{\text{kurt}}(0,1,\mathcal{S},\mathcal{K})^T\mathbf{V}_{\text{kurt}}\nabla f_{\text{kurt}}(0,1,\mathcal{S},\mathcal{K}),$$

i.e., the quadratic form

$$\begin{pmatrix}-4\mathcal{S} & -2\mathcal{K} & 0 & 1\end{pmatrix}\begin{pmatrix}1 & \mathcal{S} & \mathcal{K} & \frac{m_5}{\sigma^5}\\ \mathcal{S} & \mathcal{K}-1 & \frac{m_5}{\sigma^5}-\mathcal{S} & \frac{m_6}{\sigma^6}-\mathcal{K}\\ \mathcal{K} & \frac{m_5}{\sigma^5}-\mathcal{S} & \frac{m_6}{\sigma^6}-\mathcal{S}^2 & \frac{m_7}{\sigma^7}-\mathcal{K}\mathcal{S}\\ \frac{m_5}{\sigma^5} & \frac{m_6}{\sigma^6}-\mathcal{K} & \frac{m_7}{\sigma^7}-\mathcal{K}\mathcal{S} & \frac{m_8}{\sigma^8}-\mathcal{K}^2\end{pmatrix}\begin{pmatrix}-4\mathcal{S}\\-2\mathcal{K}\\0\\1\end{pmatrix}$$

which, after a great deal of algebraic manipulation and simplification, we can show collapses to

$$a^2=16\mathcal{S}^2(1+\mathcal{K})-8\mathcal{S}\frac{m_5}{\sigma^5}+\mathcal{K}\left(\mathcal{K}(4\mathcal{K}-1)-4\frac{m_6}{\sigma^6}\right)+\frac{m_8}{\sigma^8}. \tag{17.6}$$

We close this section by observing that if the random variables are normally distributed then, in this special case, we know that the skewness $\mathcal{S}=0$, the kurtosis $\mathcal{K}=3$ and in

addition, $\mu_5 = 0 = \mu_7 = 0$, $\mu_6 = 15$ and $\mu_8 = 105$. Substituting these values into (17.5) and (17.6) we can deduce that, under the normal assumption, we have

$$\sqrt{n}(\widetilde{\mathcal{S}}_n(X) - 0) \to Z_{\text{skew}} \sim N(0, 6) \tag{17.7}$$

and

$$\sqrt{n}(\widetilde{\mathcal{K}}_n(X) - 3) \to Z_{\text{kurt}} \sim N(0, 24) \text{ as } n \to \infty. \tag{17.8}$$

Essential Mathematics for Market Risk Management
by Simon Hubbert

18 Hypothesis Testing

A careful statistical analysis of observational data enables a skilled investigator to form a certain opinion regarding one (and very often more than one) of the distributional properties of the underlying random process. The investigator clearly hopes that, given the statistical evidence, his/her concluding opinion is correct. In this chapter our aim is to develop scientific tests which assess the validity of a proposed statistical hypothesis. The scope for this area is huge and it is hoped that this chapter will serve as a helpful introduction to this branch of statistics. The reader who wishes to discover more is encouraged to consult Cox and Hinkley's excellent textbook (1979).

18.1 THE TESTING FRAMEWORK

Let us assume that a statistical investigation has taken place and that, based upon the findings, the investigator proposes a certain hypothesis. We think of the hypothesis as a prediction of a particular distributional property that the underlying random outcome may exhibit. The purpose of this section is to outline an approach which can be used to test the validity of a given hypothesis. We begin by fixing some familiar assumptions:

- We shall consider an n-dimensional random vector $\mathbf{X} = (X_1, \ldots, X_n)^T$ whose components are the random variables representing the potential outcomes of the process over n consecutive time points.
- We assume that the joint density of the random vector $\mathbf{X}$ depends upon k parameters which we store in the vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$. As in Chapter 16, we let $\Theta \subset \mathbb{R}^k$ denote the range of all possible values that the parameter vector $\boldsymbol{\theta}$ can take.
- We let $\mathbf{x} = (x_1, \ldots, x_n)^T$ denote the vector whose components represent the actual realized values of the random vector $\mathbf{X}$. As in Chapter 16, we let $\mathcal{X} \subset \mathbb{R}^n$ denote the range of all possible outcomes of the random vector $\mathbf{X}$.

We recall, from our development of the maximum likelihood method, that, under these assumptions, the full family of possible joint density functions for the vector $\mathbf{X}$ is given by

$$\{p(\mathbf{x}|\boldsymbol{\theta}) \ (\mathbf{x} \in \mathcal{X}) : \boldsymbol{\theta} \in \Theta\}.$$

18.1.1 The null and alternative hypotheses

In general, most, if not all, of the parameters which define the joint density function of $\mathbf{X}$ will be unknown. In view of this the scientific investigator makes an assertion, or a

hypothesis, regarding the true value of the parameter vector $\boldsymbol{\theta}$. In fact, two competing hypotheses are proposed:

1. The null hypothesis.
   We choose a subset Θ_0 of the parameter space Θ and propose that $\boldsymbol{\theta}$ belongs to Θ_0; we write

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \subset \Theta.$$

2. The alternative hypothesis.
   Here we choose another subset $\Theta_1 \subset \Theta$, disjoint from Θ_0, and propose the alternative hypothesis that $\boldsymbol{\theta}$ belongs to Θ_1. We write

$$H_1 : \boldsymbol{\theta} \in \Theta_1 \subset \Theta, \quad \text{where } \Theta_1 \cap \Theta_0 \text{ is empty.}$$

   We remark that the alternative hypothesis is a direct competitor of the null hypothesis since there is no intersection between Θ_1 and Θ_0.

18.1.2 Hypotheses: simple vs compound

If a hypothesis is based upon a single vector of parameters then we say that it is a simple one. For example,

$$H_0 : \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}, \text{ i.e. , } H_0 : \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix} = \begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \\ \vdots \\ \widehat{\theta}_k \end{pmatrix}$$

represents a simple null hypothesis and, in this case, $\Theta_0 = \{\widehat{\boldsymbol{\theta}}\}$. A test that is not simple is said to be compound. An alternative hypothesis can also be simple, i.e., we can have

$$H_1 : \boldsymbol{\theta} = \widetilde{\boldsymbol{\theta}}, \text{ i.e., } H_1 : \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix} = \begin{pmatrix} \widetilde{\theta}_1 \\ \widetilde{\theta}_2 \\ \vdots \\ \widetilde{\theta}_k \end{pmatrix}$$

provided that $\Theta_1 = \{\widetilde{\boldsymbol{\theta}}\} \neq \Theta_0 = \{\widehat{\boldsymbol{\theta}}\}$.

18.1.3 The acceptance and rejection regions

Given the two competing hypotheses H_0 and H_1 we now have to make a decision; we can either accept the null hypothesis H_0 or reject it in favour of the alternative H_1. This decision is made by analysing the sample vector $\mathbf{x}$. The idea is as follows:

- Partition the sample space $\mathcal{X}$ into two distinct regions $\mathcal{A}$ and $\mathcal{R}$ such that

$$\mathcal{X} = \mathcal{A} \cup \mathcal{R} \quad \text{and} \quad \mathcal{A} \cap \mathcal{R} \text{ is empty.}$$

- We call $\mathcal{A}$ the acceptance region and $\mathcal{R} = \mathcal{X} \setminus \mathcal{A}$ the rejection region and we obey the following decision rule:

$$\text{accept } H_0 \text{ if } \mathbf{x} \in \mathcal{A}$$

and

$$\text{reject } H_0 \text{ if } \mathbf{x} \in \mathcal{R} \quad (\text{i.e., if } \mathbf{x} \notin \mathcal{A}).$$

18.1.4 Potential errors

One must take care when defining this partition of the sample space. Clearly we must avoid making errors of judgement, of which there are two possible types.

1. A type I error.
   In this case we find that we have rejected the null hypothesis when, in fact, it was true. We define the probability of committing a type I error as

$$\alpha = \mathbb{P}[\text{reject } H_0 \mid H_0 \text{ is true}] = \mathbb{P}[\mathbf{x} \notin \mathcal{A} \mid H_0 \text{ is true}]. \tag{18.1}$$

   This probability is commonly called the significance level of the test and its value is defined by the size of the acceptance region.
   The courtroom analogy of a type I error is to find the accused guilty when, in fact, they are innocent; here the null hypothesis is the presumption of innocence.
2. A type II error.
   In this case we find that we have accepted the null hypothesis when, in fact, it was false. We define the probability of committing a type II error as

$$\beta = \mathbb{P}[\text{accept } H_0 | H_0 \text{ is false}] = \mathbb{P}[\mathbf{x} \in \mathcal{A} | H_0 \text{ is false}]. \tag{18.2}$$

   The courtroom analogy of a type II error is to find the accused innocent when, in fact, they are guilty.

In practice, it is common to focus on the complimentary quantity $\widetilde{p} = 1 - \beta$, which is called the power of the test since it measures the probability that a type II error will not be committed. Mathematically, we write

$$\begin{aligned} \widetilde{p} = 1 - \beta &= \mathbb{P}[\text{reject } H_0 | H_0 \text{ is false}] \\ &= \mathbb{P}[\mathbf{x} \notin \mathcal{A} | H_0 \text{ is false}]. \end{aligned} \tag{18.3}$$

18.1.5 Controlling the testing errors/defining the acceptance region

In view of the potential for error, the statistical investigator should design a testing procedure in such a way that the associated error probabilities (18.1) and (18.2) are kept small. Unfortunately, it can be shown that these two quantities are inversely related, i.e., reducing the significance level α has the effect of increasing β (and vice versa). In view of this it is usual to proceed as follows:

1. Fix a significance level $\alpha = \widehat{\alpha}$, the probability of rejecting a true null hypothesis. Common choices are $\alpha = 0.01$ or 0.05.
2. Choose the acceptance region $\mathcal{A}$ so as to maximize the power $\widetilde{p}$ of the test while ensuring the significance is fixed at $\alpha = \widehat{\alpha}$.

The test that arises from the above procedure is said to be the most powerful test for the significance level $\widehat{\alpha}$.

18.2 TESTING SIMPLE HYPOTHESES

In the special case where both the null and alternative hypotheses are simple, we have the following theory to guide us.

Step 1. Assuming that the null and alternative hypotheses have the form

$$H_0 : \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix} = \begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \\ \vdots \\ \widehat{\theta}_k \end{pmatrix} \text{ and } H_1 : \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix} = \begin{pmatrix} \widetilde{\theta}_1 \\ \widetilde{\theta}_2 \\ \vdots \\ \widetilde{\theta}_k \end{pmatrix}, \tag{18.4}$$

we define the so-called likelihood ratio to be

$$\lambda(\mathbf{x}) = \frac{p(\mathbf{x}|\widehat{\boldsymbol{\theta}})}{p(\mathbf{x}|\widetilde{\boldsymbol{\theta}})} = \frac{p(\mathbf{x}|(\widehat{\theta}_1, \ldots, \widehat{\theta}_k))}{p(\mathbf{x}|(\widetilde{\theta}_1, \ldots, \widetilde{\theta}_k))}. \tag{18.5}$$

We note that since the likelihood ratio $\lambda(\mathbf{x})$ depends upon the observed sample vector $\mathbf{x}$, it can also be interpreted as a random variable in its own right.

Step 2. We now evoke an extremely important statistical result, the famous Neyman–Pearson lemma:

Lemma 18.1. *For a given sample vector $\boldsymbol{x}$, let $\lambda(\boldsymbol{x})$ (18.5) denote the likelihood ratio for a pair of simple null and alternative hypotheses H_0 and H_1* (18.4) *respectively. Let $\alpha \in [0, 1]$ be a given significance level and define the constant c to be the solution to*

$$\mathbb{P}[\lambda(\boldsymbol{x}) < c] = \alpha.$$

Then, the following statistical test:

$$\text{accept } H_0 \text{ if } \boldsymbol{x} \in \mathcal{A} = \{\boldsymbol{x} \in \mathcal{X} : \lambda(\boldsymbol{x}) \geq c\}$$

is the most powerful (has the smallest type II error) of all tests with significance level α.

A simple consequence of this result is that the rejection region for the most powerful test is thus defined to be

$$\mathcal{R} = \mathcal{X} \setminus \mathcal{A} = \{\mathbf{x} \in \mathcal{X} : \lambda(\mathbf{x}) < c\}.$$

18.2.1 Testing the mean when the variance is known

In order to see the Neyman–Pearson lemma in action, we consider the following set-up:

- The sequence $(X_t)_{t\in\mathbb{Z}}$ is an independent and identically distributed process. Furthermore, it is known that the process is normally distributed with known variance σ^2 but with unknown mean μ.
- The statistical investigator proposes the following hypotheses:

$$H_0 : \mu = \mu_0 \quad \text{and} \quad H_1 : \mu = \mu_1 > \mu_0.$$

- n-realized values of the process are observed and stored in the vector $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$.

In this framework the likelihood ratio is given by

$$\begin{aligned}
\lambda &= \frac{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_0)^2\right)}{\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu_1)^2\right)} \\
&= \exp\left(\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i-\mu_1)^2 - (x_i-\mu_0)^2\right]\right) \\
&= \exp\left(\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (2x_i(\mu_0-\mu_1) + (\mu_1^2-\mu_0^2))\right]\right) \\
&= \exp\left(\frac{1}{2\sigma^2}\left(2n\widehat{\mu}(\mathbf{x})(\mu_0-\mu_1) + n(\mu_1^2-\mu_0^2)\right)\right) \\
&= \exp\left(\frac{n(\mu_0-\mu_1)}{\sigma^2}\left(\widehat{\mu}(\mathbf{x}) - \frac{\mu_1+\mu_0}{2}\right)\right).
\end{aligned}$$

We note, in the penultimate line of the above development, we have used the formula for the sample mean (16.6). We recall that the alternative hypothesis assumes that $\mu_1 > \mu_0$ and thus, we can define a positive quantity δ such that $\mu_1 = \mu_0 + \delta$, and consequently we can express the likelihood ratio as

$$\lambda(\mathbf{x}) = \exp\left(-\frac{n\delta}{\sigma^2}\left(\widehat{\mu}(\mathbf{x}) - \mu_0 - \frac{\delta}{2}\right)\right).$$

In line with the Neyman–Pearson lemma we treat the sample mean of $\mathbf{x}$ as a random variable, which we denote as $\widetilde{\mu}(\mathbf{x})$, and we consider an acceptance region for the test to be of the form

$$\mathcal{A} = \left\{\mathbf{x} \in \mathcal{X} : \exp\left(-\frac{n\delta}{\sigma^2}\left[\widetilde{\mu}(\mathbf{x}) - \mu_0 - \frac{\delta}{2}\right]\right) \geq c\right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{X} : \exp\left(-\frac{\sqrt{n}\delta}{\sigma} \left[\frac{\widetilde{\mu}(\mathbf{x}) - \mu_0}{\sigma/\sqrt{n}} - \frac{\delta}{2\sigma/\sqrt{n}} \right] \right) \geq c \right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{X} : -\frac{\sqrt{n}\delta}{\sigma} \left(\frac{\widetilde{\mu}(\mathbf{x}) - \mu_0}{\sigma/\sqrt{n}} - \frac{\delta}{2\sigma/\sqrt{n}} \right) \geq \log(c) \right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{X} : \frac{\widetilde{\mu}(\mathbf{x}) - \mu_0}{\sigma/\sqrt{n}} \leq -\frac{\sigma}{\sqrt{n}\delta} \log(c) + \frac{\delta}{2\sigma/\sqrt{n}} \right\}$$

$$= \left\{ \mathbf{x} \in \mathcal{X} : \frac{\widetilde{\mu}(\mathbf{x}) - \mu_0}{\sigma/\sqrt{n}} \leq \kappa \right\},$$

where

$$\kappa = -\frac{\sigma}{\sqrt{n}\delta} \log(c) + \frac{\delta}{2\sigma/\sqrt{n}}.$$

We conclude, from the above analysis, that the definition of $\mathcal{A}$ can be written in terms of the value of the random quantity

$$T(\mathbf{X}) = \frac{\widetilde{\mu}(\mathbf{x}) - \mu_0}{\sigma/\sqrt{n}}.$$

In view of this, we can transfer attention from the seemingly complex form of the likelihood ratio and, instead, focus on the single random variable $T(\mathbf{X})$. This discovery is only useful if we know how $T(\mathbf{X})$ is distributed under the null hypothesis and, in our case, we know that $T(\mathbf{X}) \sim N(0, 1)$. This information allows us to complete the test as follows:

- Select an appropriate significance level α for the test. Then, using the above development, we notice the Neyman–Pearson requirement that we find a constant c such that

  $$\mathbb{P}[\lambda(\mathbf{x}) < c] = \alpha$$

  is equivalent to choosing a constant κ such that

  $$\mathbb{P}[T(\mathbf{X}) \geq \kappa] = \alpha. \tag{18.6}$$

  The fact that $T(\mathbf{X}) \sim N(0, 1)$ allows us to immediately deduce that $\kappa = \Phi^{-1}(1 - \alpha)$, where $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal distribution function. We note here that a distinction is made between $T(\mathbf{X})$ the random variable and $T(\mathbf{x})$ the actual observed value of the test statistic.
- We then say

  $$\text{accept } H_0 \text{ if } T(\mathbf{x}) \leq \Phi^{-1}(1 - \alpha)$$

  or equivalently,

  $$\text{if } \widehat{\mu}(\mathbf{x}) \leq \mu_0 + \frac{\sigma \Phi^{-1}(1 - \alpha)}{\sqrt{n}}. \tag{18.7}$$

  Otherwise, we reject H_0 in favour of H_1.

We remark that if the alternative hypothesis proposed that

$$H_1 : \mu_1 < \mu_0 \text{ rather than } H_1 : \mu_1 > \mu_0$$

then the same analysis as above would yield

$$\text{accept } H_0 \text{ if } T(\mathbf{x}) \geq -\Phi^{-1}(1-\alpha)$$

or equivalently,

$$\text{if } \widehat{\mu}(\mathbf{x}) \geq \mu_0 - \frac{\sigma\,\Phi^{-1}(1-\alpha)}{\sqrt{n}}. \tag{18.8}$$

The tests that we have considered here, namely

$$H_0 : \mu = \mu_0 \text{ and either } H_1 : \mu_1 > \mu_0 \text{ or } H_1 : \mu_1 < \mu_0,$$

are examples of one-sided tests since, in both cases, the alternative hypothesis is designed to clearly indicate the direction of deviation from the null hypothesis, i.e., either from above or from below.

The two-sided version of this test is given by

$$H_0 : \mu = \mu_0 \text{ and } H_1 : \mu = \mu_1 \neq \mu_0.$$

To evaluate this we can use the same test statistic as above and it can be shown that, for a significance level α, the decision to accept/reject the null hypothesis is taken on the following basis:

$$\text{accept } H_0 \text{ if } -\Phi^{-1}(1-\alpha/2) \leq T(\mathbf{x}) \leq \Phi^{-1}(1-\alpha/2).$$

In other words, accept H_0 if the value of the sample mean satisfies

$$\mu_0 - \frac{\sigma\,\Phi^{-1}(1-\alpha/2)}{\sqrt{n}} \leq \widehat{\mu}(\mathbf{x}) \leq \mu_0 + \frac{\sigma\,\Phi^{-1}(1-\alpha/2)}{\sqrt{n}}. \tag{18.9}$$

Otherwise, we reject H_0 in favour of H_1.

18.3 THE TEST STATISTIC

The previous example is illuminating because it introduces us to the notion of a test statistic. There are two ways of viewing a test statistic:

1. As a random variable.
   In this case we shall write $T(\mathbf{X}) = T(X_1, \ldots, X_n)$ to indicate that T is considered as a function of the unknown outcomes of the random vector $\mathbf{X}$.

2. As a realized value.
   If we let $\mathbf{x} = (x_1, \ldots, x_n)^T$ denote the vector of observed values of the random vector $\mathbf{X}$, then we shall write $T(\mathbf{x})$ to denote the corresponding value of the test statistic.

Let us assume that we are to test

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ against } H_1 : \boldsymbol{\theta} \in \Theta_1.$$

In order to use a test statistic to evaluate the validity of H_0, we take the following path:

(i) Let $\mathcal{X}$ denote the sample space of the observation vector $\mathbf{x}$ and propose a decision rule of the form

$$\text{accept } H_0 \text{ if } \mathbf{x} \in \mathcal{A} = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \leq \kappa\}, \tag{18.10}$$

where κ is an arbitrary constant.

(ii) Let us denote the distribution of $T(\mathbf{X})$ under the null hypothesis as

$$F(x) = \mathbb{P}[T(\mathbf{X}) \leq x | H_0 \text{ is true}].$$

We shall assume that we either know F completely or that we have a valid approximation.

(iii) For a fixed significance level α we use our knowledge of F to solve

$$1 - F(\kappa_\alpha) = \mathbb{P}[T(\mathbf{X}) > \kappa_\alpha | H_0 \text{ is true}] = \alpha,$$
$$\text{i.e.,} \quad \kappa_\alpha = F^{-1}(1 - \alpha). \tag{18.11}$$

(iv) We then fix the definitive test for significance level α by setting $\kappa = \kappa_\alpha$ in (18.10); i.e., the test is defined as

$$\text{accept } H_0 \text{ if } \mathbf{x} \in \mathcal{A} = \{\mathbf{x} \in \mathcal{X} : T(\mathbf{x}) \leq \kappa_\alpha\}, \tag{18.12}$$

where κ is given by (18.11).

We illustrate a direct application of the test statistic with the following example.

18.3.1 Example: Testing the mean when the variance is unknown

In this example we assume that the sequence $(X_t)_{t \in \mathbb{Z}}$ is an independent and normally distributed process for which the process mean and variance μ and σ^2 are both unknown. The statistical investigator sets out to test

$$H_0 : \mu = \mu_0 \text{ against } H_1 : \mu = \mu_1 > \mu_0.$$

We recall that in the previous example (where the variance σ^2 was known) we were able to decide whether or not to accept H_0 by analysing the test statistic

$$T(\mathbf{X}) = \frac{\widetilde{\mu}(\mathbf{x}) - \mu_0}{\sqrt{\sigma^2/n}}.$$

We choose to follow this approach, however we replace the known variance with its estimated value, i.e., we consider

$$T(\mathbf{X}) = \frac{\widetilde{\mu}(\mathbf{x}) - \mu_0}{\sqrt{\widetilde{\sigma}^2(\mathbf{x})/n}}, \tag{18.13}$$

where

$$\widetilde{\sigma}^2(\mathbf{x}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widetilde{\mu}(\mathbf{x}))^2\,.$$

To reveal the distributional behaviour for $T(\mathbf{X})$ we notice that we can write it as

$$T(\mathbf{X}) = \frac{\widetilde{\mu}(\mathbf{x}) - \mu_0}{\sqrt{\widetilde{\sigma}^2(\mathbf{x})/n}} = \frac{Z}{\sqrt{Y}},$$

where

$$Z = \frac{\widetilde{\mu}(\mathbf{x}) - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \;\text{ and }\; Y = \frac{\widetilde{\sigma}^2(\mathbf{x})}{\sigma^2} \sim \frac{1}{n-1}\chi^2_{n-1}.$$

In view of this discovery we can appeal to Definition 12.8 to conclude that the proposed test statistic has the student t-distribution with $n-1$ degrees of freedom. We can then follow the standard process to evaluate the validity of the null hypothesis:

- For a given significance level α, we can use statistical tables for the t-distribution (or otherwise) to find the value of κ_α that satisfies

$$t_{n-1}(\kappa_\alpha) = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\pi}}\int_{\infty}^{\kappa_\alpha}\left(1 - \frac{t^2}{n-1}\right)^{-n/2} dt = 1-\alpha,$$

$$\text{i.e., } \kappa_\alpha = t^{-1}_{n-1}(1-\alpha).$$

- Applying (18.12) we accept the null hypothesis if

$$\widehat{\mu}(\mathbf{x}) \le \mu_0 + t^{-1}_{n-1}(1-\alpha)\sqrt{\widehat{\sigma}^2(\mathbf{x})/n}.$$

We close this example by adding that the same statistic (18.13) can be used to evaluate the two-sided version of this test,

$$H_0 : \mu = \mu_0 \;\text{ and }\; H_1 : \mu = \mu_1 \neq \mu_0.$$

It can be shown, by exploiting the symmetry of the t-distribution, that the acceptance region for this test is given by

$$\mu_0 - t^{-1}_{n-1}\left(1 - \frac{\alpha}{2}\right)\frac{\widehat{\sigma}(\mathbf{x})}{\sqrt{n}} \le \widehat{\mu}(\mathbf{x}) \le \mu_0 + t^{-1}_{n-1}\left(1 - \frac{\alpha}{2}\right)\frac{\widehat{\sigma}(\mathbf{x})}{\sqrt{n}}. \tag{18.14}$$

18.3.2 The p-value of a test statistic

The success of our statistical testing procedure hinges upon the fact that the distribution of the test statistic $T(\mathbf{X})$ is known, either completely or approximately. This allows us to compute probabilities, based on $T(\mathbf{X})$, under the assumption that the null hypothesis is true. For instance, we can write

$$\mathbb{P}[T(\mathbf{X}) \geq t | H_0 \text{ is true}] = 1 - F(t).$$

The test itself corresponds to a certain confidence level $\alpha \in [0, 1]$. Once this is given, the user computes a cutoff level κ_α that satisfies

$$1 - F(\kappa_\alpha) = \alpha.$$

Then, armed with these numbers, the test statistic $T(\mathbf{x})$ is constructed (using a vector $\mathbf{x}$ of sample data) and the investigator decides

$$\text{at confidence level } \alpha \text{ the null hypothesis is rejected if } T(\mathbf{x}) \geq \kappa_\alpha.$$

The process can be summarized as:

$$\text{define } \alpha \rightarrow \text{ find } \kappa_\alpha \rightarrow \text{ check } T(\mathbf{x}) \geq \kappa_\alpha. \tag{18.15}$$

We note that as α varies from 0 to 1 then κ_α drops from its largest possible value (where $F(x) = 1$) to its smallest possible value (where $F(x) = 0$). In other words, the rejection region grows as α increases from zero to one. Suppose that, as a result of a statistical test (18.15), the investigator concludes that the null hypothesis must be rejected at confidence level α. A natural question to ask in this case is:

How small can we make α without altering the conclusion to reject the null hypothesis?

To answer this we consider the likelihood of the random variable $T(\mathbf{X})$ exceeding what we have already observed, namely $T(\mathbf{x})$. This likelihood is called the p-value associated with $T(\mathbf{x})$ and is defined by

$$p = \mathbb{P}[T(\mathbf{X}) \geq T(\mathbf{x}) | H_0 \text{ is true}] = 1 - F(T(\mathbf{x})). \tag{18.16}$$

The p-value can be viewed as a measure of the weight of evidence against the null hypothesis. If the p-value is small then we can make one of the following conclusions:

1. We have strong evidence for rejecting the null hypothesis as the chance of exceeding the original observation is slim.
2. We have been unlucky and experienced an extreme value of the test statistic even though the null hypothesis is actually true.

In view of this we see that there is no absolute proof that a null hypothesis is true or false. A small p-value however can, as indicated, be used as helpful evidence to support a decision to reject.

18.4 TESTING COMPOUND HYPOTHESES

So far in this chapter we have demonstrated, with the help of the Neyman–Pearson result, how the most powerful test for a pair of simple competing hypotheses can be constructed using the likelihood ratio defined by (18.5). We now turn to the more realistic case where both the null and alternative hypotheses are compound, i.e., we are to test

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \quad \text{against} \quad H_0 : \boldsymbol{\theta} \in \Theta_1.$$

It turns out that there is a useful recipe for constructing tests which can be viewed as a generalization of the simple Neyman–Pearson-inspired case. Before we introduce this approach we recall some facts:

- We assume the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$ belongs to the full k-dimensional parameter space $\Theta \subset \mathbb{R}^k$.
- The null hypothesis is a statement that imposes a restriction on r $(1 \leq r \leq k)$ of these parameters and, in view of this, the restricted range $\Theta_0 \subset \Theta$ is a $(k - r)$-dimensional subspace of Θ.

For a fixed observation vector $\mathbf{x}$ the general form of the likelihood function is given by

$$L(\boldsymbol{\theta}|\mathbf{x}) = p(\mathbf{x}|\theta)(\boldsymbol{\theta} \in \Theta).$$

We perform two maximization procedures:

1. Maximize over the full parameter space Θ.
   Here we solve
   $$\text{maximize } L(\boldsymbol{\theta}|\mathbf{x}) \text{ subject to } \boldsymbol{\theta} \in \Theta,$$
   and we denote the solution by $L^*_{\text{full}}(\mathbf{x})$.
2. Maximize over the restricted parameter space Θ_0.
   Here we solve
   $$\text{maximize } L(\boldsymbol{\theta}|\mathbf{x}) \text{ subject to } \boldsymbol{\theta} \in \Theta_0,$$
   and we denote the solution by $L^*_{\text{res}}(\mathbf{x})$.

We then define the generalized likelihood ratio $\lambda(\mathbf{x})$ by

$$\lambda(\mathbf{x}) = \frac{L^*_{\text{res}}(\mathbf{x})}{L^*_{\text{full}}(\mathbf{x})} = \frac{\text{maximum}\{L(\boldsymbol{\theta}|\mathbf{x}) : \boldsymbol{\theta} \in \Theta_0\}}{\text{maximum}\{L(\boldsymbol{\theta}|\mathbf{x}) : \boldsymbol{\theta} \in \Theta\}}. \tag{18.17}$$

As an initial idea we could propose to use $\lambda(\mathbf{x})$ as a test statistic. The drawback of this idea is that, in general, the distribution of $\lambda(\mathbf{X})$ is not known. However, all is not lost because we have the following result:

Proposition 18.2. *Under certain regularity conditions the statistic* $-2\log(\lambda(\boldsymbol{X}))$ *converges to a chi-squared distribution with* r *degrees of freedom as the sample size* $n \to \infty$*, i.e.,*

$$-2\log(\lambda(\boldsymbol{X})) = -2\log(\lambda(X_1, \ldots, X_n)) \to Z \sim \chi^2_r \quad \text{as} \quad n \to \infty.$$

Proof. We provide a brief sketch of the proof of the asymptotic distribution.

Consider the following:

$$\boldsymbol{\theta}_{\text{full}} = \text{ the maximizer of } L(\boldsymbol{\theta}|\mathbf{x}) \text{ subject to } \boldsymbol{\theta} \in \Theta$$

and

$$\boldsymbol{\theta}_{\text{res}} = \text{the maximizer of } L(\boldsymbol{\theta}|\mathbf{x}) \text{ subject to } \boldsymbol{\theta} \in \Theta_0.$$

A Taylor expansion of $LL(\boldsymbol{\theta}_{\text{res}}) = \log(L(\boldsymbol{\theta}_{\text{res}}))$ about $\boldsymbol{\theta}_{\text{full}}$ yields

$$LL(\boldsymbol{\theta}_{\text{res}}) \approx LL(\boldsymbol{\theta}_{\text{full}}) + (\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}_{\text{res}}) \underbrace{\nabla LL(\boldsymbol{\theta}_{\text{full}})}_{=0 \text{ by definition}} + \frac{1}{2}(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}_{\text{res}})^T \nabla^2 LL(\boldsymbol{\theta}_{\text{full}})(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}_{\text{res}}).$$

In other words, we can write

$$\begin{aligned} -2(LL(\boldsymbol{\theta}_{\text{full}}) - LL(\boldsymbol{\theta}_{\text{res}})) &= 2\log(\lambda(\mathbf{X})) \\ &\approx (\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}_{\text{res}})^T \nabla^2 LL(\boldsymbol{\theta}_{\text{full}})(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}_{\text{res}}) \\ &\approx -n(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}_{\text{res}})^T \mathbf{I}(\boldsymbol{\theta})(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}_{\text{res}}). \end{aligned}$$

We note that the final line of the above development follows by the fact that the information matrix can be estimated by the negative average of the second derivative matrix, see (16.20). We now appeal to the asymptotic behaviour of maximum likelihood estimates we developed in Chapter 16. Specifically, since $\boldsymbol{\theta}_{\text{full}}$ is the maximum likelihood estimate of the true parameter vector $\boldsymbol{\theta}$, we can appeal to (16.19) to deduce that

$$\sqrt{n}(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}) \to Z \sim N(0, \mathbf{I}(\boldsymbol{\theta})^{-1}) \quad \text{as} \quad n \to \infty.$$

This implies that

$$n(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta})^T \mathbf{I}(\boldsymbol{\theta})(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}) \to Y_k \sim \chi^2_k \quad \text{as} \quad n \to \infty.$$

Now, if our null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_{\text{res}}$ imposes r restrictions on the parameter vector, then we can infer that

$$\underbrace{n(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}_{\text{res}})^T \mathbf{I}(\boldsymbol{\theta})(\boldsymbol{\theta}_{\text{full}} - \boldsymbol{\theta}_{\text{res}})}_{\approx -2\log(\lambda(\mathbf{X}))} \to Y_r \sim \chi^2_r \quad \text{as} \quad n \to \infty.$$

□

In view of the above proposition we can say that, for large enough n, the generalized log-likelihood ratio $-2\log(\lambda(\mathbf{X}))$ is, under the null hypothesis H_0, approximately distributed as a chi-squared random variable with r degrees of freedom. We use this fact to evaluate the validity of H_0 by following the usual procedure, i.e.,

- For a significance level α let y_α denote the $100 \cdot \alpha\%$ quantile of the χ^2_r distribution. For a large enough sample we can use this value to conclude that

$$\mathbb{P}[-2\log(\lambda(\mathbf{X})) \leq z_\alpha] \approx \alpha.$$

- We then follow the decision rule

$$\text{accept } H_0 \text{ if } \mathbf{x} \in \mathcal{A} = \{\mathbf{x} \in \mathcal{X} : -2\log(\lambda(\mathbf{X})) \leq z_\alpha\}. \tag{18.18}$$

Essential Mathematics for Market Risk Management
by Simon Hubbert

19

Statistical Properties of Financial Losses

Perhaps the most challenging of all problems in mathematical finance is that of capturing the distribution function of the underlying financial variable. In risk management the random variable that is of most importance is the loss rate (or the change in value) of a portfolio of financial assets. So far in this book we have made the bold assumption that this random variable is normally distributed. Indeed, as we have discovered in Chapter 10, the normal framework is very alluring as it delivers closed-form solutions to many of the key problems of financial risk management. In this chapter our aim is to focus on the following fundamental question:

What are the true distributional properties of daily financial losses?

The first step towards tackling this question is to gather financial data and proceed as follows:

- At time t (today) we select a candidate financial asset, e.g., a stock or market index, and collect together its price history spanning the past $n+1$ days. We store this data in a sample vector

$$\mathbf{S}_t = (S_{t-1}, \ldots, S_{t-(n+1)})^T \in \mathbb{R}^{n+1}.$$

- Using $\mathbf{S}_t$ we compute the n-dimensional vector of the asset's daily log losses, which we denote by

$$\mathbf{l}_t = (l_{t-1}, \ldots, l_{t-n})^T \in \mathbb{R}^n,$$

where

$$l_{t-\tau} = -\log\left(\frac{S_{t-\tau}}{S_{t-(\tau+1)}}\right) \quad \tau = 1, \ldots, n.$$

Armed with sample data we will now investigate by following the three stages outlined below.

Stage 1: Analysis of sample statistics

As a straightforward first step we can calculate the following crucial sample statistics:

$$\text{sample mean} \quad \widehat{\mu} = \frac{1}{n}\sum_{\tau=1}^{n} l_{t-\tau};$$

$$\text{sample variance} \quad \widehat{\sigma} = \frac{1}{n-1}\sum_{\tau=1}^{n}(l_{t-\tau} - \widehat{\mu})^2;$$

$$\text{sample skewness} \quad \widehat{S} = \frac{\frac{1}{n}\sum_{\tau=1}^{n}(l_{t-\tau} - \widehat{\mu})^3}{\left(\frac{1}{n}\sum_{\tau=1}^{n}(l_{t-\tau} - \widehat{\mu})^2\right)^{3/2}};$$

$$\text{sample kurtosis} \quad \widehat{\mathcal{K}} = \frac{\frac{1}{n}\sum_{\tau=1}^{n}(l_{t-\tau} - \widehat{\mu})^4}{\left(\frac{1}{n}\sum_{\tau=1}^{n}(l_{t-\tau} - \widehat{\mu})^2\right)^{2}}. \tag{19.1}$$

We note that if the process for l_t were normal (or approximately so) then we would expect the values for the sample skewness and kurtosis to be close to 0 and 3 respectively.

Stage 2: An application of the Jarque–Bera test

At this stage of the investigation the aim is to develop a tailor-made approach whose goal is to test directly whether or not the daily log losses are normal distributed. The inspiration for the test stems from two results we have already established, namely (17.7) and (17.8). We recall that these results state that if $\widetilde{S}_n$ and $\widetilde{\mathcal{K}}_n$ denote the skewness and kurtosis estimates (based on a sample of size n) of a random process that is normally distributed, then the following asymptotic results hold:

$$\frac{\sqrt{n}\widetilde{S}_n}{\sqrt{6}} \quad \text{and} \quad \frac{\sqrt{n}(\widetilde{K}_n - 3)}{\sqrt{24}} \quad \text{are both} \quad \sim N(0,1) \quad \text{as} \quad n \to \infty. \tag{19.2}$$

In 1980, Anil Bera and Carlos Jarque (two applied economists) simply noted that if n is sufficiently large then the random variables (19.2) have, approximately, the standard normal distribution. Furthermore, since these variables are also independent then Definition 12.3 can be evoked to conclude that, for large n, the variable

$$\widetilde{JB} = n\left(\frac{\widetilde{S}_n^2}{6} + \frac{(\widetilde{K}_n - 3)^2}{24}\right) \tag{19.3}$$

is approximately a χ^2 random variable with two degrees of freedom. The random variable $\widetilde{JB}$ is commonly called the Jarque–Bera test statistic for normality Jarque and Bera (1980). We can construct the test by specifying the null hypothesis

$$H_0: \quad \text{the daily log loss process is normally distributed.}$$

Now, if we let $F_{\chi^2(2)}$ denote the distribution of a χ^2 random variable with 2 degrees of freedom then we can deduce that

$$F_{\chi^2(2)}(x) = \mathbb{P}[\widetilde{JB} \leq x | H_0 \text{ is true}].$$

For the 5% confidence level we follow (18.11) and compute the critical value

$$\kappa_{0.05} = F^{-1}_{\chi^2(2)}(0.95) = 5.99.$$

We can then use our sample data to provide an estimate $\widehat{JB}$ of the Jarque–Bera test statistic and, following (18.12), we act as follows:

$$\text{accept the null hypothesis} \quad H_0 \quad \text{if} \quad \widehat{JB} \leq 5.99. \tag{19.4}$$

We remark that if the result of the test is to reject the null hypothesis then we can also compute the corresponding p-value defined by

$$p_{JB} = 1 - F_{\chi^2(2)}(\widehat{JB}).$$

A small value for p_{JB} can be considered as further evidence for the rejection of the null hypothesis.

Stage 3. Visualization

Here we aim to produce a variety of plots in order to gain insight into the properties of daily losses. We will consider four distinct approaches.

- An empirical plot.
  Here we let the data speak for themselves and plot the empirical density function based on the sample. We can then examine whether financial losses are approximately normal by comparing the empirical plot with the appropriate normal density function.
- A quantile–quantile (Q–Q) plot.
  Here we proceed as follows:
  – Rewrite the sample losses in ascending order,

$$l^{(1)} \leq l^{(2)} \leq \cdots \leq l^{(n)}.$$

  – According to the empirical distribution

$$l^{(k)} = F_n^{-1}\left(\frac{k}{n}\right) \quad k = 1, \ldots, n.$$

  – If the data are normally distributed with mean μ and variance σ^2 then we would expect that

$$\Phi^{-1}\left(\frac{k}{n}\right) \approx \frac{l^{(k)} - \mu}{\sigma} \quad \Rightarrow \quad l^{(k)} \approx \sigma\,\Phi^{-1}\left(\frac{k}{n}\right) + \mu,$$

  i.e., if the data are normally distributed then there is an approximate linear relationship between $l^{(k)}$ and $\Phi^{-1}(k/n)$. The Q–Q plot is thus simply a plot of $\Phi^{-1}(k/n)$ against $l^{(k)}$; if a straight line emerges then this is evidence that the process is indeed normally distributed.
- A plot of the auto-correlation function.
  All serious investors are aware of the rule that *past performance is not an indicator of future return.* In order to investigate this principle scientifically, we can monitor the sample auto-correlation function

$$\text{sample auto-correlation} \quad \widehat{\rho}(k) = \frac{\frac{1}{n}\sum_{\tau=1}^{n-k}(l_{t-\tau} - \widehat{\mu})(l_{t-\tau-k} - \widehat{\mu})}{\widehat{\sigma}^2}$$

  over a range of time lags, say for $k = 1, 2, \ldots, n/4$. If the principle is accurate then we would expect that the plot will reveal a function which drops very quickly to zero.

- A volatility plot.
  We have assumed that the volatility of our loss random variable remains constant over time. To investigate this we select a candidate financial asset and gather a long history of its past realized losses covering the period $[t - T_{\text{dist}}, t - 1]$. We then fix a past reference date $T > T_{\text{dist}}$ and consider the following experiments:
  - Unconditional volatility.
    Here we estimate the volatility of the asset at time $t - \tau$ (for $\tau = 1, \ldots, T$) using all of the available data, i.e., using all realized values dating back to time T_{dist} so that

$$\widehat{\sigma}_{\text{run}}(t - \tau) = \sqrt{\frac{1}{T_{\text{dist}} - \tau - 1} \sum_{k=1}^{T_{\text{dist}}-\tau} (l_{t-\tau-k} - \widehat{\mu})^2}. \tag{19.5}$$

    We call this the running volatility as the sample window increases by one as each day passes. The running volatility is taken as a measure of the unconditional volatility of the process.
  - Conditional volatility.
    Here we estimate volatility using data from the recent past. Specifically, we fix an appropriate number n and estimate the volatility of the asset at $t - \tau$ based on the realized values from $t - \tau - 1$ to $t - \tau - n$, so that

$$\widehat{\sigma}_{\text{mov}}(t - \tau) = \sqrt{\frac{1}{n - 1} \sum_{k=1}^{n} (l_{t-\tau-k} - \widehat{\mu})^2}. \tag{19.6}$$

    We call this the moving volatility as the sample window moves along as each day passes (the final observation drops out and is replaced by the most recent one). The moving volatility is taken as a measure of the conditional volatility of the process.

The aim of our testing procedure is to shed light upon the statistical properties that appear to characterise the daily loss process of a typical financial asset. For our investigation we will examine realized daily log losses from January 2003 to January 2010 for the FTSE-100 index and Vodafone stock; this represents a total of 1830 observations.

19.1 ANALYSIS OF SAMPLE STATISTICS

We begin the testing process by displaying a plot of the historical prices and daily log losses for both the FTSE-100 index and Vodafone stock. In addition we use these data sets to compute the crucial sample statistics (19.1).

Examples: FTSE-100 and Vodafone

Plots of the daily closing prices and daily log losses for the FTSE-100 index and for Vodafone stock are displayed in Figures 19.1 and 19.2 respectively.

Using the log loss data for the FTSE-100 index, we can compute the following sample statistics:

$$\begin{aligned} &\widehat{\mu}_{\text{ftse}} = -0.00015, \quad \widehat{\sigma}_{\text{ftse}} = 0.0130, \\ &\widehat{S}_{\text{ftse}} = 0.096, \quad \widehat{\mathcal{K}}_{\text{ftse}} = 9.06. \end{aligned} \tag{19.7}$$

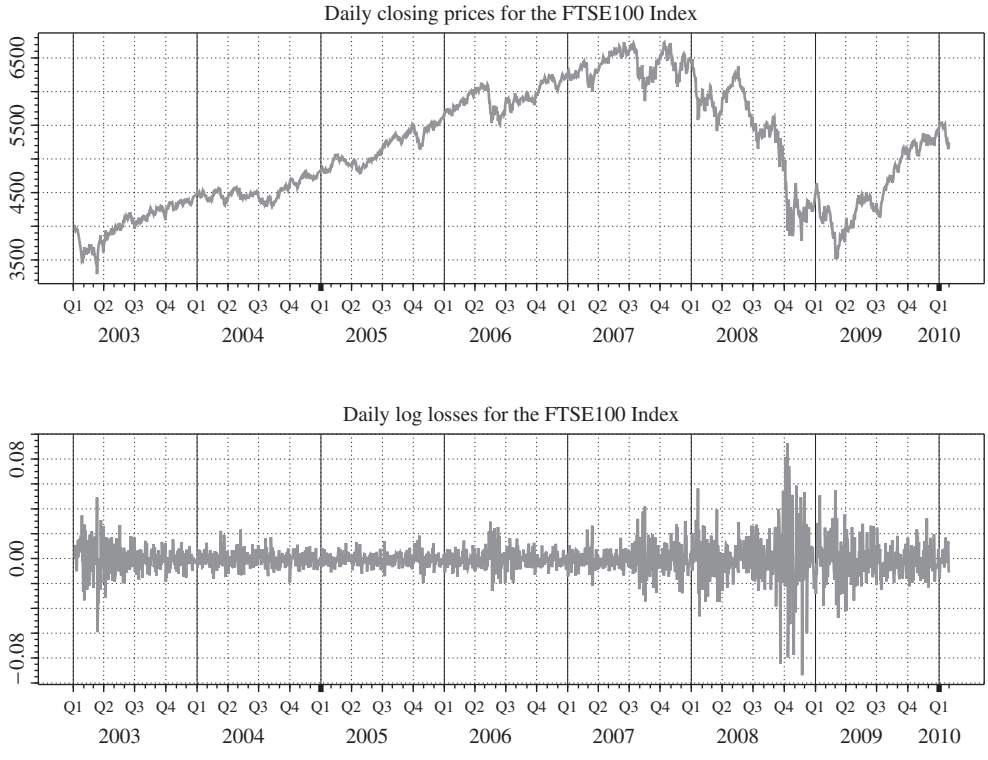


Figure 19.1 A plot of the daily closing prices and daily log losses of the FTSE-100 index.

Analogously, for Vodafone stock we have

$$\widehat{\mu}_{\text{voda}} = -0.00003, \quad \widehat{\sigma}_{\text{voda}} = 0.0184,$$
$$\widehat{S}_{\text{voda}} = 0.304, \quad \widehat{\mathcal{K}}_{\text{voda}} = 5.97. \tag{19.8}$$

Given that the above statistics are calculated on a sample of size 1830 we can readily compute the Jarque–Bera statistic (19.3) for each stock, specifically we have

$$\widehat{JB}_{\text{ftse}} = 1830 \cdot \left(\frac{0.096^2}{6} + \frac{6.06^2}{24} \right) = 2800;$$
$$\widehat{JB}_{\text{vod}} = 1830 \cdot \left(\frac{0.304^2}{6} + \frac{2.97^2}{24} \right) = 700. \tag{19.9}$$

Observations

Based upon the values of the sample statistics we make the following observations:

- The mean of the daily log losses is very small. Indeed, in comparison to the size of the volatility it is almost negligible.

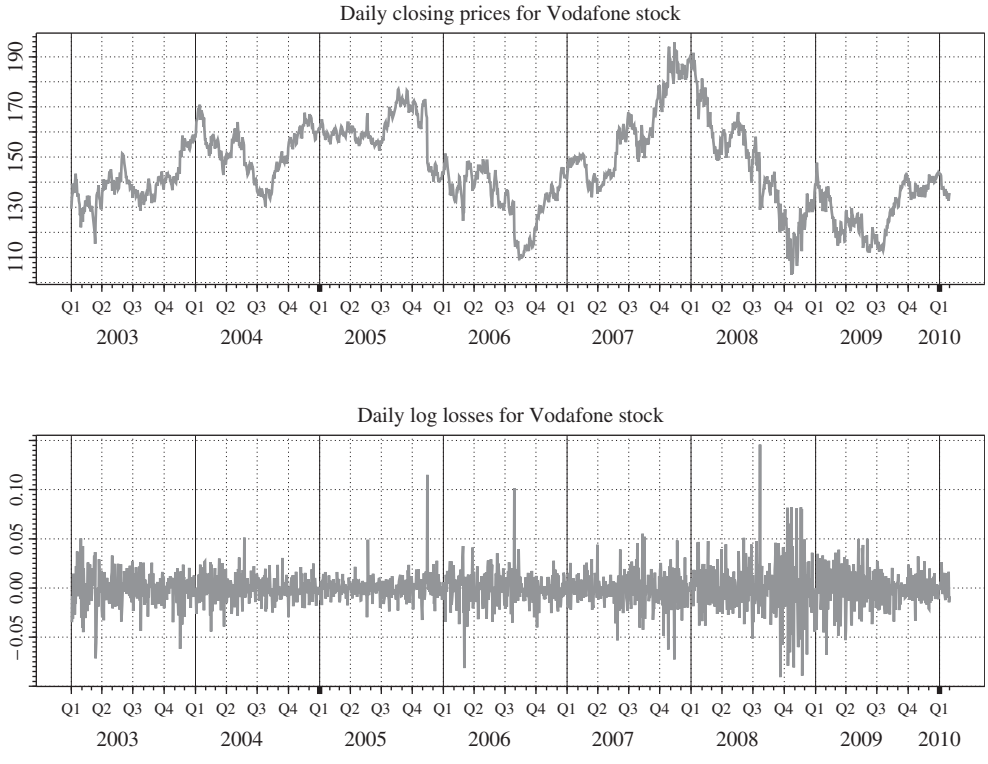


Figure 19.2 A plot of the daily closing prices and daily log losses for Vodafone stock.

- The skewness of each of our examples is positive, which indicates that losses occur more frequently than profits.
- In both our examples the kurtosis coefficient is greater than three. This indicates that extreme values of loss and profit are more likely to occur than a normally distributed model would predict.
- The Jarque–Bera statistics (19.9) for both data sets are high and indicate that we reject the hypothesis that the data are normally distributed. This decision is strengthened by zero p-values for both cases.

In addition we notice that the plots of the log losses both appear to display the following pattern:

- There are tranquil periods where the losses do not vary too much away from the mean.
- There are also well-defined periods of turbulence where the losses vary much more wildly about the mean.

This phenomenon was first observed in the early 1960s by Benoit Mandelbrot (1963), the mathematician who is now most famous for his work on fractal geometry. This observation, together with the values of the sample statistics, leads us to pose the following conjecture:

Conjecture 19.1. *The daily log loss process for a typical financial asset tends to exhibit the following properties:*

1. *Its distribution is not normal. In particular, it has fatter tails and is skewed to the left.*
2. *The loss process displays little serial correlation.*
3. *The square of the loss process (taken as a measure of the variance of the process) displays a significant element of serial correlation.*

19.2 THE EMPIRICAL DENSITY AND Q–Q PLOTS

In this section we shall provide further evidence for statement 1 of Conjecture 19.1. Specifically, we shall use historical data for each of our candidate assets to perform two tasks:

- Firstly, we construct the empirical density function for the daily log loss random variable of both the FTSE-100 index and Vodafone stock. In each case we superimpose the normal model that is implied from the sample mean and volatility estimates. In order to examine this more closely we also zoom in on the behaviour of the tail of the distribution where large losses can potentially occur. The resulting plots are displayed in Figure 19.3 (for the FTSE-100 index) and Figure 19.4 (for Vodafone stock).
- Secondly, we examine the normal assumption in more detail by providing the Q–Q plots for both the FTSE-100 index and Vodafone stock. The resulting plots are displayed in Figure 19.5.

Observations

A glance at the plots we have constructed enables us to make the following observations:

- According to the empirical density plots it appears that a normal distribution is not a good fit to financial loss data. Indeed, the real data exhibit much fatter tails, especially on the right-hand side, which corresponds to extreme loss events.
- The Q–Q plots emphasise this even further. We can see from Figure 19.5 that the Q–Q plots deviate from a straight line form at the extremes.
- Both of these observations provide yet more evidence to support statement 1 of Conjecture 19.1.

19.3 THE AUTO-CORRELATION FUNCTION

In this section we turn to statements 2 and 3 of Conjecture 19.1 and, again, we aim to provide further evidence to validate them. To do this we simply plot the auto-correlation function for both the past losses and also for the square of the losses. The results for the FTSE-100 index are displayed in Figure 19.6 and the results for Vodafone in Figure 19.7.

Observations

In both cases we can make the following observations:

- The auto-correlation function of the pure losses decays towards zero extremely quickly, within a few time lags. This supports statement 2 of Conjecture 19.1.
- On the other hand, the auto-correlation function of the squared losses takes much longer to decay, around 50–100 time lags. This supports statement 3 of Conjecture 19.1 and implies that volatility ought to be modelled as a time-dependent function.

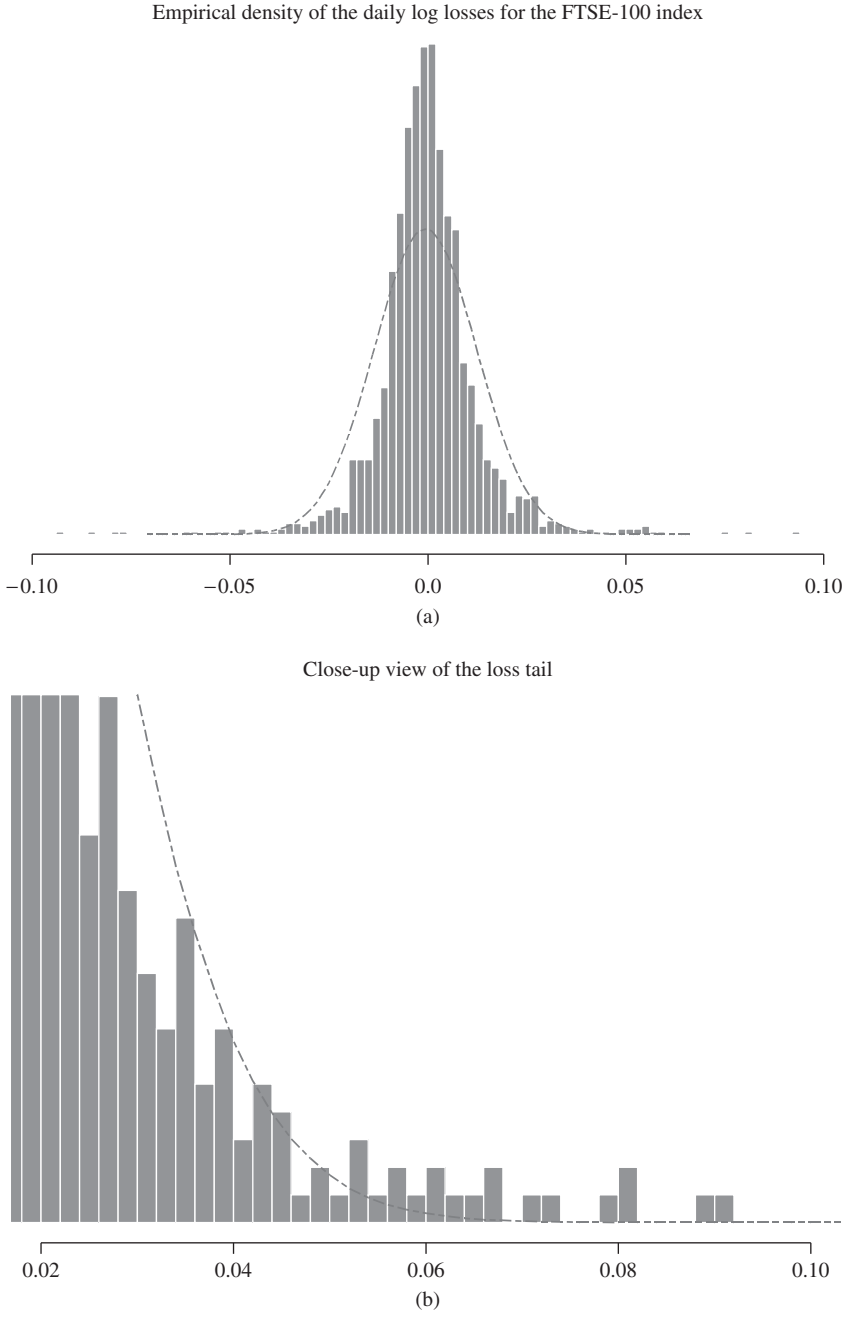


Figure 19.3 Empirical density of daily log losses for the FTSE-100 index. (a) Shape of the empirical density function for the FTSE-100 index. (b) A close-up of the extreme loss tail.

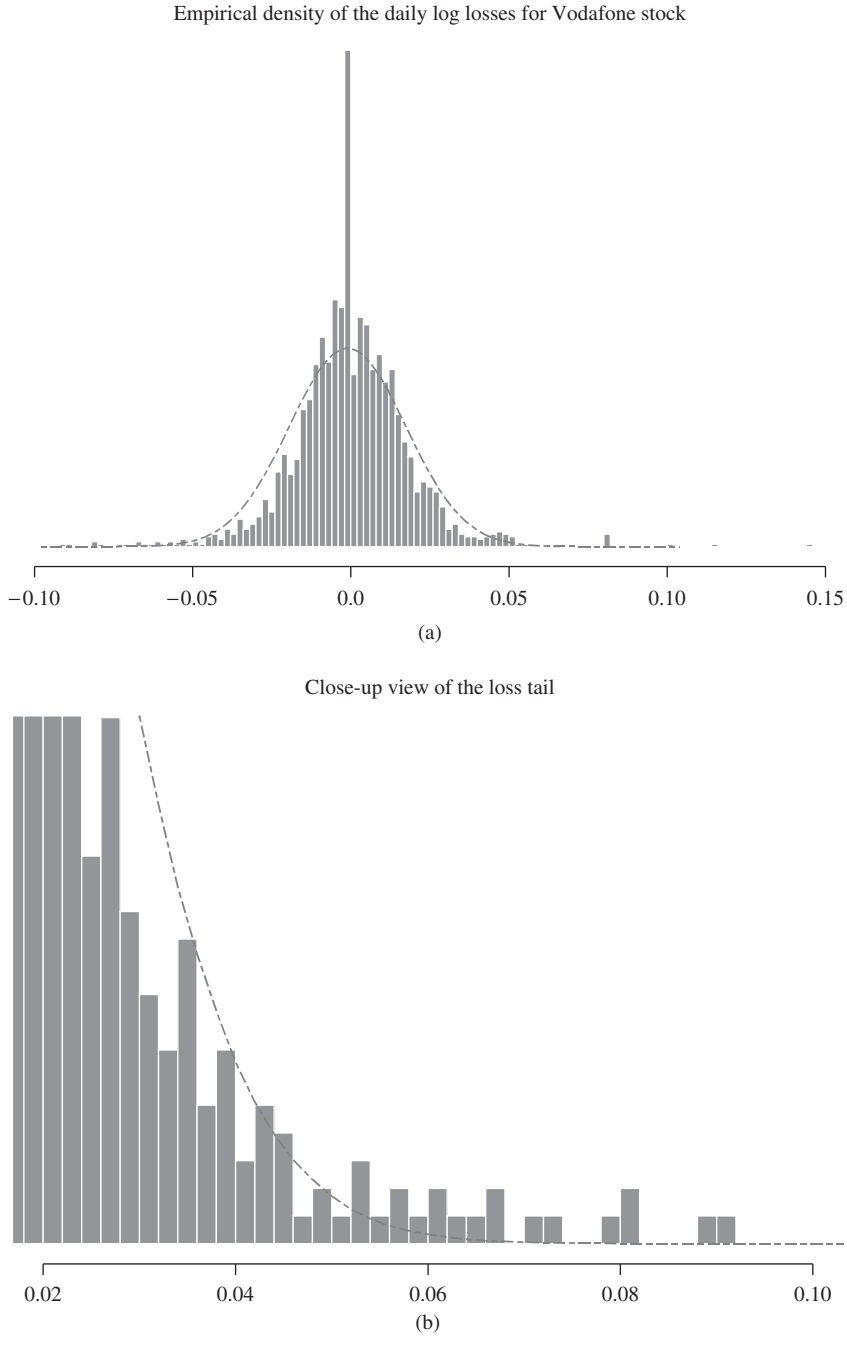


Figure 19.4 Empirical density of daily log losses for Vodafone stock. (a) Shape of the empirical density function for Vodafone stock. (b) A close-up of the extreme loss tail.

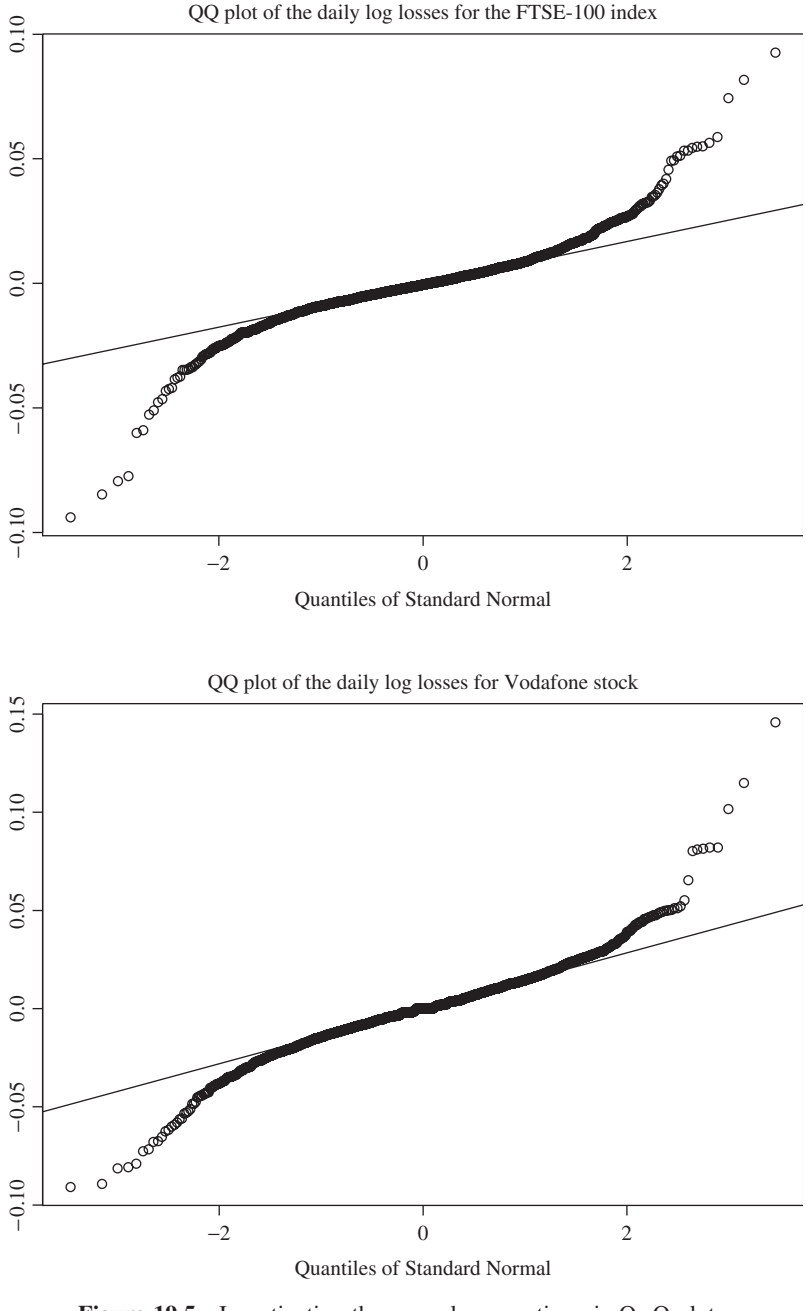


Figure 19.5 Investigating the normal assumption via Q−Q plots.

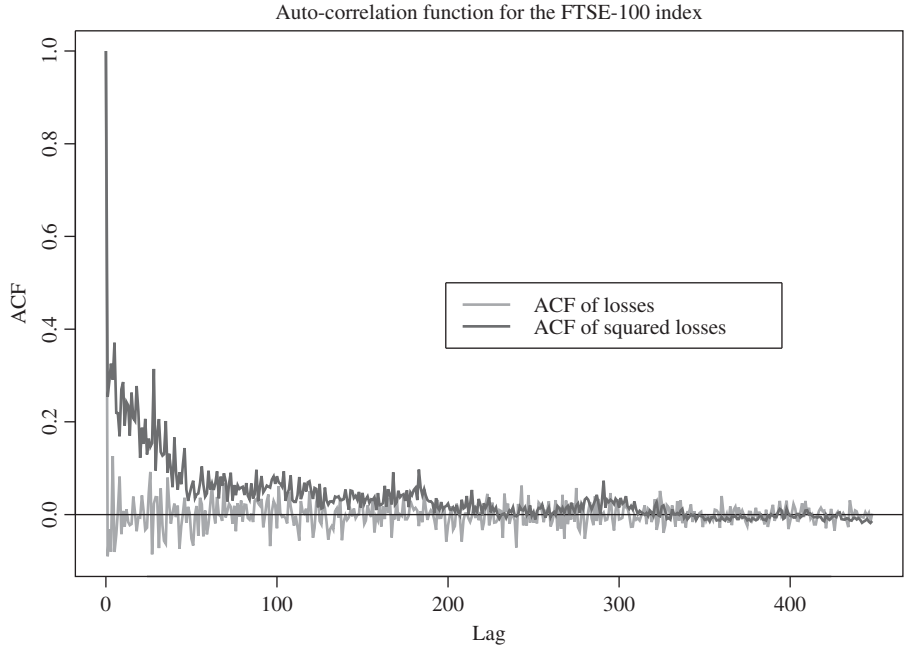


Figure 19.6 A plot of the auto-correlation function for the daily losses and squared daily losses of the FTSE-100 index.

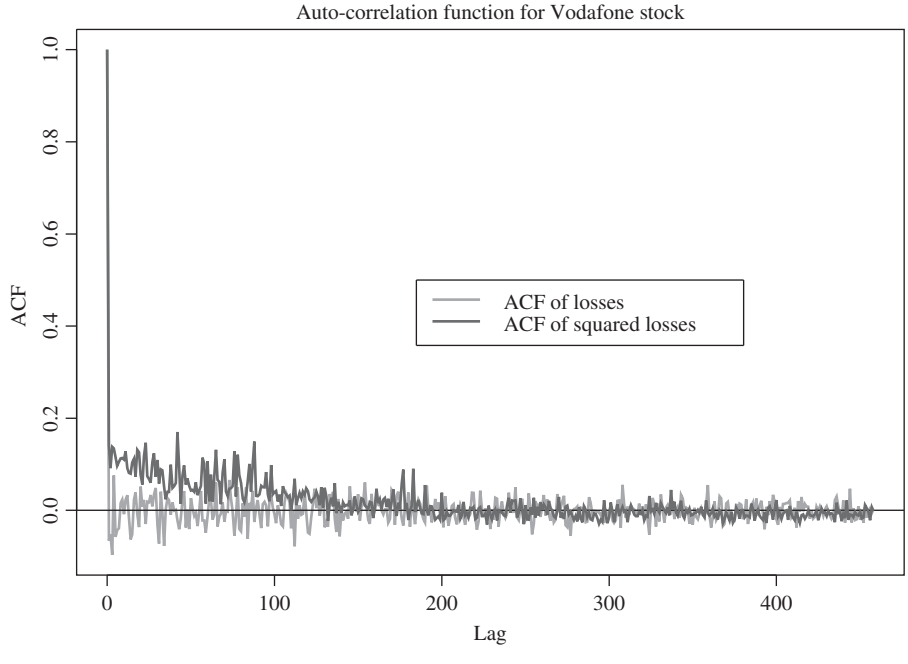


Figure 19.7 A plot of the auto-correlation function for the daily losses and squared daily losses of Vodafone stock.

19.4 THE VOLATILITY PLOT

Here we pick up from the previous section and present further evidence of the need to model volatility as a function of its recent past. We follow the experiment outlined in Section 19.1 and our aim is to distinguish between:

- the unconditional volatility of the loss process which is estimated using the full history of realized loss values, see (19.5);
- and conditional (time-dependent) volatility which is computed using only the past n realized values, see (19.6).

In our experiments we choose $n = 50$ which corresponds, approximately, to a 10-week rolling window of past values. The results from this experiment are displayed in Figures 19.8 and 19.9 for the FTSE-100 index and Vodafone stock respectively.

Observations

Based upon the visual evidence contained in Figures 19.8 and 19.9 we make the following observations:

- The unconditional volatility estimate, as expected, remains fairly flat as time evolves.
- On the other hand, the conditional estimate is a more faithful measure of risk because it reacts appropriately to periods of turbulence (where it is large) and to periods of tranquility (where its value is smaller).

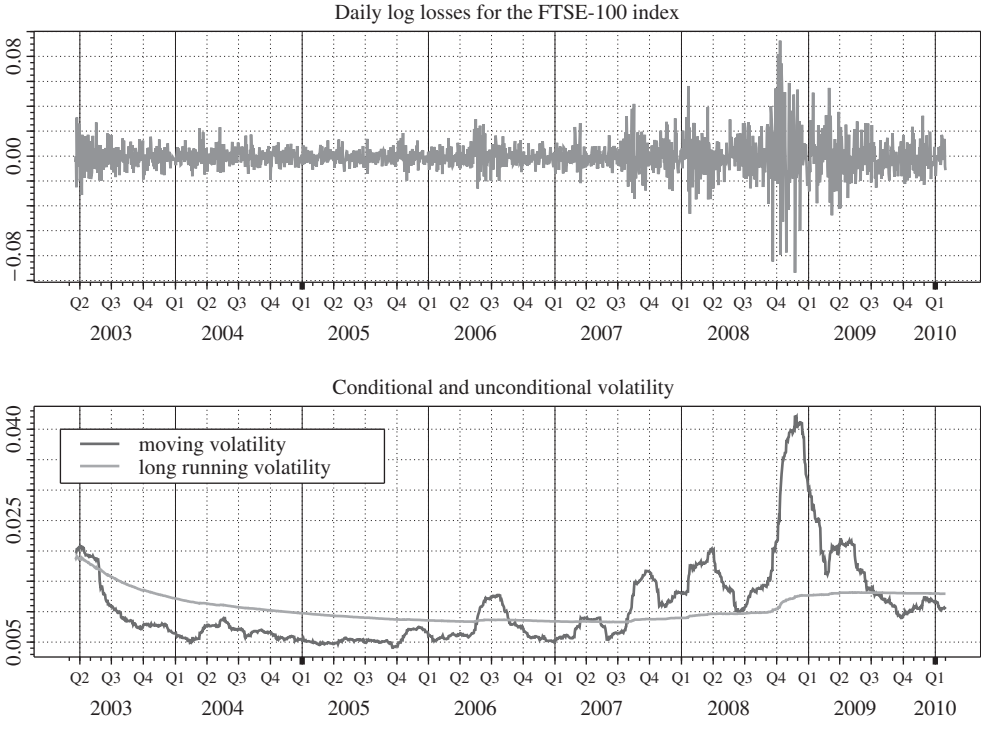


Figure 19.8 A plot to distinguish the unconditional and conditional volatility of the FTSE-100 index.

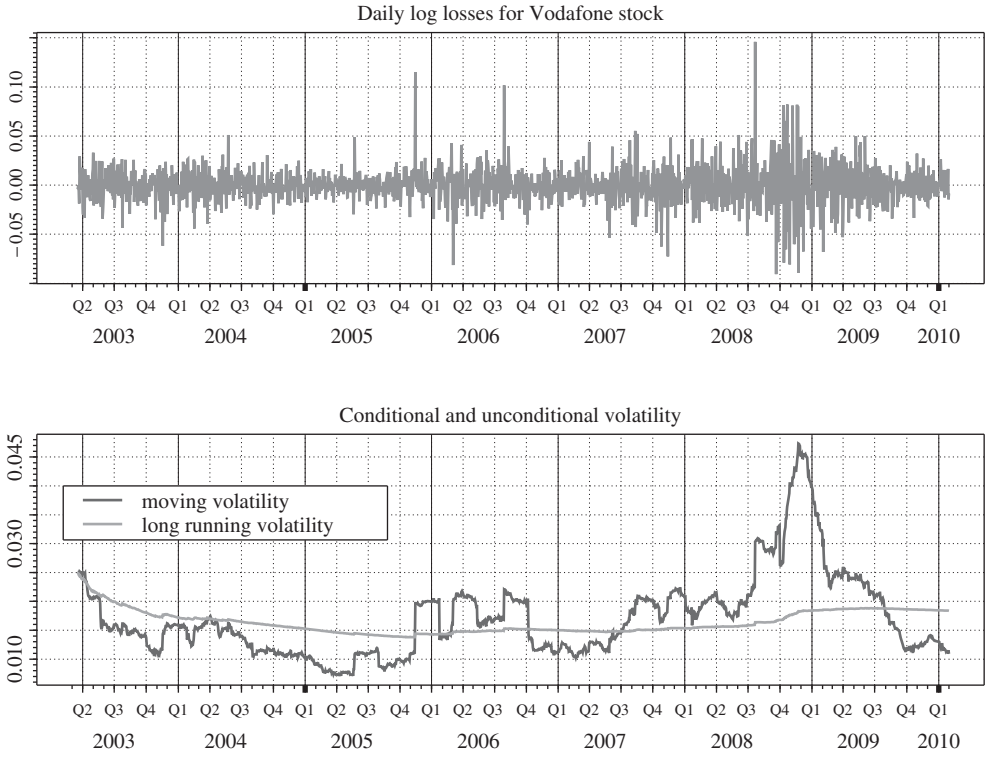


Figure 19.9 A plot to distinguish the unconditional and conditional volatility of Vodafone stock.

19.5 THE STYLIZED FACTS

The experiments we have carried out in this chapter for the FTSE-100 index and Vodafone stock represent a basic but thorough investigation into the statistical properties of financial losses. Over the years these experiments (and more sophisticated ones) have been repeated many times, for a wide range of financial assets, by academics and practitioners alike. It is remarkable that, in general, the same conclusions are made over and over again; so much so that these conclusions are now considered by the industry to be the stylized facts of financial losses. These facts are set out as follows:

- Fact 1. Negligible mean.
  The value of the mean for the daily loss random variable is small and is dominated by the value of the volatility. In view of this fact many practitioners simply ignore the mean when developing their daily loss models. However, one should be aware that for larger time intervals the mean cannot be ignored. Indeed, experiments show that the sample mean grows larger as the time interval grows.
- Fact 2. Fat tails and positive skewness.
  The unconditional distribution of the daily loss random variable is typically fat tailed and exhibits a slight positive skewness.

- Fact 3. Auto-correlation.
  The pure daily loss process exhibits very little serial correlation, i.e., its auto-correlation function decays very rapidly to zero. On the other hand, the auto-correlation function for squared losses decays slowly and so the squared losses show significant serial correlation.
- Fact 4. Volatility clustering.
  A plot of the daily loss process reveals that it encounters periods of tranquility followed by more volatile periods. We say that the process exhibits volatility clustering where extreme losses tend to trigger further extreme losses.

Essential Mathematics for Market Risk Management
by Simon Hubbert

20
Modelling Volatility

Our mathematical development so far has been built upon the assumption that a typical daily loss random variable of a financial asset or portfolio is normally distributed, i.e., we have assumed that

$$\mathcal{L}_t = \mu + \sigma \varepsilon_t, \tag{20.1}$$

where each ε_t is independently drawn from $N(0, 1)$; we often say that $(\varepsilon_t)_{t\geq 0}$ is the innovation process. However, as we now know (see previous chapter), when financial data are allowed to speak for themselves we find that (20.1) is not a good approximation and hence we should focus on developing an improved model which, at least, captures the following two important features:

- Fat tails.
  According to the second stylized fact of Section 19.6, the distribution of financial losses exhibits a tail that is thicker than the normal distribution; this indicates that extreme events (such as large losses) are more likely to occur than (20.1) would predict.
- Time-dependent volatility.
  According to the third and fourth stylized facts of Section 19.6, we conclude that the volatility of financial losses cannot be assumed to remain constant through time, as it is in (20.1). Indeed, a more realistic model must set out to capture the volatility clustering phenomenon which indicates that, in turbulent times, an extreme loss is likely to trigger more extreme losses.

By considering these two issues we seek to adjust our original model in order to make it more realistic. In particular, our aim is to construct a volatility model that takes into account the recent history of the asset and we propose the following:

$$\mathcal{L}_t = \mu + \sigma_t \varepsilon_t, \tag{20.2}$$

where ε_t is the usual innovation process which is assumed to be independent of the conditional volatility σ_t, whose square (the conditional variance) is given by

$$\sigma_t^2 = \mathbb{E}[(\mathcal{L}_t - \mu)^2 | I_{t-1}],$$

where I_{t-1} denotes the complete information set (for the loss process) that is available at time $t - 1$. The mathematical challenge is to develop an accurate model for σ_t^2 and our plan of attack is as follows:

- Specify a function f whose value is determined by a vector of free parameters; i.e., f is a function of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ say.
- We then use the information set I_{t-1} and attempt to find a vector $\boldsymbol{\theta}^*$ say, such that

$$\sigma_t^2 = \mathbb{E}[(\mathcal{L}_t - \mu)^2 | I_{t-1}] = f(I_{t-1} | \boldsymbol{\theta}^*). \tag{20.3}$$

The success of this method hinges upon how we choose the function f. Obviously if a risk manager is equipped with an accurate model then future decisions can be made with a greater degree of confidence; thus the challenge of finding a successful function f is a rewarding one.

20.1 THE RISKMETRICS MODEL

We take up the task of specifying a volatility function f (20.3) by presenting a few relatively simple ideas. By identifying the weaknesses of these basic proposals, we will find that we can gradually make improvements and so arrive at more sophisticated models. In particular, by taking this route of trial and improvement we are able to establish the famous RiskMetrics volatility model which is widely used in practice. In order to present a neat mathematical development we shall assume that the constant mean μ is equal to zero; we note that, at short intervals, this is a valid assumption (see the first stylized fact from the previous chapter).

Our first very simple idea is to model the variance (the square of volatility) as a simple moving average, i.e., we propose that

$$\sigma_t^2 = f(I_{[t-n,t-1]}) = \sum_{k=1}^{n} \frac{1}{n} \mathcal{L}_{t-k}^2. \tag{20.4}$$

We remark that this model is similar to what was used in the previous chapter, see (19.6), to reveal that volatility is indeed a time-varying function. The simple average serves as a quick and easy tool for analysing real data, however it should not be used as a serious model for volatility. The function itself is parameter-free and only depends upon a window of information that spans the previous n days (denoted by $[t-n, t-1]$); these two facts present their own problems.

- Equal weights.
  The fact that the model puts equal weight (equal to $1/n$) on the past observations yields unwarranted results. It implies that more distant events in the moving window will have the same influence on the volatility estimates as more recent ones. Clearly, this is not what we would expect.
- Choosing the size of the window.
  If the window is chosen to be relatively small then, statistically speaking, we cannot expect the estimate to be an accurate one. On the other hand, choosing a window that is too large also presents problems. For example, if an extreme shock occurs then it will continue to influence the future volatility estimates for a long time. In particular, if the true market has settled back into a tranquil period then this will not be reflected in the volatility estimates. Furthermore, when the extreme shock finally does drop out of the window we can expect a sudden drop in the next day's estimate; we want to avoid these scenarios.

In order to overcome the weaknesses of (20.4) we decide to replace the equal weights with a sequence of declining (positive) weights

$$w_1 > w_2 > \cdots > w_n > 0,$$

so that more weight is attached to recent observations and less to more distant ones. If the weights sum to one, then our volatility model has the form

$$\sigma_t^2 = f(I_{[t-n,t-1]}|(w_1, \ldots, w_n)) = \sum_{k=1}^{n} w_k \mathcal{L}_{t-1}^2$$

and if the weights do not sum to one we simply normalize them and consider

$$\sigma_t^2 = f(I_{[t-n,t-1]}|(w_1, \ldots, w_n)) = \frac{\sum_{k=1}^{n} w_k \mathcal{L}_{t-k}^2}{\sum_{k=1}^{n} w_k}. \tag{20.5}$$

The obvious drawback of this model is that we have now introduced a potentially large number of parameters to which we must assign appropriate values, and this can be a computational burden. In order to ease this problem we propose that the weights have the form

$$w_k = \lambda^{k-1} \quad \text{where} \quad \lambda \in (0, 1), \quad \text{for} \quad k = 1, \ldots, n.$$

These weights have the declining property and their sum is given by

$$\sum_{k=1}^{n} \lambda^{k-1} = 1 + \lambda + \cdots + \lambda^{n-1} = \frac{1 - \lambda^n}{1 - \lambda}.$$

In this framework, (20.5) becomes the so-called exponentially weighted moving average model given by

$$\sigma_t^2 = f(I_{[t-n,t-1]}|\lambda) = \frac{1 - \lambda}{1 - \lambda^n} \sum_{k=1}^{n} \lambda^{k-1} \mathcal{L}_{t-k}^2. \tag{20.6}$$

This model now exhibits two helpful properties:

1. The weights are designed to attach more importance to recent events than to those in the distant past.
2. Only one parameter $\lambda \in (0, 1)$ is needed to completely define the model.

Unfortunately, the problem of how large to choose the window still remains. In order to overcome this we look at the model from a mathematical perspective and note that by taking the limit of (20.6) as $n \to \infty$, we incorporate all known information into the model (and so circumvent the question of how to choose n). This limiting process leads to the following (intriguing) development:

$$\begin{aligned} \sigma_t^2 &= \lim_{n \to \infty} f(I_{[t-n,t-1]}|\lambda) \\ &= (1 - \lambda) \sum_{k=1}^{\infty} \lambda^{k-1} \mathcal{L}_{t-k}^2 \end{aligned}$$

$$= (1-\lambda)\mathcal{L}_{t-1}^2 + (1-\lambda)\underbrace{\sum_{k=2}^{\infty}\lambda^{k-1}\mathcal{L}_{t-k}^2}_{\text{re-index sum}}$$

$$= (1-\lambda)\mathcal{L}_{t-1}^2 + (1-\lambda)\sum_{k=1}^{\infty}\lambda^{k}\mathcal{L}_{t-(k+1)}^2$$

$$= (1-\lambda)\mathcal{L}_{t-1}^2 + \lambda\underbrace{(1-\lambda)\sum_{k=1}^{\infty}\lambda^{k-1}\mathcal{L}_{t-1-k}^2}_{\text{equals } \sigma_{t-1}^2 \text{ by (20.6)}}.$$

Hence we can write

$$\sigma_t^2 = f(I_{t-1}|\lambda) = (1-\lambda)\mathcal{L}_{t-1}^2 + \lambda\sigma_{t-1}^2. \tag{20.7}$$

This is the famous RiskMetrics volatility model and it states that a new volatility estimate is modelled as a linear combination of the previous day's estimate and the previous day's squared loss. The weight λ is chosen to minimize the error between the estimated and measured volatilities over some sample period. One of the appealing advantages of this method is the fact that the RiskMetrics group, the developers of the model, have made freely available a large amount of their technical research. Indeed, this research includes results regarding how well the model performs on a wide range of asset classes and even provides guidance on how best to choose an appropriate λ for a specific problem. In the late 1990s the RiskMetrics model established itself as the industry standard. Over the years there have been several revisions to the basic model and it is still widely used today. We will return to this model later in the chapter.

20.2 ARCH MODELS

The ARCH family of volatility models was first proposed by Robert Engle (1982) in the early 1980s. In the ARCH framework the idea is that the volatility function (20.3) should be a linear combination of the squared losses over the previous p days, i.e., we propose that

$$\begin{aligned}\sigma_t^2 &= f(I_{t-1}|\omega, \alpha_1, \ldots, \alpha_p)\\ &= \omega + \alpha_1\mathcal{L}_{t-1}^2 + \cdots + \alpha_p\mathcal{L}_{t-p}^2.\end{aligned} \tag{20.8}$$

Thus, the ARCH model is specified by the $p+1$ parameters $\omega, \alpha_1, \ldots, \alpha_p$. In order to guarantee that the estimated volatility (variance) is positive at all times, we insist that

$$\begin{aligned}&\text{the intercept} \quad \omega > 0\\ &\text{and also that} \quad \alpha_k \geq 0 \quad \text{for} \quad k = 1, \ldots, p.\end{aligned} \tag{20.9}$$

The reason why this class of model has been assigned the quirky name ARCH is explained as follows:

- The parameters $\omega, \alpha_1, \ldots, \alpha_p$ are commonly referred to as the *auto-regressive* coefficients as their estimation is based on the past values of the process. This accounts for the A and the R of ARCH.
- The model itself is designed to capture the conditional volatility of the process as a function of time. In probabilistic terminology such a random process is said to be *conditionally heteroscedastic* and it is this terminology that accounts for the C and the H of ARCH.

An ARCH model that relies on the previous p days' losses is commonly referred to as ARCH(p). Now, when such a model is incorporated into (20.2), with $\mu = 0$, it leads to

$$\mathcal{L}_t = \left(\sqrt{\omega + \sum_{k=1}^{p} \alpha_k \mathcal{L}_{t-k}^2} \right) \varepsilon_t, \tag{20.10}$$

where the auto-regressive parameters satisfy (20.9). In fact, as we shall see, a stronger condition on these parameters is needed if we are to insist that the loss process $\mathcal{L}_t$ be a stationary one, i.e., that the unconditional mean and variance remain constant through time.

We observe, from (20.10), that the unconditional mean for the ARCH-driven process is zero, since

$$\mathbb{E}[\mathcal{L}_t] = \mathbb{E}\left[\sqrt{\omega + \sum_{k=1}^{p} \alpha_k \mathcal{L}_{t-k}^2} \right] \underbrace{\mathbb{E}[\varepsilon_t]}_{=0} = 0.$$

To investigate unconditional variance we consider the following development:

$$\begin{aligned} \sigma^2 = \mathbb{E}[\mathcal{L}_t^2] &= \mathbb{E}[\sigma_t^2] \underbrace{\mathbb{E}[\varepsilon_t^2]}_{=1} \\ &= \mathbb{E}\left[\omega + \sum_{k=1}^{p} \alpha_k \mathcal{L}_{t-k}^2 \right] \\ &= \omega + \sum_{k=1}^{p} \alpha_k \underbrace{\mathbb{E}[\mathcal{L}_{t-k}^2]}_{=\sigma^2} \\ &= \omega + \sigma^2 \sum_{k=1}^{p} \alpha_k. \end{aligned} \tag{20.11}$$

Rearranging this we find that the long-run, unconditional, variance is given by

$$\sigma^2 = \frac{\omega}{1 - \sum_{k=1}^{p} \alpha_k}. \tag{20.12}$$

Thus, for the loss process to be stationary, we must impose the stronger conditions that

$$\omega > 0, \quad \alpha_k \geq 0 \quad (1 \leq k \leq p) \quad \text{and} \quad \sum_{k=1}^{p} \alpha_k < 1. \tag{20.13}$$

We remark that, in view of our findings in the previous chapter, the assumption of stationarity is a valid one and so any implementation of the ARCH(p) model ought to insist that its parameters satisfy (20.13).

20.2.1 The ARCH(1) volatility model

The success or otherwise of any volatility model will be judged on how well it captures the basic properties of the loss process of a real financial asset/portfolio. In order to gain more insight into the ARCH class of models we choose to examine the simplest case where $p = 1$. In this setting our loss process is modelled as $\mathcal{L}_t = \sigma_t \varepsilon_t$, where ε_t denotes an innovation process and where σ_t is determined by the ARCH(1) model

$$\sigma_t^2 = \omega + \alpha \mathcal{L}_{t-1}^2 \quad \text{where} \quad \omega > 0 \quad \text{and} \quad 0 \leq \alpha < 1.$$

Our aim is to investigate this model and specifically we will aim to establish answers to the following three questions:

1. Does the model lead to a fat-tailed loss distribution?
2. Does the model capture the volatility clustering phenomenon?
3. Does the loss model exhibit negligible auto-correlation? In addition, do the squared losses display a significant degree of auto-correlation?

20.2.1.1 Investigating the tail thickness

In order to examine the tail thickness of the ARCH(1) loss model we set out to compute its kurtosis coefficient and thus we need to find:

- The unconditional variance, i.e., the number σ^2 that satisfies

$$\mathbb{E}[\mathcal{L}_t^2] \quad \text{for all} \quad t.$$

- The unconditional fourth moment, i.e., m_4 that satisfies

$$m_4 = \mathbb{E}[\mathcal{L}_t^4] \quad \text{for all} \quad t.$$

We begin this calculation process by observing that, according to (20.12), the imposed parameter conditions imply that the unconditional variance of the ARCH(1) model is given by

$$\sigma^2 = \frac{\omega}{1 - \alpha}. \tag{20.14}$$

We compute m_4 in the same way as we computed the unconditional variance, see (20.11), and we begin with the following development:

$$\begin{aligned} m_4 = \mathbb{E}[\mathcal{L}_t^4] &= \mathbb{E}[\sigma_t^4]\underbrace{\mathbb{E}[\varepsilon_t^4]}_{=3} \\ &= 3\mathbb{E}\left[\left(\omega + \alpha\mathcal{L}_{t-1}^2\right)^2\right] \\ &= 3\left[\omega^2 + 2\omega\alpha\underbrace{\mathbb{E}[\mathcal{L}_{t-1}^2]}_{=\sigma^2} + \alpha^2\underbrace{\mathbb{E}[\mathcal{L}_{t-1}^4]}_{=m_4}\right] \\ &= 3\left[\omega^2 + 2\omega\alpha\sigma^2 + \alpha^2 m_4\right]. \end{aligned} \tag{20.15}$$

Rearranging the above we find that

$$m_4 = 3\left[\frac{\omega^2 + 2\omega\alpha\sigma^2}{1 - 3\alpha^2}\right] = 3\omega\left[\frac{\omega + 2\alpha\sigma^2}{(1 - 3\alpha^2)}\right].$$

Furthermore, (20.14) tells us that $\omega = (1 - \alpha)\sigma^2$ and substituting this into the above formula gives

$$m_4 = \frac{3(1 - \alpha^2)\sigma^4}{(1 - 3\alpha^2)}, \tag{20.16}$$

which is well defined provided $0 \leq \alpha \leq 1/\sqrt{3}$, and in which case, the kurtosis coefficient is

$$\mathcal{K}_{\text{ARCH}} = \frac{m_4}{\sigma^4} = 3\left[\frac{1 - \alpha^2}{1 - 3\alpha^2}\right] > 3. \tag{20.17}$$

The above development allows us to conclude that a zero-mean loss process (20.2) whose conditional volatility σ_t is driven by the ARCH(1) model

$$\sigma_t^2 = \omega + \alpha\mathcal{L}_{t-1}^2 \quad \text{with} \quad \omega > 0 \quad \text{and} \quad 0 \leq \alpha < \frac{1}{\sqrt{3}} \tag{20.18}$$

has a distribution whose tail is fatter than the normal distribution.

20.2.1.2 Forecasting with ARCH(1)

We now examine the forecasting ability of our ARCH(1) model. We know, for instance, that under the ARCH(1) model, the τ-day-ahead estimate of conditional variance satisfies

$$\begin{aligned} \sigma_{t+\tau}^2 &= \omega + \alpha\mathcal{L}_{t+\tau-1}^2 \\ &= \sigma^2(1 - \alpha) + \alpha\mathcal{L}_{t+\tau-1}^2 \quad \text{using (20.14)}, \end{aligned}$$

and this can be rearranged to yield

$$\sigma_{t+\tau}^2 - \sigma^2 = \alpha(\mathcal{L}_{t+\tau-1}^2 - \sigma^2). \tag{20.19}$$

If the information set I_t up to date t is available, then the 1-day forecast is known and is given by

$$\sigma_{t+1}^2 = \sigma^2 + \alpha(\mathcal{L}_t^2 - \sigma^2).$$

For $\tau > 1$ we let

$$\widehat{\sigma_{t+\tau}^2} = \mathbb{E}[\mathcal{L}_{t+\tau}^2 | I_t] = \mathbb{E}[\sigma_{t+\tau}^2 | I_t] \tag{20.20}$$

denote the τ-day-ahead variance forecast which then, using (20.19), satisfies the following one-step recursion:

$$\widehat{\sigma_{t+\tau}^2} - \sigma^2 = \alpha(\widehat{\sigma_{t+\tau-1}^2} - \sigma^2). \tag{20.21}$$

We can apply (20.21) repeatedly to establish that

$$\widehat{\sigma_{t+\tau}^2} - \sigma^2 = \alpha^{\tau-1}(\sigma_{t+1}^2 - \sigma^2). \tag{20.22}$$

This forecasting formula is useful because it enables us to shed further light upon the mechanics of the ARCH(1) model. For instance, we note that the forecast uses only the most recent component of our information set, namely the current variance estimate σ_{t+1}^2. Furthermore, the forecast is obtained by projecting $(\sigma_{t+1}^2 - \sigma)$ out into the future by a factor $\alpha^{\tau-1}$. Thus, the larger the value of α the more influence the current estimate will have on the future, as a result we often refer to α as the persistence parameter. Finally, since $0 \leq \alpha < 1/\sqrt{3}$ we observe that, eventually, in the distant future, the influence of the current estimate will be almost negligible. Indeed, taking the limit as $\tau \to \infty$ we find that

$$\lim_{\tau \to \infty} \mathbb{E}\left[\sigma_{t+\tau}^2 | I_t\right] = \sigma^2 \quad \text{(the unconditional variance)}.$$

20.2.1.3 Auto-correlation

We know from the previous chapter that a general daily loss process tends not to display a significant correlation with its recent past; the auto-correlation function decays rapidly. To investigate whether the ARCH(1) process captures this property we define a time lag τ and examine the covariance between $\mathcal{L}_t$ and $\mathcal{L}_{t-\tau}$. We evoke the following form of the iterated law of expectations (for a general random process X_t):

$$\mathbb{E}[X_t] = \mathbb{E}\left[\mathbb{E}[X_t | I_{t-1}]\right]$$

to deduce that, for $\tau > 0$, we have

$$\begin{aligned}
\mathbb{E}[\mathcal{L}_t \mathcal{L}_{t-\tau}] &= \mathbb{E}\left[\mathbb{E}[\mathcal{L}_t \mathcal{L}_{t-\tau} | I_{t-1}]\right] \\
&= \mathbb{E}\left[\mathcal{L}_{t-\tau} \underbrace{\mathbb{E}[\mathcal{L}_t | I_{t-1}]}_{=0}\right] \\
&= 0.
\end{aligned} \tag{20.23}$$

Thus, we can conclude that the ARCH(1) loss process is serially uncorrelated.

The next task is to analyse how the squared losses of the ARCH(1) process are correlated. We recall from the previous chapter that the squared losses of real financial time series tend to exhibit a significant auto-correlation. We begin our investigation with the following useful development:

$$\begin{aligned}\mathcal{L}_t^2 &= \sigma_t^2 + \left(\mathcal{L}_t^2 - \sigma_t^2\right) \\ &= \omega + \alpha \mathcal{L}_{t-1}^2 + \left(\sigma_t^2 \varepsilon_t^2 - \sigma_t^2\right) \\ &= \omega + \alpha \mathcal{L}_{t-1}^2 + v_t, \quad \text{where} \quad v_t = \sigma_t^2(\varepsilon_t^2 - 1). \end{aligned} \tag{20.24}$$

We remark that the above squared loss model bears a close resemblance to the familiar AR(1) model that we originally encountered in Chapter 15. Indeed, the following result establishes that (20.24) does in fact coincide with an AR(1) process:

Proposition 20.1. *The residual process v_t appearing in (20.24) is a white noise process.*

Proof. We prove the result by verifying that v_t possesses the three properties which characterize a white noise process.

1. Zero mean:

$$\mathbb{E}[v_t] = \mathbb{E}[\sigma_t^2(\varepsilon_t^2 - 1)] = \mathbb{E}[\sigma_t^2] \underbrace{\mathbb{E}[\varepsilon_t^2 - 1]}_{=0} = 0.$$

2. Constant variance:

$$\begin{aligned}\mathbb{E}[v_t^2] = \mathbb{E}[\sigma_t^4(\varepsilon_t^2 - 1)^2] &= \underbrace{\mathbb{E}[\sigma_t^4]}_{=m_4/3} \underbrace{\mathbb{E}[\varepsilon_t^4 - 2\varepsilon^2 + 1]}_{=2} \\ &= \frac{2(1-\alpha^2)\sigma^4}{1-3\alpha^2}, \quad \text{using (20.16).}\end{aligned}$$

3. Serially uncorrelated:
   let $\tau > 0$ then we have

$$\begin{aligned}\operatorname{cov}(v_t, v_{t+\tau}) &= \mathbb{E}[\sigma_t^2(\varepsilon_t^2 - 1)\sigma_{t+\tau}^2(\varepsilon_{t+\tau}^2 - 1)] \\ &= \mathbb{E}[\sigma_t^2 \sigma_{t+\tau}^2(\varepsilon_t^2 - 1)] \underbrace{\mathbb{E}[\varepsilon_{t+\tau}^2 - 1]}_{=0} = 0.\end{aligned}$$

□

We can now exploit our knowledge of AR(1) models (see Chapter 15) to make the following deductions:

- Using (15.10) we rediscover our formula for the unconditional variance of the process, namely that

$$\mathbb{E}[\mathcal{L}_t^2] = \frac{\omega}{1-\alpha} = \sigma^2.$$

- Using (15.13) we verify that the squared loss process does indeed exhibit a significant degree of auto-correlation given by

$$\rho_{\text{arch}}(\tau) = \frac{\mathbb{E}[\mathcal{L}_t^2 \mathcal{L}_{t-\tau}^2]}{\text{var}(\mathcal{L}_t^2)} = \alpha^\tau \quad \tau > 0. \tag{20.25}$$

20.2.1.4 Overview

Our investigations have revealed that the ARCH(1) loss model (20.18) possesses many appealing properties. Indeed, it has been constructed so as to capture the crucial stylized facts including fat tails, volatility clustering and a lack of serial correlation. The model does, however, have its drawbacks. Firstly, we have observed that the model parameter α is required to satisfy $0 \leq \alpha < 1/\sqrt{3}$. Unfortunately, this is often too restrictive and one finds that the real clustering phenomenon is not accurately captured. This problem can be overcome by employing the more general ARCH(p) model (with $p > 1$). Practical experiments show that a large value of p is often required and this adds a further computational burden of estimating a potentially large number of parameters.

20.3 GARCH MODELS

In the mid-1980s Robert Engle, discoverer of the ARCH model, set his PhD student Tim Bollerslev onto a project to generalize ARCH with a view to finding an approach that is easier to implement. We recall that one of the drawbacks of ARCH is that, in practice, it is often the case that a large value of p is needed to capture the volatility clustering phenomenon. The idea that Bollerslev pursued was actually to augment the ARCH(p) model by allowing the conditional variance to depend upon the past squared losses and also the past variance estimates. Mathematically speaking, he proposed the following model (Bollerslev, 1986):

$$\begin{aligned} \mathcal{L}_t = \sigma_t \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim N(0,1) \quad \text{independently drawn} \\ \text{and} \quad \sigma_t^2 = \omega + \sum_{k=1}^{p} \alpha_k \mathcal{L}_{t-k}^2 + \sum_{k=1}^{q} \beta_k \sigma_{t-k}^2 \\ \text{where} \quad \omega > 0, \quad \alpha_k \geq 0 \quad (1 \leq k \leq p) \\ \text{and} \quad \beta_k \geq 0 \quad (1 \leq k \leq q). \end{aligned} \tag{20.26}$$

This generalized ARCH model is better known as GARCH(p, q). We note that rather than reduce the computational burden the model appears to do the opposite as q additional parameters are introduced. However, the reason why this generalization has gained significant attention is that, in practice, one finds that relatively low values of p and q are needed for the model to faithfully reflect the clustering phenomenon; indeed, the simplest case of GARCH(1, 1) is by far the most commonly used version in financial applications.

20.3.1 The GARCH(1, 1) volatility model

Owing to its popularity we shall now devote attention to the GARCH(1, 1) model of volatility. To fix our notation we shall express the model as follows:

$$\begin{aligned} \mathcal{L}_t = \sigma_t \varepsilon_t \quad &\text{where} \quad \varepsilon_t \sim N(0,1) \quad \text{independently drawn} \\ &\text{and} \quad \sigma_t^2 = \omega + \alpha \mathcal{L}_{t-1}^2 + \beta \sigma_{t-1}^2 \\ &\text{where} \quad \omega > 0 \quad \text{and} \quad \alpha, \beta \geq 0. \end{aligned} \tag{20.27}$$

Our aim is to deliver the crucial statistical properties of the above model and we will follow the same approach as we applied to the ARCH(1) model.

20.3.1.1 Unconditional mean and variance

The unconditional mean of the GARCH(1, 1) process is zero since

$$\mathbb{E}[\mathcal{L}_t] = \mathbb{E}[\sigma_t \varepsilon_t] = \mathbb{E}[\sigma_t] \underbrace{\mathbb{E}[\varepsilon_t]}_{=0} = 0.$$

The unconditional variance is defined by

$$\begin{aligned} \sigma^2 = \mathbb{E}[\mathcal{L}_t^2] &= \mathbb{E}[\sigma_t^2 \varepsilon_t^2] \\ &= \mathbb{E}[\sigma_t^2] \mathbb{E}[\varepsilon_t^2] \\ &= \mathbb{E}[\sigma_t^2] \quad \text{for all} \quad t \geq 0. \end{aligned}$$

Using this definition and (20.27) we find that σ^2 must satisfy

$$\begin{aligned} \sigma^2 &= \mathbb{E}[\omega + \alpha \mathcal{L}_{t-1}^2 + \beta \sigma_{t-1}^2] \\ &= \omega + (\alpha + \beta)\sigma^2. \end{aligned}$$

Solving this equation for σ^2 yields

$$\sigma^2 = \frac{\omega}{1 - (\alpha + \beta)}. \tag{20.28}$$

We note that this formula is the analogue of (20.14) from the ARCH(1) setting. The formula itself tells us that (20.27) describes a stationary process whenever $\alpha + \beta < 1$.

20.3.1.2 Investigating tail thickness

To calculate the kurtosis coefficient of (20.27) we begin by deriving its fourth unconditional moment, defined by

$$\begin{aligned} m_4 = \mathbb{E}[\mathcal{L}_t^4] &= \mathbb{E}[\sigma_t^4 \varepsilon_t^4] \\ &= \mathbb{E}[\sigma_t^4] \underbrace{\mathbb{E}[\varepsilon_t^4]}_{=3} = 3\mathbb{E}[\sigma_t^4] \quad \text{for all} \quad t \geq 0. \end{aligned} \tag{20.29}$$

We now substitute the GARCH(1, 1) model and compute as follows:

$$\begin{aligned} m_4 &= 3\mathbb{E}\left[\left(\omega + \alpha\mathcal{L}_{t-1}^2 + \beta\sigma_{t-1}^2\right)^2\right] \\ &= 3\mathbb{E}\left[\omega^2 + \alpha^2\mathcal{L}_{t-1}^4 + \beta^2\sigma_{t-1}^4 \right. \\ &\left. +2\left(\alpha\omega\mathcal{L}_{t-1}^2 + \beta\omega\sigma_{t-1}^2 + \alpha\beta\underbrace{\mathcal{L}_{t-1}^2\sigma_{t-1}^2}_{=\sigma_{t-1}^4\varepsilon_{t-1}^2}\right)\right]. \end{aligned} \tag{20.30}$$

Now, $\sigma^2 = \mathbb{E}[\mathcal{L}_t^2] = \mathbb{E}[\sigma_t^2]$ denotes the unconditional variance and so, with this in mind, we can take expectations above to reveal that

$$m_4 = 3\omega^2 + 3\alpha^2\underbrace{\mathbb{E}[\mathcal{L}_{t-1}^4]}_{=m_4} + \beta^2\underbrace{\left(3\mathbb{E}[\sigma_t^4]\right)}_{=m_4} + 6\omega(\alpha+\beta)\sigma^2 + 2\alpha\beta\cdot\underbrace{\left(3\mathbb{E}[\sigma_t^4\varepsilon_t^2]\right)}_{=m_4},$$

i.e., we have

$$m_4\left(1 - 2\alpha^2 - (\alpha+\beta)^2\right) = 3\omega\left(\omega + 2(\alpha+\beta)\sigma\right).$$

Finally, we can use the fact that $\omega = \sigma^2(1-(\alpha+\beta))$ to deduce that the fourth moment, given by

$$m_4 = \frac{3\sigma^4\left(1-(\alpha+\beta)^2\right)}{1-2\alpha^2-(\alpha+\beta)^2},$$

is well defined provided that $(\alpha+\beta)^2 < 1 - 2\alpha^2$ and, in which case, the corresponding kurtosis coefficient is given by

$$\mathcal{K}_{\text{GARCH}} = \frac{m_4}{\sigma^4} = \frac{3\left(1-(\alpha+\beta)\right)^2}{1-2\alpha^2-(\alpha+\beta)^2} > 3. \tag{20.31}$$

Thus, in analogy to the ARCH(1) process, we are able to conclude that a loss process (20.2) whose conditional volatility σ_t is driven by the GARCH(1, 1) model:

$$\begin{aligned} &\sigma_t^2 = \omega + \alpha\mathcal{L}_{t-1}^2 + \beta\sigma_{t-1}^2 \\ &\text{with} \quad \omega, \alpha, \beta > 0, \quad \alpha+\beta < 1 \quad \text{and} \quad (\alpha+\beta)^2 < 1 - 2\alpha^2, \end{aligned} \tag{20.32}$$

has a distribution whose tail is fatter than the normal distribution.

20.3.1.3 Forecasting with GARCH(1, 1)

We recall that in the ARCH(1) setting we exploited the recursive nature of the process to forecast future estimates of its variance; see (20.22). The exact same arguments can be applied to the GARCH(1, 1) model and these lead to the following analogous formula which holds for $\tau > 1$:

$$\widehat{\sigma^2_{t+\tau}} - \sigma^2 = (\alpha + \beta)^{\tau-1}(\sigma^2_{t+1} - \sigma^2). \tag{20.33}$$

We note that this model, as with ARCH(1), has the property that it predicts the variance of the process will converge to its unconditional value σ^2 (20.28) the further we look into the future. As a result we say that both ARCH(1) and GARCH(1, 1) are mean-reverting processes.

20.3.1.4 Auto-correlation

Given that the GARCH(1, 1) loss process has zero conditional mean we can establish, in the same way as for the ARCH case (20.23), that the process exhibits no serial correlation, i.e.,

$$\mathbb{E}[\mathcal{L}_t \mathcal{L}_{t-\tau}] = 0 \quad \text{for} \quad \tau > 0.$$

The investigation into the auto-correlation of the squared process follows the same lines as that for the ARCH case but, as we shall see, it leads to a different conclusion. We recall that we have demonstrated that, under the ARCH(1) model, the squared loss process can be viewed as an AR(1) process. We will show that, in contrast, if the loss process follows a GARCH(1, 1) model then its square can be viewed as an ARMA(1, 1) process. The derivation is as follows:

$$\begin{aligned}\mathcal{L}^2_t &= \sigma^2_t + (\mathcal{L}^2_t - \sigma^2_t) \\ &= \omega + \alpha \mathcal{L}^2_{t-1} + \beta \sigma^2_{t-1} + \underbrace{\sigma^2_t(\varepsilon^2_t - 1)}_{v_t}.\end{aligned}$$

We have already established that $v_t = \sigma^2_t(\varepsilon^2_t - 1)$ is a white noise process; see Proposition 20.1. We now notice that

$$v_{t-1} = \sigma^2_{t-1}(\varepsilon^2_{t-1} - 1) = \mathcal{L}^2_{t-1} - \sigma^2_{t-1}$$

and so we have

$$\begin{aligned}\mathcal{L}^2_t &= \omega + \alpha \mathcal{L}^2_{t-1} + \beta(\mathcal{L}^2_{t-1} - v_{t-1}) + v_t \\ &= \omega + (\alpha + \beta)\mathcal{L}^2_{t-1} - \beta v_{t-1} + v_t.\end{aligned}$$

Now, we can substitute the formula $\omega = (1 - (\alpha + \beta))\sigma^2$ into the above expression and rearrange to give

$$\mathcal{L}^2_t - \sigma^2 = (\alpha + \beta)\left(\mathcal{L}^2_{t-1} - \sigma^2\right) + v_t - \beta v_{t-1}. \tag{20.34}$$

We now compare this with (15.15) to conclude that if the loss process is described by a GARCH(1, 1) model, then the shifted squared loss process defined by

$$Y_t := \mathcal{L}_t^2 - \sigma^2$$

can be viewed as an ARMA(1, 1) process. Since we have already encountered such processes in Chapter 15 we can take advantage of what we know to reveal some interesting properties of the squared loss process. In particular, we know that ARMA(1, 1) processes have zero mean and so, in our case, this implies

$$\mathbb{E}[\mathcal{L}_t^2] = \frac{\omega}{1 - (\alpha + \beta)},$$

and so verifies our earlier calculation of the unconditional variance (20.28). Furthermore, using (15.22), the ARMA(1, 1) autocorrelation formula, we can deduce that the auto-correlation of the squared losses is given by

$$\rho_{\text{GARCH}}(\tau) = (\alpha + \beta)^{\tau - 1} \frac{\alpha(1 - \beta(\alpha + \beta))}{1 + \beta^2 - 2\alpha(1 - \beta(\alpha + \beta))} \quad \tau > 0. \tag{20.35}$$

Thus we conclude that the GARCH(1, 1) loss model (20.32) captures the main features of real financial losses: a fat-tailed distribution, negligible auto-correlation of pure losses and significant auto-correlation in the squared losses (the clustering effect). The model itself is widely used in practice and can be viewed as the prototype for a whole host of new and improved approaches.

20.3.2 The RiskMetrics model revisited

The very first well-established volatility process that we encountered in this chapter was the RiskMetrics model (20.7). We recall that this famous model is completely specified by a single parameter $\lambda \in (0, 1)$ and it takes the following form:

$$\mathcal{L}_t = \sigma_t \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim N(0, 1) \quad \text{independently drawn}$$

$$\text{and} \quad \sigma_t^2 = (1 - \lambda)\mathcal{L}_{t-1}^2 + \lambda \sigma_{t-1}^2. \tag{20.36}$$

Our derivation of the RiskMetrics model was accomplished without any knowledge of ARCH or GARCH processes. However, in retrospect, when we compare (20.36) with (20.27) we can see that RiskMetrics is in fact a special kind of GARCH(1, 1) process where the intercept $\omega = 0$ and where the sum of the parameters α and β is set equal to one. In view of this new interpretation we use our knowledge of the GARCH framework to make the following remarks:

- The RiskMetrics process is not weakly stationary since it has infinite variance due to formula (20.28). However, it can be shown that the process is strictly stationary; see Definition 15.1.
- The RiskMetrics model remains popular with practitioners despite the fact that its long-term volatility is infinite. One reason why practitioners have not been put off RiskMetrics is that, in practice, one commonly finds that the estimated GARCH(1, 1) parameters α and β have the property that $\alpha + \beta \approx 1$. Thus, it is often judged that simple RiskMetrics implementation should be chosen over the GARCH(1, 1) model.

20.3.3 Summary

Engle's ARCH and Bollerslev's GARCH models are both designed to track the way the volatility of a typical loss process varies over time. However, the GARCH framework has the crucial advantage that many fewer parameters need to be estimated in order to establish a model that delivers faithful forecasts. As a result, the GARCH methodology is widely implemented in practise, usually the GARCH(1, 1) version. Indeed, implementations of the GARCH model will be found in almost all financial institutions where its output can be used in a whole host of applications, from derivative pricing to risk management.

20.4 EXPONENTIAL GARCH

Robert Engle first described his ARCH framework for modelling time-varying volatility in 1982. This work was to prove ground-breaking as, at that time, it was the first model of its kind with the ability to accurately capture what we now know as the stylized facts of asset returns/losses. Furthermore, the success of Tim Bollerslev's more general GARCH framework then sparked a flurry of interest in the field of volatility modelling, from both academics and practitioners alike. As a result there now exists a whole host of ARCH- and GARCH-type models, each of which is designed to improve or capture some additional features. In this section we shall illustrate the development of one of the more popular extensions, exponential GARCH. To motivate our discussion we highlight a drawback that is common to both ARCH and GARCH.

The symmetry of ARCH and GARCH

In order to construct a model that delivers a non-negative value for volatility, the ARCH and GARCH frameworks both propose that conditional variance be dependent upon a non-negative linear combination of the past squared losses. We recall that the daily loss process is given by

$$\mathcal{L}_t = \sigma_t \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim N(0, 1) \quad \text{independently drawn}$$

and, based on this process, we point out the following obvious fact:

$$\varepsilon_t > 0 \Rightarrow \text{asset price is falling}$$
$$\text{and } \varepsilon_t < 0 \Rightarrow \text{asset price is rising.}$$

In the real world volatility tends to be higher in a falling market than a rising market, i.e., we expect that a large loss will create significantly more volatility than a large profit of the same magnitude; in financial jargon, this phenomenon is commonly called the leverage effect. Unfortunately, neither ARCH nor GARCH models capture this effect; they are both driven by past squared losses and so they do not distinguish between a positive and a negative shock. In the early 1990s several researchers attempted to modify the original GARCH model in order to more accurately capture the leverage effect. One of the more popular extensions is due to Nelson who, in 1991, observed that another way of ensuring a model delivers a

non-negative volatility estimate is to take a logarithmic approach, specifically he considered a model of the form (Nelson, 1991)

$$\mathcal{L}_t = \sigma_t \varepsilon_t \quad \text{where} \quad \varepsilon_t \sim N(0,1) \quad \text{independently drawn}$$
$$\text{and} \quad \log(\sigma_t) = \omega + \alpha g(\varepsilon_{t-1}) + \beta \log(\sigma_{t-1}). \tag{20.37}$$

Here g is a function of the innovation sequence $(\varepsilon_t)_{t\geq 0}$ which is designed to capture the direction of the underlying asset price. In light of the leverage effect it is clear that g should be constructed so that it contributes more to the volatility estimate when the innovations are positive. Nelson proposed that g be given by

$$g(\varepsilon_t) = a\varepsilon_t + b\left(|\varepsilon_t| - \mathbb{E}[|\varepsilon_t|]\right).$$

We have assumed that our innovation sequence has the standard normal distribution and, in this setting, one can show that $\mathbb{E}[|\varepsilon_t|] = \sqrt{2/\pi}$, and as a result Nelson's GARCH extension, which is commonly known as Exponential GARCH, takes the form

$$\log(\sigma_t) = \alpha_0 + \alpha_1(\varepsilon_t + \lambda|\varepsilon_t|) + \beta_1 \log(\sigma_{t-1}). \tag{20.38}$$

This model has proven to be popular in practice and this is partly due to the fact that, unlike ARCH and GARCH, there is no need to impose any restrictions upon the parameters; the non-negative volatility is found by simply taking the exponential of (20.38). The model captures the volatility clustering phenomenon and, in addition, the leverage effect is accounted for because one tends to find that the estimates for α_1 and λ will be positive numbers.

There are several other GARCH extensions that have been designed to account for the leverage effect. We mention the following commonly cited examples:

- Asymmetric GARCH.
  In 1993 Engle and Ng argued that the lack of symmetry could be fixed by introducing a new shift parameter $\lambda > 0$ and proposed the following modified model:

$$\sigma_t^2 = \omega + \alpha(\mathcal{L}_{t-1} + \lambda)^2 + \beta\sigma_{t-1}^2.$$

- GJR-GARCH.
  Another candidate GARCH model was also introduced in 1993, by Glosten, Jagannathan and Runkle. This contribution, which is named after its founders, also extends the GARCH(1, 1) model by introducing a new term, with a new parameter $\gamma > 0$, whose influence is only felt if a loss occurs. Specifically, the GJR model proposes that volatility be modelled according to

$$\sigma_t^2 = \omega + \alpha\mathcal{L}_{t-1}^2 + \beta\sigma_{t-1}^2 + \gamma \max(\mathcal{L}_{t-1}, 0).$$

The success of the basic GARCH volatility model has led to a profusion of GARCH-type models; each one arising from an attempt to capture the stylized facts more accurately and efficiently. The reader who wants to discover more is advised to consult Alexander (2008a).

Essential Mathematics for Market Risk Management
by Simon Hubbert

21 Extreme Value Theory

In this chapter we change our emphasis from day-to-day risk calculations and instead we consider how to cope with those rare occasions when catastrophic events hit the financial markets. Unfortunately, the standard risk measures, e.g., VaR and TVaR, are not well suited to this scenario as their values are derived from the whole of the loss distribution. What we need is an alternative approach that makes the most of the probabilistic information contained in the extreme tail of the distribution. This chapter is devoted to the mathematical exploration of extreme values and the development of a VaR-type risk measure which accounts for the danger of suffering an extreme loss.

21.1 THE MATHEMATICS OF EXTREME EVENTS

The mathematics of extreme events, commonly called extreme value theory, is a well-established branch of probability theory which has found applications in many areas of the natural sciences. For instance, in low-lying countries such as the Netherlands a fundamental problem is to determine the height of a dam so as to prevent a potential flood resulting from a period of extreme rainfall. An application of extreme value theory can be employed here to help solve this problem.

A relatively recent development has seen extreme value theory employed to solve financial risk problems. In analogy with the rainfall problem we can equate the height of the dam to the size of a buffer fund of risk capital; the monetary amount set aside to help absorb a potential extreme loss resulting from a rare financial storm.

To motivate the mathematical development we recall the central limit theorem (11.35), which deals with the central part of a probability distribution.

The Central Limit Theorem Revisited

- Let $(X_k)_{k=1}^{\infty}$ denote an infinite sequence of independent random variables each possessing the same distribution function F.
- We now introduce two sequences

$$\begin{aligned} &\text{a shift sequence:} && (a_n)_{n=1}^{\infty} && \text{where} && a_n \in \mathbb{R} \\ &\text{and a scaling sequence:} && (b_n)_{n=1}^{\infty} && \text{where} && b_n > 0 \end{aligned} \tag{21.1}$$

and consider the following new sequence of shifted and scaled averages by

$$A_n = \left(\frac{\frac{1}{n}\sum_{k=1}^{n} X_k - a_n}{b_n} \right) \quad n = 1, 2, \ldots$$

We then investigate the properties of the limit of this sequence, i.e., we ask

$$\text{what can we say about } \lim_{n\to\infty} A_n?$$

- The central limit theorem tells us that if

$$\mu = \mathbb{E}[X_k] \quad \text{and} \quad \sigma^2 = \mathbb{E}[(X_k - \mu)^2] < \infty \quad k = 1, 2, \ldots$$

and we let

$$a_n = \mu \quad \text{and} \quad b_n = \frac{\sigma}{\sqrt{n}} \quad n = 1, 2, \ldots$$

then

$$\lim_{n\to\infty} A_n \sim N(0, 1).$$

In the risk management framework we are dealing with sequences of loss random variables $(\mathcal{L}_k)_{k\geq 0}$. We follow the approach of the central limit theorem, i.e., we assume that the losses are independently drawn from the same distribution F; the precise form of which is not necessarily known. However, instead of focusing on averages we turn attention to extremes and define the following sequence of maxima:

$$\mathcal{L}_n^{\max} = \max\{\mathcal{L}_0, \ldots, \mathcal{L}_n\} \quad n = 1, 2, \ldots$$

Our aim is then to investigate whether there is an analogue of the central limit theorem for extreme values, i.e., we shall attempt to answer the following question.

The Extreme Value Problem

Let $(\mathcal{L}_k)_{k\geq 1}$ denote an infinite sequence of loss random variables that are independently drawn from the same distribution F. Does there exist a shift sequence $(a_n)_{n=1}^{\infty}$ (where each $a_n \in \mathbb{R}$) and a scaling sequence $(b_n)_{n=1}^{\infty}$ (where each $b_n > 0$) such that the new sequence of shifted and scaled maxima

$$M_n = \frac{\mathcal{L}_n^{\max} - a_n}{b_n}$$

converges to a random variable that has a recognizable distribution?

Before we make a start on this problem we shall define x_F to be the right endpoint of the underlying distribution F, i.e.,

$$x_F = \sup\{x \in \mathbb{R} : F(x) < 1\}. \tag{21.2}$$

We think of x_F as the point at which all of the probability weight of the distribution has accumulated, i.e.

$$x \geq x_F \Rightarrow F(x) = 1 \quad \text{and} \quad x < x_F \Rightarrow 0 \leq F(x) < 1. \tag{21.3}$$

21.1.1 A naive attempt

Let us suppose that we do not shift and scale the sequence of maxima, i.e., we simply examine the limit of $\mathcal{L}_n^{\max}$ as $n \to \infty$. In this case we can deduce that, for any given n, the distribution of $\mathcal{L}_n^{\max}$, denoted by $F_n^{\max}$, is given by

$$\begin{aligned} F_n^{\max}(x) &= \mathbb{P}[\mathcal{L}_n^{\max} \leq x] \\ &= \mathbb{P}[\max\{\mathcal{L}_1, \ldots, \mathcal{L}_n\} \leq x] \\ &= \mathbb{P}[\mathcal{L}_1 \leq x] \cdots \mathbb{P}[\mathcal{L}_n \leq x] \quad \text{(by independence)} \\ &= F^n(x). \end{aligned}$$

This leads us to conclude that

$$\lim_{n\to\infty} F_n^{\max}(x) = \begin{cases} 0 & \text{if} \quad x < x_F \\ 1 & \text{if} \quad x \geq x_F \end{cases}$$

and we say that $F_n^{\max}$ converges to a degenerate distribution as $n \to \infty$.

The above naive investigation highlights the importance of the shift and scale sequences. Our aim is to identify, if possible, shift and scale sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ such that

$$\mathbb{P}\left[\frac{\mathcal{L}_n^{\max} - a_n}{b_n} < x\right] = F^n(a_n + b_n x) \tag{21.4}$$

converges to a non-degenerate extremal distribution G as $n \to \infty$. It is in no way obvious that such non-degenerate functions exist and so to provide further insight our plan is to examine some concrete examples.

21.1.2 Example 1: Exponentially distributed losses

Assume that the loss random variables are exponentially distributed, i.e.,

$$\mathbb{P}[\mathcal{L}_k \leq x] = F(x) = 1 - \exp(-x) \quad \text{for all} \quad k = 1, 2, \ldots$$

In this case we let

$$a_n = \log n \quad \text{and} \quad b_n = 1 \quad \text{for all } n = 1, 2, \ldots$$

and we find that

$$F^n(\log n + x) = (1 - \exp(-\log n - x))^n = \left(1 - \frac{e^{-x}}{n}\right)^n.$$

We can now evoke the following useful limit result (which actually serves as a definition of the exponential function):

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = \exp(x), \quad \text{for } x \in \mathbb{R} \tag{21.5}$$

to deduce that

$$\lim_{n\to\infty} F^n(\log n + x) \to G(x) = \exp(-e^{-x}).$$

21.1.3 Example 2: Normally distributed losses

In this case we assume that the loss random variables have the standard normal distribution, i.e.,

$$\mathbb{P}[\mathcal{L}_k \le x] = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du \quad \text{for all} \quad k = 1, 2, \ldots$$

In order to make our task a little easier it is useful to introduce the so-called error function, a special function from mathematical physics, defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-\frac{u^2}{2}\right) du. \tag{21.6}$$

Using the error function we can express the standard normal distribution function in a more compact form, as

$$\begin{aligned}\Phi(x) &= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right) \\ &= 1 - \frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right).\end{aligned} \tag{21.7}$$

The error function arises in many real-world problems and consequently its properties are well understood, indeed the seventh chapter of Abramowitz and Stegun (1964) is devoted to its properties. Since our task is to investigate the limiting behaviour of the distribution then, in view of (21.7), the following asymptotic property is relevant:

$$\frac{1}{2}(1 - \operatorname{erf}(x)) \sim \frac{1}{\sqrt{4\pi}} \frac{\exp\left(x^2\right)}{x} \text{ as } x \to \infty. \tag{21.8}$$

Returning to the task in hand, we let

$$a_n = \sqrt{2\log n} - \frac{1}{2}\left(\frac{\log(4\pi \log n)}{\sqrt{2\log n}}\right) \quad \text{and} \quad b_n = \frac{1}{\sqrt{2\log n}}$$

for $n = 1, 2, \ldots$, and we see that

$$\Phi^n(a_n + b_n x) = \left[1 - \frac{1}{2}\left(1 - \operatorname{erf}\left(\frac{4\log n - \log(4\pi \log n) + 2x}{4\sqrt{\log n}}\right)\right)\right]^n.$$

We note that the argument of the error function grows to infinity as $n \to \infty$ and so we can employ (21.8) to deduce that, as n grows large, the term in the square bracket can be written as

$$1 - \frac{1}{\sqrt{4\pi}} \frac{\exp\left(-\left(x + \log n - 1/2 \log(4\pi \log n) + \varepsilon_n(x)\right)\right)}{\sqrt{\log n} + \delta_n(x)}$$
$$= 1 - \frac{\sqrt{\log n}}{n} \frac{\exp(-(x + \varepsilon_n(x)))}{(\sqrt{\log n} + \delta_n(x))},$$

where $\varepsilon_n(x)$ and $\delta_n(x)$ are sequences that both $\to 0$ as $n \to \infty$. Taking the limit in stages we finally discover

$$\lim_{n\to\infty} \Phi^n(a_n + b_n x) = \lim_{n\to\infty} \left[1 - \frac{\sqrt{\log n}}{n} \frac{\exp(-(x + \varepsilon_n(x)))}{(\sqrt{\log n} + \delta_n(x))} \right]^n$$
$$= \lim_{n\to\infty} \left(1 - \frac{e^{-x}}{n} \right) = \exp(-e^{-x}).$$

21.1.4 Example 3: Pareto distributed losses

Here we assume that the loss random variables have the Pareto distribution, i.e., for positive constants K and α this means

$$\mathbb{P}[\mathcal{L}_k \le x] = F(x) = \begin{cases} 1 - Kx^{-\alpha} & \text{if } x > K^{1/\alpha} \\ 0 & \text{otherwise} \end{cases} \qquad \text{for all } k = 1, 2, \ldots$$

In this case we let

$$a_n = 0 \quad \text{and} \quad b_n = (Kn)^{1/\alpha} \quad \text{for all } n = 1, 2, \ldots$$

and find that

$$F^n\left((Kn)^{1/\alpha} x\right) = \begin{cases} \left(1 - \frac{x^{-\alpha}}{n}\right)^n & \text{if } x > 1/n^{1/\alpha} \\ 0 & \text{otherwise.} \end{cases}$$

Now, applying (21.5) as before, we deduce that

$$F^n\left((Kn)^{1/\alpha} x\right) \to G(x) = \begin{cases} \exp(-x^{-\alpha}) & \text{if } x \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

21.1.5 Example 4: Uniformly distributed losses

Here we make the theoretical assumption that the loss random variables are uniformly distributed on [0,1] and write $\mathcal{L}_k \sim U[0, 1]$. We remark that, from a practical point of view,

this assumption is nonsensical because it states that each loss lies in the interval [0,1] and all realizations are equally likely to occur. In any case, the uniform distribution function is given by

$$\mathbb{P}[\mathcal{L}_k \leq x] = F(x) = \begin{cases} 1 & \text{if } x \geq 1 \\ x & \text{if } x \in [0,1] \\ 0 & \text{if } x \leq 0 \end{cases} \quad \text{for all } k = 1, 2, \ldots$$

Here we set

$$a_n = 1 \quad \text{and} \quad b_n = \frac{1}{n} \quad \text{for all } n = 1, 2, \ldots$$

and find that

$$F^n\left(1 + \frac{x}{n}\right) = \begin{cases} 1 & \text{if } x \geq 0 \\ \left(1 + \frac{x^n}{n}\right)^n & \text{if } x \in [-n, 0] \\ 0 & \text{if } x \leq -n. \end{cases}$$

An application of (21.5) allows us to deduce that

$$F^n\left(1 + \frac{x}{n}\right) \to G(x) = \begin{cases} \exp(x) & \text{if } x \leq 0 \\ 1 & \text{otherwise.} \end{cases}$$

21.1.6 Example 5: Cauchy distributed losses

In this final example we shall assume that the loss random variables have the Cauchy distribution. A Cauchy distributed random variable is, in fact, simply a t-distributed random variable with one degree of freedom, and so using (12.15) we can deduce that its density is given by

$$p(x) = \frac{1}{\pi(1 + x^2)} \quad (\text{set } n = 1 \text{ in } (12.15)) \tag{21.9}$$

and consequently its distribution function is

$$F(x) = \mathbb{P}[\mathcal{L}_k \leq x] = \int_{-\infty}^{x} \frac{1}{\pi(1 + u^2)} du.$$

This integral can be solved by making the following substitution:

$$\text{let } u = \tan\theta \quad \text{then } 1 + u^2 = 1 + \tan^2\theta = \sec^2\theta$$
$$\text{and } du = \sec^2\theta d\theta,$$

and this allows us to derive the following closed-form expression:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\pi(1 + u^2)} du$$

$$= \frac{1}{\pi} \int_{-\pi/2}^{\tan^{-1} x} \frac{\sec^2 \theta d\theta}{\sec^2 \theta} = \frac{1}{\pi} \tan^{-1} x + \frac{1}{2}.$$

In this case we let

$$a_n = 0 \quad \text{and} \quad b_n = \frac{n}{\pi} \quad \text{for all} \quad n = 1, 2, \ldots$$

and we find that

$$F^n \left(\frac{nx}{\pi} \right) = \left(\frac{1}{2} + \frac{1}{\pi} \tan^{-1} \left(\frac{nx}{\pi} \right) \right)^n.$$

We now employ the trigonometric identity

$$\frac{1}{2} + \frac{1}{\pi} \tan^{-1}(x) = \begin{cases} 1 - \frac{1}{\pi} \tan^{-1} \left(\frac{1}{x} \right) & \text{if } x > 0 \\ -\frac{1}{\pi} \tan^{-1} \left(\frac{1}{x} \right) & \text{otherwise} \end{cases}$$

followed by the series expansion for the inverse tangent

$$\tan^{-1}(x) = x - \frac{x^3}{3} + \frac{x^5}{5} - \frac{x^7}{7} + \cdots$$

to deduce that

$$F^n \left(\frac{nx}{\pi} \right) = \begin{cases} \left(1 - \frac{x^{-1}}{n} + \varepsilon_n \right)^n & \text{if } x > 0 \\ -\frac{x^{-1}}{n} + \varepsilon_n & \text{if } x < 0 \end{cases}$$

where the sequence $\varepsilon_n \to 0$ as $n \to \infty$. Once again, we can call upon (21.5) to deduce that

$$F^n \left(\frac{nx}{\pi} \right) \to G(x) = \begin{cases} \exp(-x^{-1}) & \text{if } x > 0 \\ 0 & \text{if } x < 0. \end{cases}$$

21.1.7 The extreme value theorem

The examples above lead us to conjecture that there are only three types of non-degenerate extreme distributions, namely

$$\begin{aligned} &\text{(Fréchet) } \Phi_\alpha(x) = \begin{cases} 0 & \text{if } x \le 0 \\ \exp(-x^{-\alpha}) & \text{if } x > 0 \end{cases} \quad \text{for } \alpha > 0; \\ &\text{(Gumbel) } \Lambda(x) = \exp(-e^{-x}), \quad x \in \mathbb{R}; \\ &\text{(Weibull) } \Psi_\alpha(x) = \begin{cases} \exp(-(-x)^\alpha) & \text{if } x < 0 \\ 1 & \text{if } x \ge 0 \end{cases} \quad \text{for } \alpha > 0. \end{aligned} \tag{21.10}$$

This conjecture is indeed true and its discovery, by Fisher and Tippet (1928), dates back to the late 1920s. Their famous theorem is as follows:

Fisher–Tippet Theorem 21.1. *Let* $(\mathcal{L}_k)_{k\geq 0}$ *denote an infinite sequence of independent random variables each having the same distribution function* F. *Let*

$$\mathcal{L}_n^{\max} = \max\{\mathcal{L}_1, \ldots, \mathcal{L}_n\} \quad \text{for} \quad n = 1, 2, \ldots$$

denote the sequence of maxima. If there exists a shift sequence $\{a_n \in \mathbb{R} : n = 1, 2, \ldots\}$ *and a scaling sequence* $\{b_n > 0 : n = 1, 2, \ldots\}$ *such that*

$$F^n(a_n + b_n x) \to G(x) \quad \text{as} \quad n \to \infty,$$

where G *is a non-degenerate distribution, then* G *is either the Fréchet, the Gumbel or the Weibull distribution; see* (21.10).

21.2 DOMAINS OF ATTRACTION

In an attempt to capture all three distributions in one all-encompassing formula, we introduce a new parameter $\xi \in \mathbb{R}$ and consider the function

$$H_\xi(x) = \exp\left[-(1+\xi x)^{-\frac{1}{\xi}}\right] \quad \text{for} \quad 1 + \xi x \geq 0. \tag{21.11}$$

We make the following observations:

(i) If $\xi > 0$ then H_ξ coincides with the Fréchet distribution with $\alpha = 1/\xi$ since

$$\xi > 0 \Rightarrow \Phi_{\frac{1}{\xi}}(x) = H_\xi\left(\frac{x-1}{\xi}\right).$$

(ii) If $\xi < 0$ then H_ξ coincides with the Weibull distribution again with $\alpha = 1/\xi$ since

$$\xi < 0 \Rightarrow \Psi_{\frac{1}{\xi}}(x) = H_\xi\left(-\frac{x+1}{\xi}\right).$$

(iii) The case where $\xi \to 0$ captures the remaining Gumbel distribution since

$$(1+\xi x)^{-\frac{1}{\xi}} = \frac{1}{(1+\xi x)^{\frac{1}{\xi}}} \to \frac{1}{e^x} = e^{-x} \quad \text{as} \quad \xi \to 0,$$

thus we conclude that

$$\lim_{\xi\to 0} H_\xi(x) = \lim_{\xi\to 0} \exp\left[-(1+\xi x)^{-\frac{1}{\xi}}\right] = \exp(-e^{-x}).$$

In conclusion, the function H_ξ (21.11) is successful in capturing all possible extreme distributions; we say H_ξ is the standard generalized extremal distribution with shape

parameter ξ. In more general terms we can also allow for a shift and scale of the distribution and we provide the following definition:

Definition 21.2. *The function*

$$H_{\xi,\mu,\beta}(x) = \exp\left[-\left(1+\xi\left(\frac{x-\mu}{\beta}\right)\right)^{-\frac{1}{\xi}}\right] \text{ for } 1+\xi\left(\frac{x-\mu}{\beta}\right) \geq 0 \tag{21.12}$$

is called the generalized extremal distribution with shape parameter $\xi \in \mathbb{R}$, location parameter $\mu \in \mathbb{R}$ and scale parameter $\beta > 0$.

Given that there are only three possible extremal distributions, a natural question to ask is:

Is it possible to pin down distributional properties of F which ensure that $F^n(a_n + b_n x)$ converges to a particular extremal distribution?

In order to tackle this problem we introduce the notion of an extremal domain of attraction. For instance, we write $F \in \mathcal{DA}(\xi > 0)$ (and say F belongs to the Fréchet domain of attraction) if there exist sequences $(a_n)_{n\geq 0}$ and $(b_n)_{n\geq 0}$ such that $F^n(a_n + b_n x)$ converges to the Fréchet distribution; the domains of attraction for the Gumbell and Weilbull distributions ($\mathcal{DA}(\xi = 0)$ and $\mathcal{DA}(\xi < 0)$) are defined in the same way. With this terminology in place we can conclude, from our earlier worked examples, that:

(a) The Cauchy and the Pareto distributions both belong to $\mathcal{DA}(\xi > 0)$.
(b) The normal and the exponential distributions both belong to $\mathcal{DA}(\xi = 0)$.
(c) The uniform distribution belongs to $\mathcal{DA}(\xi < 0)$.

These findings suggest that it is the tail behaviour of F that determines the corresponding domain of attraction. Indeed, the evidence suggests that:

- Fat-tailed distributions (e.g., Cauchy, Pareto) belong to the Fréchet domain of attraction. Here we think of fat-tailed distributions as those whose endpoint x_F (21.2) is infinite and whose tail decays at a polynomial rate, i.e.,

$$1 - F(x) \approx Cx^{-\alpha} \quad \text{for large } x, \tag{21.13}$$

  where C is a constant and $\alpha > 0$.
- Medium-tailed distributions (e.g., normal, exponential) belong to the Gumbel domain of attraction. Here we think of medium-tailed distributions as those with $x_F = \infty$ and whose tail $1 - F(x)$ decays at an exponentially fast rate as $x \to \infty$.
- Thin-tailed distributions (e.g., uniform) belong to the Weibull domain of attraction. Here we think of thin-tailed distributions as those whose endpoint x_F is finite.

It turns out that the above guidelines serve as a very good rule of thumb for determining which domain of attraction a particular distribution belongs to. Indeed, in the early 1940s Gnedenko published a breakthrough paper in which he carefully constructed precise tail conditions needed for a distribution F to belong to a particular domain of attraction (Gnedenko, 1941). The technical machinery needed to develop these conditions is beyond the scope of this book, however the interested reader will find an excellent and accessible account in Section 3.3 of Embrechts, Klüppelberg and Mikosch (1997).

21.2.1 The Fréchet domain of attraction

A typical loss distribution for a financial portfolio, as we know, will exhibit fat tails, or in extreme value terminology, is most likely to belong to the Fréchet domain of attraction. In order to shed more light upon the theoretical properties of such distributions we appeal to Theorem 3.3.7 of Embrechts, Klüppelberg and Mikosch (1997), where it is established that $F \in \mathcal{DA}(\xi > 0)$ if and only if

$$1 - F(x) = \frac{\lambda(x)}{x^{1/\xi}},$$

where λ is a slowly varying function which, formally speaking, is a function that satisfies

$$\lim_{x \to \infty} \frac{\lambda(xt)}{\lambda(t)} = 1 \quad \text{for all } t \in [0, \infty).$$

For our purposes we may think of λ as being approximately constant and so a less rigorous, but more user-friendly, version of the above condition states

$$F \in \mathcal{DA}(\xi > 0) \Leftrightarrow \text{ the tail of } F \text{ decays like } \frac{1}{x^{1/\xi}}.$$

The familiar check for a fat-tailed random variable is to determine whether its kurtosis coefficient is > 3. The criterion for fat-tailedness in the extreme value environment is slightly different as it is linked to the existence (or otherwise) of the so-called upper moments of the distribution, defined by

$$\begin{aligned} \text{upper } k^{\text{th}} \text{ moment } &= \mathbb{E}\left[\max(\mathcal{L}^k, 0)\right] \\ &= \int_0^\infty F'(x) x^k dx \quad k = 0, 1, 2, \ldots \end{aligned}$$

For this integral to exist it is necessary that $\max_{x \in [0,\infty)} F'(x) < \infty$ and that the integrand, $F'(x)x^k$, decays faster than $1/x$ as $x \to \infty$.

Now, if the distribution F belongs to the Fréchet domain of attraction then we have that

$$F'(x)x^k \text{ decays like } \frac{1}{x^{1+\frac{1}{\xi}-k}} \text{ where } \xi > 0.$$

Using this analysis we can deduce that if $\mathcal{L}$ is a loss random variable whose distribution $F \in \mathcal{DA}(\xi > 0)$, then not all of its upper moments exist. Indeed, it follows that

$$\mathbb{E}\left[\max\left(\mathcal{L}^k, 0\right)\right] < \infty \Leftrightarrow 0 < \xi < \frac{1}{k}. \tag{21.14}$$

We now examine the Fréchet distribution itself, which we write as

$$\Phi_\xi(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \exp\left(-x^{-\frac{1}{\xi}}\right) & \text{if } x > 0. \end{cases}$$

To find the associated Fréchet density function, we differentiate to give

$$p_F(x) = \Phi'_\xi(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{1}{\xi} x^{-\left(\frac{1}{\xi}+1\right)} \exp\left(-x^{-\frac{1}{\xi}}\right) & \text{if } x > 0. \end{cases} \tag{21.15}$$

In order to find where p_F achieves its maximum, we differentiate it to find

$$p'_F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ \frac{1}{\xi^2} x^{-2\left(\frac{1}{\xi}+1\right)} \exp\left(-x^{-\frac{1}{\xi}}\right)\left[1 - x^{\frac{1}{\xi}}(1+\xi)\right] & \text{if } x > 0. \end{cases}$$

Setting this equal to zero we find that the maximum occurs at

$$x_{\max} = \left(\frac{1}{1+\xi}\right)^\xi,$$

and the value of this maximum is

$$p_F(x_{\max}) = \exp(-(1+\xi))\frac{(1+\xi)^{1+\xi}}{\xi}.$$

We note that $x_{\max} \to 1$ as $\xi \to 0$. To investigate how the density function behaves at this limit we consider a Taylor expansion of $(1+\xi)^{1+\xi}$ to show that, for small ξ, we have

$$p_F(x_{\max}) = \frac{\exp(-(1+\xi))}{\xi}\left[1 + (1+\xi)\xi + O(\xi^3)\right]$$

and thus, from this, we can see that

$$p_F(x_{\max}) \to \infty \quad \text{as} \quad \xi \to 0,$$

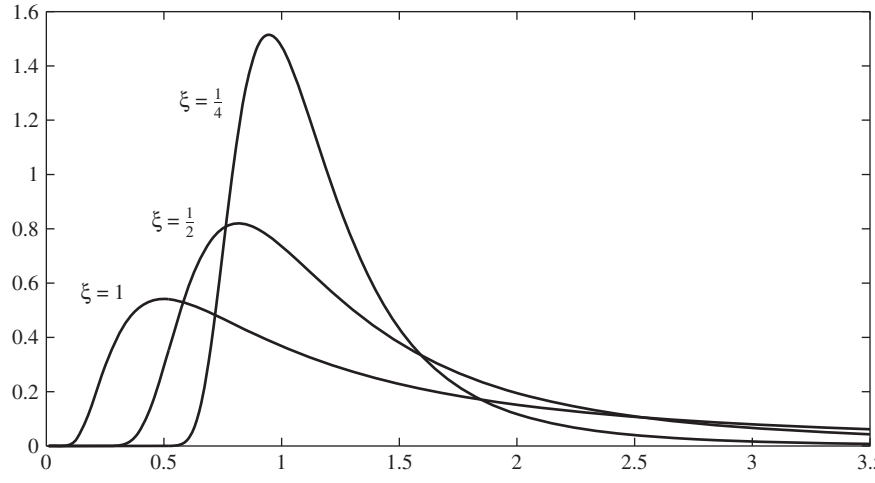


Figure 21.1 Plots of the Fréchet density function.

i.e., as ξ approaches zero the Fréchet density function resembles an infinite spike at $x = 1$; we can see this tendency in Figure 21.1.

To investigate the weight in the tail of the Fréchet distribution we can use (21.15) to explicitly compute its moments; they are given by

$$\mu_k = \int_0^\infty x^k \exp\left(-x^{-\frac{1}{\xi}}\right) \frac{x^{-(\frac{1}{\xi}+1)}}{\xi} dx \quad k = 1, 2, \ldots$$

We now make the substitution

$$u = x^{-\frac{1}{\xi}} \Rightarrow du = -\frac{x^{-(\frac{1}{\xi}+1)}}{\xi}$$

$$\text{and } \Rightarrow x^k = u^{-\xi k}$$

and, as a result, we have that

$$\mu_k = \int_0^\infty u^{-k\xi} \exp(-u) du \quad k = 1, 2, \ldots$$

We now appeal to (12.1) and observe that, provided $1 - k\xi > 0$, we have

$$\mu_k = \int_0^\infty u^{(1-k\xi)-1} \exp(-u) du = \Gamma(1 - k\xi).$$

In the case where $k \geq 1/\xi$ the above integral is not well defined and so we conclude that the Fréchet moments of order $k \geq 1/\xi$ do not exist. We remark that, in view of (21.14), this

indicates that the underlying distribution $F \in \mathcal{DA}(\xi > 0)$ and its limiting Fréchet distribution $\Phi_{1/\xi}$ both possess the same number of moments μ_k, namely those for which $0 < 1/\xi < k$, and, for this reason, the parameter ξ is commonly referred to as the shape or tail index of F.

21.3 EXTREME VALUE AT RISK

Now that we have established a theoretical framework for extremes, the next task is to demonstrate how it can be used to solve practical problems. To motivate a discussion we consider the following scenario:

- The risk management unit of a financial institution decides upon a large monetary loss, u say. We think of u as representing a huge loss the like of which is only experienced on very rare occasions; e.g., once every 5–10 years.
- The concern of the risk manager is then attached to any occasion when the loss random variable $\mathcal{L}$ exceeds u. As a result, the new focus of attention is on modelling the conditional distribution of $Y = \mathcal{L} - u$ given that the loss exceeds u. We will return to the task of selecting an appropriate value for u later in this chapter.
- In mathematical terms, if F denotes the distribution of $\mathcal{L}$, then the new task is to investigate

$$F_u(y) = \frac{F(y+u) - F(u)}{1 - F(u)} \quad 0 \le y \le x_F - u. \tag{21.16}$$

As a first step we compute the so-called mean excess function, i.e., the mean of this conditional distribution as a function of u:

$$\mathrm{ME}(u) = \int_0^{x_F - u} y F_u'(y) dy = \int_0^{x_F - u} \frac{y F'(y+u)}{1 - F(u)} dy.$$

Now, setting $x = y + u$ we have that

$$\mathrm{ME}(u) = \frac{1}{1 - F(u)} \int_u^{x_F} (x - u) F'(x) dx.$$

This integral can be developed further by observing that

$$\frac{d}{dx} \left((x - u) \cdot F(x) \right) = (x - u) F'(x) - F(x)$$

and so we can deduce

$$\begin{aligned} \mathrm{ME}(u) &= \frac{1}{1 - F(u)} \left[\int_u^{x_F} \frac{d}{dx} \left((x - u) \cdot F(x) \right) dx - \int_u^{x_F} F(x) dx \right] \\ &= \frac{1}{1 - F(u)} \left[x_F - u - \int_u^{x_F} F(x) dx \right] \end{aligned} \tag{21.17}$$

$$= \frac{1}{1 - F(u)} \int_u^{x_F} (1 - F(x))\, dx.$$

We remark that the mean excess function plays an important role in practice. Indeed, as we shall discover later in this chapter, it can be used as a useful diagnostic tool to enable the implementation of practical solutions based upon theoretical findings. In order to make further progress in our investigation we now equip ourselves with one final theoretical result which was established independently in the mid-1970s by Pickands (1975) and Balkema and de Hahn (1974). The result itself provides insight into how the conditional distribution F_u behaves as the threshold parameter grows large.

Theorem 21.3. *Let F denote the distribution function of a loss random variable $\mathcal{L}$. If $F \in \mathcal{DA}(\xi > 0)$ then there exists a scale parameter $\beta > 0$ such that, for large u, we have*

$$F_u(y) \approx G_{\xi,\beta}(y) = 1 - \left(1 + \frac{\xi}{\beta} y\right)^{-\frac{1}{\xi}}, \quad y \geq 0. \tag{21.18}$$

Armed with this theorem we now revisit the problem of computing portfolio VaR at a confidence level α. We recall that, in its simplest form, the Value at Risk is the number VaR_α that satisfies the equation $F(\text{VaR}_\alpha) = \alpha$ where F denotes the underlying loss distribution. Our new approach to discovering an estimate of VaR_α can now be achieved in three steps.

- Step 1. A new formula for $F(x)$.
  Here we simply set $y = x - u$ in expression (21.16) and discover, after rearrangement, that

$$F(x) = (1 - F(u))F_u(x - u) + F(u). \tag{21.19}$$

- Step 2. A statistical component.

Here we assume that we have access to a long history of realized losses, i.e., if we let t denote today's date then we assume we have the past N days' losses $\{\mathcal{L}_{t-\tau} : \tau = 1, \ldots, N\}$. For a given threshold u we define the excess indicator function by

$$I_u(t - \tau) = \begin{cases} 0 & \text{if } \mathcal{L}_{t-\tau} \leq u \\ 1 & \text{if } \mathcal{L}_{t-\tau} > u. \end{cases}$$

Thus, for a fixed u the value

$$n_u = \sum_{\tau=1}^{N} I_u(t - \tau)$$

counts the number of times, over the past N days, that the loss exceeded the threshold u. Thus, using the data we can estimate the true distribution F with the empirical version F_N (see Chapter 22 for more details) by making the assumption that

$$F(u) = \mathbb{P}[\mathcal{L}_t \leq u] \approx \frac{N - n_u}{N} = F_N(u). \tag{21.20}$$

We observe that the empirical distribution can be used in formula (21.19) to approximate the factors $F(u)$ and $1 - F(u)$.

- Step 3. A theoretical component.
  Here we argue that if the threshold value is suitably large then we can evoke Theorem 21.3 to yield the approximation

$$F_u(x-u) \approx G_{\xi,\beta}(x-u) \tag{21.21}$$

and thus, together with the empirical distribution, we unearth the following approximation:

$$F(x) \approx \widehat{F}(x) = \frac{n_u}{N} G_{\xi,\beta}(x-u) + 1 - \frac{n_u}{N}. \tag{21.22}$$

We can now use (21.22) to provide new estimates for VaR_α. Specifically, we seek the number $\text{VaR}_{\alpha,\xi}$ that is the unique solution to

$$\widehat{F}(\text{VaR}_{\alpha,\xi}) = \alpha.$$

To achieve this we substitute the expression (21.18) in (21.22) to yield

$$\begin{aligned}\widehat{F}(x) &= \frac{n_u}{N}\left(1 - \left(1 + \frac{\xi}{\beta}(x-u)\right)^{-\frac{1}{\xi}}\right) + 1 - \frac{n_u}{N} \\ &= 1 - \frac{n_u}{N}\left(1 + \frac{\xi}{\beta}(x-u)\right)^{-\frac{1}{\xi}}.\end{aligned} \tag{21.23}$$

Now, solving

$$\widehat{F}(\text{VaR}_{\alpha,\xi}) = \alpha \Rightarrow 1 - \frac{n_u}{N}\left(1 + \frac{\xi}{\beta}(\text{VaR}_{\alpha,\xi} - u)\right)^{-\frac{1}{\xi}} = \alpha$$

we find that

$$\text{VaR}_{\alpha,\xi} = u + \frac{\beta}{\xi}\left(\left[\frac{N}{n_u}(1-\alpha)\right]^{-\xi} - 1\right). \tag{21.24}$$

In addition, we can use this expression in (9.12) to calculate the corresponding Tail Value at Risk:

$$\begin{aligned}\text{TVaR}_{\alpha,\xi} &= \frac{1}{1-\alpha}\int_\alpha^1 \left(u + \frac{\beta}{\xi}\left(\left[\frac{N}{n_u}(1-x)\right]^{-\xi} - 1\right)\right) dx \\ &= u + \frac{\beta}{\xi}\left(\frac{N}{n_u}\right)^{-\xi} \frac{1}{1-\alpha}\int_\alpha^1 (1-x)^{-\xi} dx - \frac{\beta}{\xi} \\ &= u + \frac{\beta}{\xi}\left(\frac{N}{n_u}\right)^{-\xi} \frac{(1-\alpha)^{-\xi}}{1-\xi} - \frac{\beta}{\xi}\end{aligned}$$

$$= \frac{1}{1-\xi}\left[u(1-\xi) + \frac{\beta}{\xi}\left(\frac{N}{n_u}(1-\alpha)\right)^{-\xi} - \frac{\beta}{\xi}(1-\xi)\right]$$

$$= \frac{1}{1-\xi}\left[\underbrace{u + \frac{\beta}{\xi}\left(\left[\frac{N}{n_u}(1-\alpha)\right]^{-\xi} - 1\right)}_{=\mathrm{VaR}_{\alpha,\xi}} + \beta - u\xi\right].$$

Thus, we can conclude that

$$\mathrm{TVaR}_{\alpha,\xi} = \frac{\mathrm{VaR}_{\alpha,\xi}}{1-\xi} + \frac{\beta - u\xi}{1-\xi}. \tag{21.25}$$

21.4 PRACTICAL ISSUES

As we follow the development leading to the VaR and TVaR estimates (21.24) and (21.25) we notice there are two key practical issues that we must address:

- How do we estimate the parameters β and ξ that appear in the resulting expressions?
- How should we select the threshold value u?

21.4.1 Parameter estimation

We begin by presenting a simple maximum likelihood approach that we can use to estimate the size of the shape parameter ξ. The method we present (which can also be found in Christoffersen (2003), Section 4.5.3) relies on the fact that when $\xi > 0$ we can assume that, for a sufficiently large u, the tail of the underlying loss distribution can be approximated by

$$1 - F(x) \approx cx^{-1/\xi} \quad \text{for} \quad x > u.$$

We can use this to approximate f_u, the conditional density function (on the event that the loss exceeds u):

$$f_u(x) = \frac{F'(x)}{1-F(u)} \approx \frac{\frac{1}{\xi}x^{-1-1/\xi}}{u^{-1/\xi}} = \frac{1}{x\xi}\left(\frac{x}{u}\right)^{-\frac{1}{\xi}}, \quad \text{for} \quad x > u.$$

We now appeal to our historical data and let $\{\widehat{\mathcal{L}}_1, \ldots, \widehat{\mathcal{L}}_{n_u}\}$ denote the n_u realized losses that exceed the threshold value u. Using this data we can set up a maximum likelihood problem for the parameter ξ. Specifically, the likelihood function is given by

$$L_u(\xi) = \prod_{k=1}^{n_u} \frac{1}{\widehat{\mathcal{L}}_k \xi}\left(\frac{\widehat{\mathcal{L}}_k}{u}\right)^{-\frac{1}{\xi}},$$

which in log form reads

$$\log\left(L_u(\xi)\right) = -\sum_{k=1}^{n_u}\left(\log\xi - \left(1+\frac{1}{\xi}\right)\log\widehat{\mathcal{L}}_k + \frac{1}{\xi}\log u\right).$$

To find the optimal value $\widehat{\xi}$ we differentiate with respect to ξ and solve the first-order conditions

$$\frac{d}{d\xi}\log\left(L_u(\widehat{\xi})\right) = -\sum_{k=1}^{n_u}\left(\frac{1}{\widehat{\xi}} + \frac{1}{\widehat{\xi}^2}\log\widehat{\mathcal{L}}_k - \frac{1}{\widehat{\xi}^2}\log u\right) = 0.$$

Multiplying through by $\widehat{\xi}^2$ and rearranging yields

$$\widehat{\xi} = \frac{1}{n_u}\sum_{k=1}^{n_u}\log\widehat{\mathcal{L}}_k - \log u. \tag{21.26}$$

The above estimate is commonly called the Hill estimator for ξ; it is the log average of the n_u extreme observations which exceed the threshold u, minus the observation taken at the threshold u.

A more precise approach is to set up and solve the true maximum likelihood problem, in which case the conditional density function of the excess losses is found by differentiating the generalized Pareto distribution (21.18); doing so gives

$$p_{\xi,\beta}(x) = \frac{1}{\beta}\left(1+\frac{\xi}{\beta}(x-u)\right)^{-(1+\frac{1}{\xi})},$$

The corresponding log-likelihood function is then given by

$$LL_u(\xi,\beta) = n_u\left(1+\frac{1}{\xi}\right)\log\beta + \left(1+\frac{1}{\xi}\right)\sum_{k=1}^{n_u}\log\left(1+\frac{\xi}{\beta}(\widehat{\mathcal{L}}_k - u)\right).$$

This function can be optimized using numerical techniques to provide the maximum likelihood estimates $\widehat{\xi}$ and $\widehat{\beta}$ for the shape and scale parameters respectively.

21.4.2 The choice of threshold

We note that our estimate for extreme VaR is dependent upon the threshold value u that is selected at the outset. One of the key considerations we take when selecting u is that it should be high enough for the approximation (21.21) to be reasonably accurate. We know that the approximation improves as u grows however, unfortunately, u can grow so large that the statistical estimate (21.20) is rendered useless. Thus, there is an uncertainty principle at work: if u is large then approximation (21.21) is accurate but as n_u drops we have less

faith in our statistical parameter estimates. On the other hand, we can lower u in order to achieve better statistical estimates however, at the same time incur a deterioration in our generalized Pareto distribution approximation.

One way of gaining some insight into the validity of approximation (21.18) is to pin down a theoretical property belonging to the generalized Pareto distribution which we can check with real data. It turns out that the mean excess function of $G_{\xi,\beta}$ is very useful in this regard because, as the result below shows, it has a particularly simple form; a straight line with positive slope.

Proposition 21.4. *The mean excess function corresponding to the generalized Pareto distribution $G_{\xi,\beta}$ for $0 < \xi < 1$ is given by*

$$\mathrm{ME}(u : \xi, \beta) = \frac{\beta + \xi u}{1 - \xi} \quad \text{for} \quad \beta + \xi u > 0. \tag{21.27}$$

Proof. The right-hand endpoint of $G_{\xi,\beta}$ is infinite and thus, recalling equation (21.17) and employing representation (21.18), we have

$$\begin{aligned}
\mathrm{ME}(u : \xi, \beta) &= \int_u^\infty \frac{1 - G_{\xi,\beta}(x)}{1 - G_{\xi,\beta}(u)} dx \\
&= \frac{1}{\left(1 + \frac{\xi u}{\beta}\right)^{-\frac{1}{\xi}}} \int_u^\infty \left(1 + \frac{\xi}{\beta} x\right)^{-\frac{1}{\xi}} dx \\
&= \frac{1}{\left(1 + \frac{\xi u}{\beta}\right)^{-\frac{1}{\xi}}} \left[\frac{\beta}{\xi - 1}\left(1 + \frac{\xi x}{\beta}\right)^{1-\frac{1}{\xi}}\right]_u^\infty \\
&= \frac{\beta}{1 - \xi}\left(1 + \frac{\xi u}{\beta}\right) \\
&= \frac{\beta + \xi u}{1 - \xi}.
\end{aligned}$$

□

We can now turn to our loss data and compute estimates of the mean excess function for a range of different threshold values. Specifically, we let $\mathcal{L}_{min} = \min\{\mathcal{L}_{t-\tau} : \tau = 1, \ldots, N\}$, $\mathcal{L}_{max} = \max\{\mathcal{L}_{t-\tau} : \tau = 1, \ldots, N\}$ and, for a given $u \in (\mathcal{L}_{min}, \mathcal{L}_{max})$ we compute the empirical mean excess function defined by

$$ME_N(u) = \frac{\sum_{\tau=1}^N \max(\mathcal{L}_{t-\tau} - u, 0)}{\sum_{\tau=1}^N I_u(t - \tau).}$$

We recall that the theory tells us that the excess loss (over a sufficiently high threshold u) resembles a random variable with the generalized Pareto distribution, which we now know (see Proposition 21.4) has a mean excess function that is linear and upwardly sloping. In view of this we would expect that, if the loss data truly support the model then a simple plot of u against $ME_N(u)$ ought to resemble an upwardly sloping straight line. Indeed, this simple diagnostic check is commonly used by practitioners to help select an appropriate value of u for implementation. Specifically the threshold level should correspond to a region of the empirical mean excess plot which clearly resembles a straight line; for further details and practical examples the reader can consult McNeil, Frey and Embrechts (2005).

Essential Mathematics for Market Risk Management
by Simon Hubbert

22 Simulation Models

Throughout this book we have exposed and developed mathematical and statistical techniques in order to build and improve models of financial risk. In this chapter we add to this story by describing an alternative framework for calculating Value at Risk (and the wider class of spectral risk measures) that is based upon numerical simulation. In this setting our aim is to develop a numerical algorithm which is designed to deliver the so-called empirical distribution, an approximation to the true loss distribution. In general, the algorithm will rely on historical data or, alternatively, a mathematical model that is calibrated to match the behaviour of the loss random variable. Once the empirical distribution is available the VaR measure is then taken to be the appropriate quantile estimate. In view of this we open the chapter with an introduction to the statistical approach to quantile estimation.

22.1 ESTIMATING THE QUANTILE OF A DISTRIBUTION

Let us suppose that $(X_j)_{j=1}^{\infty}$ is a sequence of i.i.d. random variables with a common distribution function F. We shall assume that F is strictly increasing, continuously differentiable (we let $f = dF/dx$ denote its density function) and possesses a continuous inverse. Consequently we know that, for any given $p \in [0, 1]$, there exists a unique number x_p, which we call the p-quantile of the distribution, that satisfies

$$F(x_p) = p. \tag{22.1}$$

In many cases F is unknown and so, for a given $p \in [0, 1]$, the solution to (22.1) needs to be estimated from sample values. This process is straightforward, the estimate is found as follows:

- Take a sample of n observed values of the process $\{x_1, \ldots, x_n\}$ and rearrange into ascending order, we write

$$x^{(1)} < x^{(2)} < \cdots < x^{(n)}.$$

  In statistical terminology we say

$$x^{(k)} \text{ is the realization of the } k\text{th-order statistic } X^{(k)}. \tag{22.2}$$

- Given the sample values we construct the empirical distribution function, defined as

$$F_n(x) = \begin{cases} 0 & \text{if } x \leq x^{(1)}; \\ k/n & \text{if } x \in [x^{(k)}, x^{(k+1)}) \quad \text{for} \quad k = 1, \ldots, n-1; \\ 1 & \text{if } x \geq x^{(n)}. \end{cases}$$

  We note that F_n is a staircase function with n steps each of height $1/n$. The staircase starts at zero, ends at one and the width of step k is $x^{(k+1)} - x^{(k)}$, $k = 1, \ldots, n-1$; see Figure 22.1.

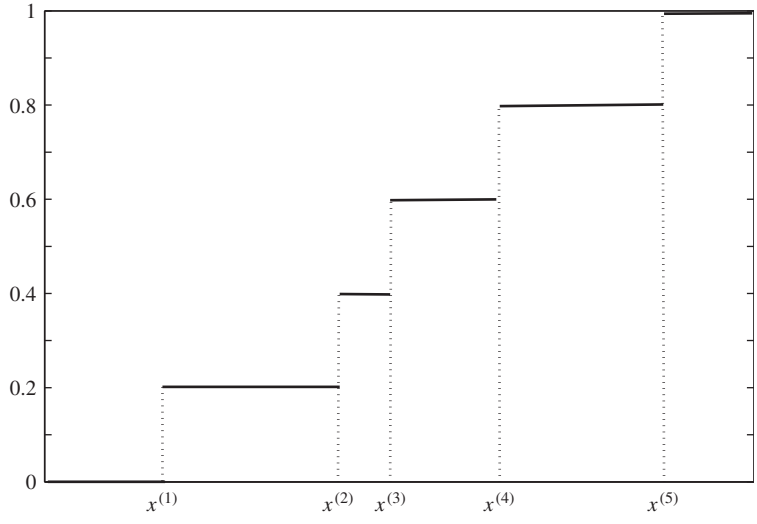


Figure 22.1 An empirical distribution with step size $1/5$.

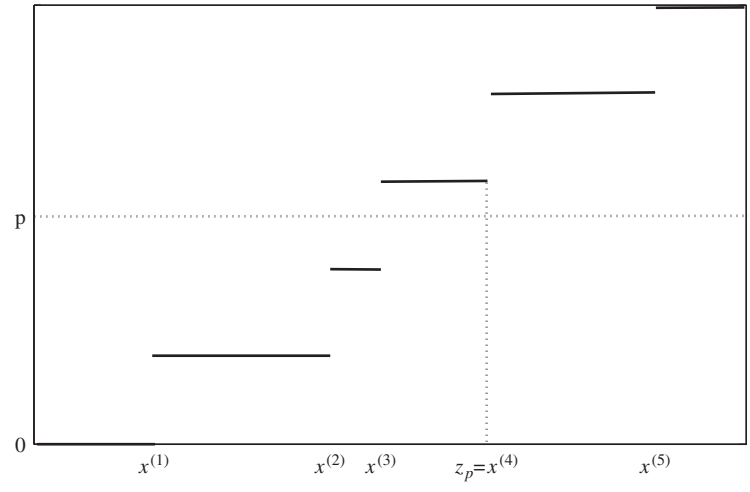


Figure 22.2 To illustrate the calculation of the sample quantile estimate.

- The estimate $z_p(n)$ of the p-quantile is defined to be the value that satisfies $F_n(z) = p$. We can see from Figure 22.2 that, depending upon the level of p, this equation will have either:

  (i) No unique solution; this occurs when np is not an integer. In this case we let

  $$k = [np] = \text{the greatest integer } \leq np, \quad \text{and set } z_p(n) = x^{(k+1)}.$$

  (ii) Or infinitely many solutions; this occurs when Tp is an integer. In this case we let

  $$k = np \quad \text{and set } z_p(n) = x^{(k+1)}.$$

22.1.1 Asymptotic behaviour

As with all sample statistics the quantile estimate $Z = z_p(n)$ is a random variable which depends upon the underlying sample. In order to assess the accuracy we examine the asymptotics of this random quantity, i.e., how it behaves as the sample size grows without bound. To kick-start this investigation we let $g(z)$ denote the density function of Z. Our aim is to derive an explicit expression for this density and we begin by considering the probability element $g(z)dz$ which describes the probability that the outcome of Z is situated in an infinitesimal interval $(z, z + dz)$. Another way of viewing this quantity is to note that it is identical to the probability that, out of the n sample values:

- $k = [np]$ are below z,
- $n - k - 1$ are above $z + dz$,
- and the one remaining value falls between z and $z + dz$.

Using our distributional assumptions we may thus deduce that

$$g(z)dz = \frac{n!}{k!1!(n-k-1)!}(F(z))^k(1-F(z))^{n-k-1}f(z)dz,$$

and so the density function of $Z = z_p(n)$ is given by

$$g(z) = \frac{n!}{k!(n-k-1)!}(F(z))^k(1-F(z))^{n-k-1}f(z).$$

In order to investigate the distribution in more detail we consider a shifted and scaled version of Z given by

$$Y = \sqrt{\frac{n(f(x_p))^2}{pq}}\Big(Z - x_p\Big), \tag{22.3}$$

where $q = 1 - p$ and x_p denotes the true value of the p-quantile, i.e., the unique value that satisfies $F(x_p) = p$. Using (11.13) we can deduce that the density function of Y is given by

$$\begin{aligned} g_{\text{shift}}(y) &= \sqrt{\frac{pq}{n(f(x_p))^2}}\, g\left(\sqrt{\frac{pq}{n(f(x_p))^2}}\, y + x_p\right) \\ &= G_1(n)G_2(y,n)G_3(y,n), \end{aligned}$$

where the three factors are given explicitly by

$$G_1(n) = \sqrt{\frac{pq}{n}}\frac{n!}{k!(n-k-1)!}p^k q^{n-k-1};$$

$$G_2(y,n) = \frac{f\left(\sqrt{\frac{pq}{n(f(x_p))^2}}\, y + x_p\right)}{f(x_p)};$$

$$G_3(y,n) = \left(\frac{F\left(\sqrt{\frac{pq}{n(f(x_p))^2}} y + x_p \right)}{p} \right)^k \left(\frac{1 - F\left(\sqrt{\frac{pq}{n(f(x_p))^2}} y + x_p \right)}{q} \right)^{n-k-1}.$$

We shall now examine the limiting behaviour of each of these factors as $n \to \infty$.

- The limit of $G_1(n)$ as $n \to \infty$.
  We begin by observing that, after a little manipulation, we can express the factor $G_1(n)$ as

$$G_1(n) = \sqrt{\frac{p(1-p)}{n}} \left[\frac{n!}{k!(n-k)!} \right] p^k (1-p)^{n-k} \frac{n-k}{1-p} \tag{22.4}$$

  where we have used the fact that $q = 1 - p$.
  We remark that as n grows large then so does the value $k = [np]$. In view of this we start our analysis by focusing on the term in the square brackets. The following result is Stirling's formula ([1], formula 6.1.37) which describes the asymptotic behaviour of the factorial function:

$$m! \sim \left(\frac{m}{e} \right)^m \sqrt{2\pi m}. \tag{22.5}$$

  Employing this we find that

$$\frac{n!}{k!(n-k)!} \sim \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \frac{n^n}{k^k (n-k)^{n-k}}. \tag{22.6}$$

  For large n we shall assume that $k = np$, in which case we can deduce that

$$\begin{aligned} \frac{n!}{k!(n-k)!} &\sim \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{n^2 p(1-p)}} \frac{n^n}{n^n p^k (1-p)^{n-k}} \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{1}{np(1-p)}} \frac{1}{p^k (1-p)^{n-k}}. \end{aligned}$$

  Plugging this into (22.4) we find that as n grows large we have

$$\begin{aligned} G_1(n) &\sim \frac{1}{\sqrt{2\pi}} \frac{n-k}{n} \frac{1}{1-p} \\ &= \frac{1}{\sqrt{2\pi}} \frac{n(1-p)}{n} \frac{1}{1-p} = \frac{1}{\sqrt{2\pi}}, \end{aligned}$$

  and so we have demonstrated that

$$G_1(n) \to \frac{1}{\sqrt{2\pi}} \quad \text{as } n \to \infty. \tag{22.7}$$

- The limit of $G_2(n)$ as $n \to \infty$.
  This limit can be found as follows:

$$\lim_{n\to\infty} G_2(y,n) = \lim_{n\to\infty} \frac{f\left(\sqrt{\frac{pq}{n(f(x_p))^2}} y + x_p \right)}{f(x_p)} = \frac{f(x_p)}{f(x_p)} = 1. \tag{22.8}$$

- The limit of $G_3(n)$ as $n \to \infty$.
  We shall investigate the behaviour of G_3 by developing a second-order Taylor series approximation of F about x_p, that is we write

$$F(t + x_p) \approx F(x_p) + tf(x_p) + \frac{t^2}{2} f'(x_p) + \text{higher-order terms.}$$

  Setting

$$t = \sqrt{\frac{pq}{n(f(x_p))^2}},$$

  we find that

$$F\left(\sqrt{\frac{pq}{f^2(x_p)n}}y + x_p\right) = p + \sqrt{\frac{pq}{n}}y + \frac{pq}{n}\frac{f'(x_p)}{f^2(x_p)}y^2 + \varepsilon_n,$$

  where $\varepsilon_n \to 0$ as $n \to \infty$. We can use this to show that

$$G_3(y, n) = \frac{\left[\left(1 + \sqrt{\frac{q}{pn}}y + \frac{qf'(x_p)}{n(f(x_p))^2}y^2 + \varepsilon_n\right)^p \left(1 - \sqrt{\frac{p}{qn}}y - \frac{pf'(x_p)}{n(f(x_p))^2}y^2 + \varepsilon_n\right)^q\right]^n}{1 - \sqrt{\frac{p}{qn}}y + \frac{pf'(x_p)}{n(f(x_p))^2}y^2 + \varepsilon_n}.$$

  We note that in the above expression we have used the assumption that for large n we have $k = np$.

We notice immediately that the denominator of the above expression converges to 1 as $n \to \infty$. To investigate the numerator we shall expand the two factors in the square brackets using the binomial theorem, i.e., using

$$(1 + t)^\alpha = 1 + \alpha t + \frac{\alpha(1 - \alpha)}{2}t^2 + \text{higher-order terms.}$$

However, in each case, we shall let the generic ε_n represent all terms that decay faster than a rate of $1/n$. Thus, the expression in the square bracket can be approximated by

$$\begin{aligned} &1 + \sqrt{\frac{pq}{n}}y + \frac{pqf'(x_p)}{n(f(x_p))^2}y^2 + \frac{p(p-1)}{2}\frac{q}{pn}y^2 + \varepsilon_n \\ &= 1 + \sqrt{\frac{pq}{n}}y + \frac{pqf'(x_p)}{n(f(x_p))^2}y^2 - \frac{q^2}{2n}y^2 + \varepsilon_n \end{aligned}$$

multiplied by

$$\begin{aligned} &1 - \sqrt{\frac{qp}{n}}y - \frac{qpf'(x_p)}{n(f(x_p))^2}y^2 + \frac{q(q-1)}{2}\frac{p}{qn}y^2 + \varepsilon_n \\ &= 1 - \sqrt{\frac{qp}{n}}y - \frac{qpf'(x_p)}{n(f(x_p))^2}y^2 - \frac{p^2}{2n}y^2 + \varepsilon_n. \end{aligned}$$

Performing this multiplication we find that

$$G_3(y,n) = \frac{\left[1 - \left(\frac{pq}{n} + \frac{p^2+q^2}{2n}\right) y^2 + \varepsilon_n\right]^n}{1 - \sqrt{\frac{p}{nq}}\, y + \frac{pf'(x_p)}{n(f(x_p))^2} y^2 + \varepsilon_n}$$
$$= \frac{\left[1 - \left(\frac{(p+q)^2}{2n} y^2\right) + \varepsilon_n\right]^n}{1 - \sqrt{\frac{p}{qn}}\, y + \frac{pf'(x_p)}{n(f(x_p))^2} y^2 + \varepsilon_n}$$
$$= \frac{\left[1 - \frac{y^2}{2n} + \varepsilon_n\right]^n}{1 - \sqrt{\frac{p}{qn}}\, y + \frac{pf'(x_p)}{n(f(x_p))^2} y^2 + \varepsilon_n}.$$

Now using (21.5) we find that

$$\lim_{n\to\infty} G_3(y,n) = \exp\left(-\frac{y^2}{2}\right). \tag{22.9}$$

We can now collect our findings (22.9), (22.8) and (22.7) and piece them together to deduce that

$$g_{\text{shift}}(y) = G_1(n)G_2(y,n)G_3(y,n) \to \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \quad \text{as} \quad n \to \infty,$$

i.e., the random variable Y (22.3) converges to a standard normal variable as $n \to \infty$. As a consequence we can deduce that the sample distribution of the p-quantile estimate $Z = z_p(n)$ satisfies

$$z_p(n) \to \mathcal{Z} \sim N\left(x_p, \frac{p(1-p)}{n(f(x_p))^2}\right) \quad \text{as} \quad n \to \infty. \tag{22.10}$$

22.2 HISTORICAL SIMULATION

Our first simulation approach is based upon a very straightforward idea which uses past loss data to construct an empirical distribution. Thus, rather than imposing a theoretical model the data are allowed to speak for themselves. Once we are equipped with the empirical loss distribution we can then deliver a VaR measure by estimating its appropriate quantile. Indeed, a risk manager who wants to monitor risk using an historical simulation framework can follow the recipe we set out here.

- Step 1. Gather data.
  We assume that our portfolio consists of n risky assets whose values at time t are denoted by $S_i(t)$ for $i = 1, \ldots, n$. For each asset we gather its past price history spanning the previous $T + 1$ days, i.e., we extract

$$\{S_i(t-1), S_i(t-2), \ldots, S_i(t-T-1)\} \quad \text{for} \quad i = 1, \ldots, n.$$

Using the price data we then compute the past T realized log losses

$$\{l_i(t-1), l_i(t-2), \ldots, l_i(t-T)\} \quad \text{for} \quad i = 1, \ldots, n,$$

according to the formula

$$l_i(t-\tau) = -\log\left(\frac{S_i(t-\tau)}{S_i(t-\tau-1)}\right) \text{ for } \quad \tau = 1, \ldots, T.$$

- Step 2. Simulate and order potential future losses.
  Given portfolio weights $\{w_1, \ldots, w_n\}$ we use the historical losses we have collected to create the following set of potential future portfolio losses:

$$\{\mathcal{L}_{t-\tau} = V(t-\tau)\sum_{i=1}^{n} w_i l_i(t-\tau)\}_{\tau=1}^{T}.$$

  We then reorder the potential future losses so that they are in ascending order. This calls for a naming convention which we define as

$$\mathcal{L}^{(1)} < \mathcal{L}^{(2)} < \cdots < \mathcal{L}^{(T)}.$$

- Step 3. Derive the Value at Risk estimate.
  Given a confidence level α the corresponding Value at Risk estimate $\text{VaR}_\alpha^{\text{HS}}$ is just the α-quantile, i.e., we let

$$k = [n\alpha] = \text{the greatest integer } \leq n\alpha$$

  and set

$$\text{VaR}_\alpha^{\text{HS}}(n) = \mathcal{L}^{(k+1)}. \tag{22.11}$$

  We remark that if we assume that the losses are independent of each other and that they share the same probability density function f, then, using (22.10), we can deduce that as n grows large then $\text{VaR}_\alpha^{\text{HS}}(n)$, given by (22.11), resembles a normal random variable whose mean is the true VaR_α and whose variance is

$$\sigma^2(\text{VaR}_\alpha^{\text{HS}}(n)) = \frac{1}{f^2(\text{VaR}_\alpha)}\frac{\alpha(1-\alpha)}{n}.$$

  In other words, we can expect the simulation estimate to converge to the true estimate but at a relatively slow rate, i.e.,

$$|\text{VaR}_\alpha^{\text{HS}}(n) - \text{VaR}_\alpha| = O\left(\frac{1}{\sqrt{n}}\right).$$

- Step 4. Compute the Tail Value at Risk.
  We recall that the TVaR for a given portfolio is given by the conditional expectation of $\mathcal{L}_t$, the portfolio loss random variable, i.e.,

$$\begin{aligned}\text{TVaR}_\alpha(\mathcal{L}_t) &= \mathbb{E}[\mathcal{L}_t | \mathcal{L}_t > \text{VaR}_\alpha] \\ &= \frac{1}{1-\alpha}\int_{\text{VaR}_\alpha}^{\infty} x f(x) dx,\end{aligned}$$

where f denotes the probability density function of $\mathcal{L}_t$. The formula for TVaR leads us to think of it as the average of the losses inside the $100(1-\alpha)\%$ tail. In the historical simulation framework we can compute an estimate of this quantity by simply taking the average of the extreme losses, i.e.,

$$\begin{aligned} &\text{we let } k = [n\alpha] \quad \text{and} \quad \mathrm{VaR}_\alpha^{\mathrm{HS}}(n) = \mathcal{L}^{(k+1)} \\ &\text{then define } \mathrm{TVaR}_\alpha^{\mathrm{HS}}(n) = \frac{1}{n-k-1}\sum_{i=k+1}^{n} \mathcal{L}^{(i)}. \end{aligned} \tag{22.12}$$

The historical simulation technique will be found in the risk calculation engine of the majority of financial institutions. It delivers quick and easy estimates of both VaR and TVaR and it is based upon the assumption that the history of the portfolio can be used to provide accurate scenarios for its future path. Despite its obvious allure the historical simulation approach is unlikely to be the definitive risk calculator as it is not without its drawbacks. In order to implement the method, the following non-trivial issues must be addressed.

- Do we have access to historical data for all products?
  A typical trading portfolio is likely to contain a vast range of diverse financial instruments. The task of gathering historical price data for each instrument is a significant undertaking. All values need to be delivered to a central database and, if some are missing, then one must employ a suitable interpolation method to fill the gaps.
- How do we choose the size of the historical database?
  This question presents the same kind of problems that we encountered in modelling conditional volatility, namely:
  - if the historical window is too small then the resulting VaR is likely to be weak;
  - if the window is too large then potentially irrelevant values from the very distant past will have a misleading influence on the VaR estimate.

One attempt to overcome this problem is to take the lead from volatility modelling and impose a declining sequence of weights to the historical values. The exponential weighting scheme is a popular one, this involves choosing a parameter $\lambda \in (0,1)$ and fixing the weights

$$\frac{1-\lambda}{1-\lambda^n} > \frac{(1-\lambda)\lambda}{1-\lambda^n} > \frac{(1-\lambda)\lambda^2}{1-\lambda^n} > \cdots > \frac{(1-\lambda)\lambda^{n-1}}{1-\lambda^n}.$$

We then apply these to yield the following set of weighted losses:

$$\left\{\widetilde{\mathcal{L}_{t-k}} = \frac{(1-\lambda)\lambda^{k-1}}{1-\lambda^n}\mathcal{L}_{t-k} : k = 1,\ldots,n\right\}.$$

The VaR is then calculated on the weighted losses rather than the plain losses. Clearly the results will depend upon the choice of λ and the estimation of this parameter deserves further research; the interested reader can find out more on this approach by consulting Christoffersen (2003).

22.3 MONTE CARLO SIMULATION

The historical simulation approach to risk is mathematically light, it does not rely upon a model for the daily losses and, as a result, it is completely parameter free. The Monte Carlo simulation method is an alternative approach to risk management which is similar in spirit to historical simulation however, rather than relying only on past data, the Monte Carlo method does take advantage of the mathematical loss models that are on the market. The basic recipe for the Monte Carlo simulation method is given as follows:

- Step 1. Choose an appropriate model for the daily loss random variable for the assets which make up the portfolio.
- Step 2. Use the model to simulate potential future asset prices and hence deliver a potential value for the daily portfolio loss.
- Step 3. Repeat step 2 many times and use the simulated values to create an approximate distribution.
- Step 4. For a given confidence level calculate VaR and TVaR using the same approach as with the historical simulation method.

To illustrate the Monte Carlo method we shall consider a familiar environment, namely we shall assume that our portfolio does not contain derivative products and that the n-dimensional vector of loss rates has the joint normal distribution, i.e., we assume that

$$\begin{pmatrix} l_1(t) \\ \vdots \\ l_n(t) \end{pmatrix} = \mathbf{l}_t \sim N(\mathbf{e}, \mathbf{V})$$

where, as usual, $\mathbf{e} \in \mathbb{R}^n$ denotes the vector of expected returns and $\mathbf{V} \in \mathbb{R}^{n \times n}$ denotes the covariance matrix. We shall assume that we are equipped with both $\mathbf{e}$ and $\mathbf{V}$, for instance they can be estimated from historical data.

In this setting we know that, given a vector $\mathbf{w} \in \mathbb{R}^n$ of portfolio weights, the corresponding VaR, for a confidence level α, is given by

$$\mathrm{VaR}_\alpha^{\mathrm{true}} = V(t)\Big[\mathbf{w}^T\mathbf{e} + \sqrt{\mathbf{w}^T\mathbf{V}\mathbf{w}}\Phi^{-1}(\alpha)\Big], \tag{22.13}$$

where $V(t)$ denotes the current value of the portfolio.

We now aim to use this toy example to shed some light on the Monte Carlo methodology for risk measurement. We set this development out as follows:

- Decomposing the covariance matrix.
  We shall assume that the covariance matrix $\mathbf{V}$ is positive definite and thus, according to Theorem 2.6, it possesses a unique Choleski decomposition, i.e.,

$$\mathbf{V} = \mathbf{R}\mathbf{R}^T \quad \text{where } \mathbf{R} \text{ is lower triangular.} \tag{22.14}$$

- The inverse of the covariance matrix.
  The inverse of $\mathbf{V}$ is also positive definite and, using (22.14), it can be written as

$$\mathbf{V}^{-1} = \mathbf{R}^{-T}\mathbf{R}^{-1} \quad \text{where } \mathbf{R}^{-T} = (\mathbf{R}^{-1})^T \text{ is upper triangular.} \tag{22.15}$$

- A Choleski transformation.
  We can transform the loss random vector $\mathbf{l}_t$ by shifting about its mean and multiplying it by $\mathbf{R}^{-1}$, i.e., we consider

$$\mathbf{z}_t = \mathbf{R}^{-1}(\mathbf{l}_t - \mathbf{e}).$$

  We notice that the new vector has zero mean and its covariance matrix is the identity, i.e.,

$$\mathbb{E}[\mathbf{z}_t] = \mathbf{R}^{-1}(\mathbb{E}[\mathbf{l}_t] - \mathbf{e}) = \mathbf{0}$$

  and

$$\begin{aligned}
\mathbb{E}[\mathbf{z}_t\mathbf{z}_t^T] &= \mathbb{E}[\left(\mathbf{R}^{-1}(\mathbf{z}_t - \mathbf{e})\right)\left(\mathbf{R}^{-1}(\mathbf{z}_t - \mathbf{e})\right)^T] \\
&= \mathbb{E}[\mathbf{R}^{-1}(\mathbf{z}_t - \mathbf{e})(\mathbf{z}_t - \mathbf{e})^T\mathbf{R}^{-T}] \\
&= \mathbf{R}^{-1}\mathbb{E}[(\mathbf{z}_t - \mathbf{e})(\mathbf{z}_t - \mathbf{e})^T]\mathbf{R}^{-T} \\
&= \mathbf{R}^{-1}\mathbf{V}\mathbf{R}^{-T} = \mathbf{R}^{-1}\mathbf{R}\mathbf{R}^T\mathbf{R}^{-T} = \mathbf{I}_n.
\end{aligned}$$

- Scenario generation.
  The above development is helpful as it tells us that

$$\begin{aligned}
&\text{if } \mathbf{l}_t = (l_1(t), \ldots, l_n(t))^T \sim N(\mathbf{e}, \mathbf{V}) = N(\mathbf{e}, \mathbf{R}\mathbf{R}^T) \\
&\text{then } \mathbf{z}_t = \mathbf{R}^{-1}(\mathbf{l}_t - \mathbf{e}) \sim N(\mathbf{0}, \mathbf{I}_n) \text{ standardized.}
\end{aligned} \tag{22.16}$$

  In particular we can deduce from (22.16) that realizations of the loss random vector can be written as

$$\mathbf{l}_t = \mathbf{R}\mathbf{z}_t + \mathbf{e},$$

  where $\mathbf{z}_t \sim N(\mathbf{0}, \mathbf{I}_n)$.

22.3.1 The Choleski algorithm

The example above demonstrates the importance of the Cholseki factorization in generating future scenarios for a financial portfolio. In view of this we now turn to the practical challenge of developing an algorithm, which can be implemented on a computer, that is designed to deliver the Choleski factor of a given positive definite matrix. We begin by examining the 2×2 case, where we are looking for components r_{11}, r_{21} and r_{22} such that

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix} = \begin{pmatrix} r_{11} & 0 \\ r_{21} & r_{22} \end{pmatrix} \begin{pmatrix} r_{11} & r_{21} \\ 0 & r_{22} \end{pmatrix}.$$

Multiplying this out, we find that

$$\begin{aligned}
\sigma_{11} = r_{11}^2 &\;\Rightarrow\; r_{11} = \sqrt{\sigma_{11}}; \\
\sigma_{21} = r_{21}r_{11} &\;\Rightarrow\; r_{21} = \frac{\sigma_{21}}{r_{11}}; \\
\sigma_{22} = r_{21}^2 + r_{22}^2 &\;\Rightarrow\; r_{22} = \sqrt{\sigma_{22} - r_{21}^2}.
\end{aligned}$$

More generally, we have

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix} = \begin{pmatrix} r_{11} & 0 & \cdots & 0 \\ r_{21} & r_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix} \begin{pmatrix} r_{11} & r_{21} & \cdots & r_{n1} \\ 0 & r_{22} & \cdots & r_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix}.$$

Now, applying the rules of matrix multiplication we find that

$$r_{jj} = \sqrt{\sigma_{jj} - \sum_{k=1}^{j-1} r_{jk}^2} \quad \text{for} \quad j = 1, \ldots, n,$$

and

$$r_{ij} = \frac{\sigma_{ij} - \sum_{k=1}^{j-1} r_{ik} r_{jk}}{r_{jj}} \quad \text{for} \quad i = j+1, \ldots, n.$$

We make the following deductions:

- The first column of $\mathbf{R}$ is given by

$$\begin{pmatrix} r_{1,1} \\ r_{2,1} \\ \vdots \\ r_{i,1} \\ \vdots \\ r_{n,1} \end{pmatrix} = \begin{pmatrix} \sqrt{\sigma_{11}} \\ \sigma_{12}/r_{11} \\ \vdots \\ \sigma_{1i}/r_{11} \\ \vdots \\ \sigma_{1n}/r_{11} \end{pmatrix}.$$

- Using the first column we can compute the second column and so on. In general, given that we have calculated the first $(j-1)$ columns then the jth column is given by

$$\begin{pmatrix} r_{1,j} \\ r_{2,j} \\ \vdots \\ r_{j-1,j} \\ r_{j,j} \\ r_{j+1,j} \\ \vdots \\ r_{n,j} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ (\sigma_{jj} - \sum_{k=1}^{j-1} r_{j,k}^2)^{1/2} \\ (\sigma_{j,j+1} - \sum_{k=1}^{j-1} r_{j,k} r_{j+1,k})/r_{j,j} \\ \vdots \\ (\sigma_{j,n} - \sum_{k=1}^{j-1} r_{j,k} r_{n,k})/r_{j,j} \end{pmatrix}.$$

These observations enable us to write a computer program to derive **R** for a given covariance matrix **V**.

22.3.2 Generating random numbers

In order to mimic the probability laws of any statistical model we need to generate a long sequence of random variables from the required distribution. In order to kick start the process we consider the task of generating realizations $u_1, u_2, \ldots$ of random variables $U_1, U_2, \ldots$ that are independent and have the uniform distribution $U(0, 1)$. The reason for focusing on $U(0, 1)$ is that it is often possible (as we shall see) to transform independent $U(0, 1)$ sequences into new sequences which have a different distribution.

Our random numbers will be generated by a deterministic computer algorithm and so, from a philosophical point of view, we cannot really think of the output as being truly random. Indeed, the realistic target is to develop algorithms that are carefully designed to approximate the randomness of the uniform distribution; such algorithms are known collectively as pseudo-random number generators. More precisely, we give the following definition:

Definition 22.1. *A pseudo-random sequence* $(u_n)_{n\geq 1}$ *is a deterministic sequence of numbers in* $[0, 1]$ *having the same relevant statistical properties as a sequence of independent random numbers chosen from the uniform distribution* $U(0, 1)$.

Computers generate the pseudo-random sequence by iterating a deterministic formula starting from an initial value u_0 commonly called the seed of the algorithm. Formally we have a finite set E and a function $f : E \to E$, then given the seed u_0, we iteratively generate the sequence

$$u_0, \ \underbrace{f(u_0)}_{u_1}, \ \underbrace{f^2(u_0)}_{u_2}, \ \ldots \ , \ \underbrace{f^k(u_0)}_{u_k}, \ \ldots$$

The science behind the specification of an appropriate function f is beyond the scope of this book. Fortunately, almost all modern mathematical and statistical software packages are equipped with reliable pseudo-random number generators. In order to give an indicator of what makes a good choice of f we make the following observations:

- Periodicity.
  The function f is defined upon a finite set and thus, eventually, the sequence will repeat itself; the number of iterations before a repetition is called the period of f. The mark of a good pseudo-random number generator is that its period is long, most of the commonly used packages will have a period of the order $> 2^{40}$.
- Speed.
  A desirable feature of any pseudo-random sequence is that the numbers are delivered quickly, thus one should ensure that the computational time needed to evaluate the function f is minimal. Needless to say, the commonly used packages are designed to be time efficient.
- Testing.
  There exists a well-defined batch of statistical tests that are purpose built to assess how accurately the generated sequence mimics a uniformly distributed sequence. Again, the most commonly used packages are those that have performed consistently well under testing.

Once we have generated a sequence $(u_k)_{k\geq 1}$ of pseudo-random numbers there exist a variety of transformations which result in a new sequence of numbers that exhibit the statistical properties of a different distribution. As an example we demonstrate the famous Box–Muller transformation which yields a new sequence $(z_n)_{n\geq 1}$ that exhibits the properties of the standard normal distribution.

22.3.2.1 The Box–Muller transform

We begin by considering the following two functions:

$$F(u, v) = \sqrt{-2\log u}\cos(2\pi v) \quad \text{and} \quad G(u, v) = \sqrt{-2\log u}\sin(2\pi v),$$

where $0 < u, v < 1$. We note that the natural logarithm is negative on $(0, 1)$ and so these functions are well defined. We now create two new random variables X and Y say, by setting

$$X = F(U, V) \quad \text{and} \quad Y = G(U, V)$$

and we note that the original pair (U, V) can be recovered from X and Y using appropriate mappings. In particular,

$$\phi(x, y) = \exp\left(-\frac{1}{2}(x^2 + y^2)\right)$$

allows us to recover U, since

$$\begin{aligned}
\phi(X, Y) &= \phi\left(F(U, V), G(U, V)\right) \\
&= \phi\left(\sqrt{-2\log U}\cos(2\pi V), \sqrt{-2\log U}\sin(2\pi V)\right) \\
&= \exp\left(-\frac{1}{2}\left(\log U^{-2}\underbrace{(\cos^2(2\pi V) + \sin^2(2\pi V))}_{=1}\right)\right) \\
&= \exp\left(-\frac{1}{2}\log U^{-2}\right) = \exp(\log(U)) = U
\end{aligned}$$

and

$$\psi(x, y) = \frac{1}{2\pi}\tan^{-1}\left(\frac{y}{x}\right)$$

allows us to recover V, since

$$\begin{aligned}
\psi(X, Y) &= \psi\left(F(U, V), G(U, V)\right) \\
&= \psi\left(\sqrt{-2\log U}\cos(2\pi V), \sqrt{-2\log U}\sin(2\pi V)\right)
\end{aligned}$$

$$= \frac{1}{2\pi} \tan^{-1} \left(\frac{\sqrt{-2 \log U} \sin(2\pi V)}{\sqrt{-2 \log U} \cos(2\pi V)} \right)$$

$$= \frac{1}{2\pi} \tan^{-1} (\tan(2\pi V)) = V.$$

We notice that we have created, using the functions F and G, a new pair of random variables X and Y in such a way that the originals (U and V) can be recovered using the transformations ϕ and ψ. In this situation we can use the same calculus we developed in our derivation of the F-distribution (see Section 12.4) to show that the joint density of (X, Y) is (using 12.12)

$$q(x, y) = 1 \times J(x, y) = \text{absolute value of det} \begin{pmatrix} \partial\phi/\partial x & \partial\psi/\partial x \\ \partial\phi/\partial y & \partial\psi/\partial y \end{pmatrix}$$

$$= \text{absolute value of det} \begin{pmatrix} -x\phi(x, y) & \frac{-y}{2\pi(x^2+y^2)} \\ -y\phi(x, y) & \frac{x}{2\pi(x^2+y^2)} \end{pmatrix}$$

$$= \left| -\frac{x^2\phi(x, y)}{2\pi(x^2 + y^2)} - \frac{y^2\phi(x, y)}{2\pi(x^2 + y^2)} \right|$$

$$= \frac{1}{2\pi}\phi(x, y) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x^2 + y^2) \right),$$

which we recognize as the density function for a pair of independent $N(0, 1)$ random variables. This transformation dates back to the late 1950s and is commonly referred to as the Box–Muller transformation in respect of its founders Box and Muller (1958). It is extremely useful, and to emphasize its importance we capture the result in the following theorem:

Theorem 22.2. *If U and V denote two independent and uniformly distributed random variables then the variables X and Y defined by*

$$X = \sqrt{-2 \ln U} \cos(2\pi V) \quad \text{and} \quad Y = \sqrt{-2 \ln U} \sin(2\pi V)$$

are independent and have the standard normal distribution.

We are now in a position to present the Monte Carlo simulation method for our basic model.

- MC Step 1.
  Take the covariance matrix $\mathbf{V}$ and find its Choleski factor $\mathbf{R}$.
- MC Step 2.
  Employ a pseudo-random number generator together with a suitable mapping (e.g., the Box–Muller transform) to produce an n-dimensional vector $\mathbf{z}_t$ which simulates a realization of the random vector from $N(\mathbf{0}, \mathbf{I}_n)$.

- MC Step 3.
  Compute a simulation of the portfolio loss using

$$\mathcal{L}_t = V(t)\left[\mathbf{w}^T \mathbf{R}\mathbf{z}_t + \mathbf{w}^T \mathbf{e}\right].$$

- MC Step 4.
  Repeat steps 2 and 3 over again until N hypothetical portfolio losses are available, i.e., we have $\{\mathcal{L}_t^k : k = 1, \ldots, N\}$.
  Use the simulated loss data to derive the $100\alpha\%$ quantile, i.e., compute

$$\text{VaR}_\alpha^{\text{MC}}(N) = \text{quantile}\left\{(\mathcal{L}_t^k)_{k=1}^N : \ 100\alpha\%\right\}. \tag{22.17}$$

To see the Monte Carlo recipe in action we consider the following simple example. We assume that we are able to invest in three assets whose loss random variables $\{l_1, l_2, l_3\}$ have the joint normal distribution

$$\begin{pmatrix} l_1 \\ l_2 \\ l_3 \end{pmatrix} \sim N\left[\begin{pmatrix} 3 \\ 4 \\ 5 \end{pmatrix}, \begin{pmatrix} 2 & 1 & 2 \\ 1 & 4 & 3 \\ 2 & 3 & 6 \end{pmatrix}\right].$$

We choose to form an equally weighted portfolio and, in this case, the portfolio loss random variable is

$$\mathcal{L}_t = \frac{1}{3}(l_1 + l_2 + l_3).$$

The mean and variance of $\mathcal{L}_t$ are computed as follows:

$$\mu = \frac{1}{3}(3 + 4 + 5) = 4;$$
$$\sigma^2 = \frac{1}{9}(2 + 1 + 2 + 1 + 4 + 3 + 2 + 3 + 6) = \frac{8}{3}.$$

The true VaR at the 95% confidence level is given by

$$\text{VaR}_{0.95}^{\text{true}} = 4 + \frac{2\sqrt{2}}{\sqrt{3}} 1.65 = 6.69.$$

Applying the Choleski algorithm we find that

$$\begin{pmatrix} r_{11} \\ r_{21} \\ r_{31} \end{pmatrix} = \begin{pmatrix} \sqrt{2} \\ 1/\sqrt{2} \\ \sqrt{2} \end{pmatrix},$$

$$\begin{pmatrix} r_{12} \\ r_{22} \\ r_{32} \end{pmatrix} = \begin{pmatrix} 0 \\ \sqrt{4 - r_{21}^2} \\ \frac{3 - r_{21} r_{31}}{\sqrt{4 - r_{21}^2}} \end{pmatrix} = \begin{pmatrix} 0 \\ \sqrt{4 - 1/2} \\ \frac{3-1}{\sqrt{4-1/2}} \end{pmatrix}$$

and

$$\begin{pmatrix} r_{13} \\ r_{23} \\ r_{33} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \sqrt{6 - r_{31}^2 - r_{32}^2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \sqrt{6 - 2 - \frac{4}{4-1/2}} \end{pmatrix}.$$

Thus, the Choleski factor is given by

$$\mathbf{R} = \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 1/\sqrt{2} & \sqrt{7/2} & 0 \\ \sqrt{2} & \sqrt{8/7} & \sqrt{20/7} \end{pmatrix}.$$

We can now use a pseudo-random number generator to deliver a vector $\mathbf{z} = (z_1, z_2, z_3)^T \in \mathbb{R}^3$ whose components are designed to mimic three independent standard normal random variables. A potential value of the portfolio loss random variable is then given by the simulation

$$\mathcal{L}_t^{\text{sim}} = 4 + \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 1/\sqrt{2} & \sqrt{7/2} & 0 \\ \sqrt{2} & \sqrt{8/7} & \sqrt{20/7} \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}.$$

This process is then repeated a large number of times, so as to build up a picture of the true distribution, and the VaR is calculated according to (22.17).

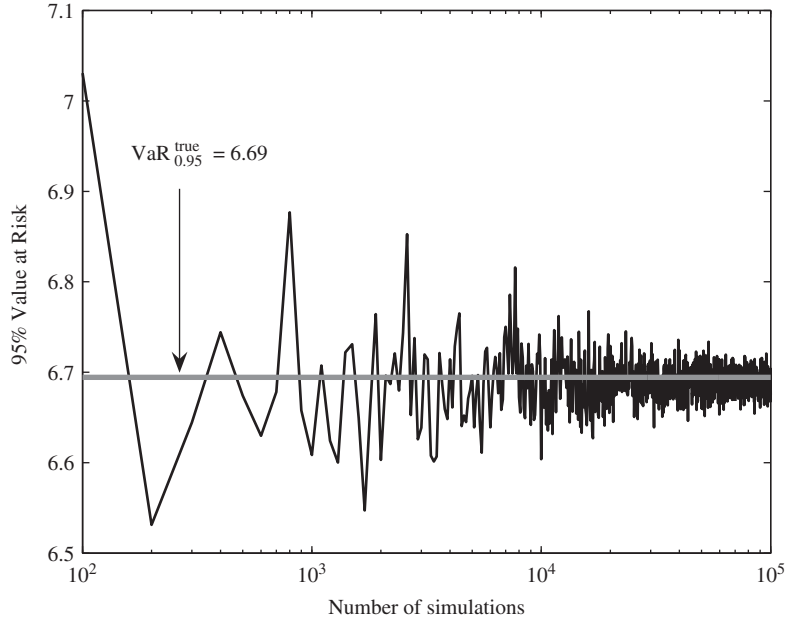


Figure 22.3 The convergence of Monte Carlo VaR.

Figure 22.3 establishes that Monte Carlo VaR estimates do converge to the true value, albeit at a rather slow rate. We remark that this is not surprising as we know from the previous section that the rate of convergence of VaR estimates based upon an empirical distribution of size n is of the order $1/\sqrt{n}$. Now, despite its slow convergence, the Monte Carlo simulation method ranks as one of the most popular approaches for estimating VaR. We have only demonstrated the technique in the simplified framework of normality and constant volatility. Clearly a risk manager would put more trust in a VaR estimate that is calculated from a more accurate loss model, e.g., one with GARCH-type features. The beauty of the Monte Carlo simulation approach is that it can be applied universally to any model and it is this flexibility that has established the Monte Carlo methodology as one of the most important tools in risk management. An enormous amount of academic research has been directed towards the use of Monte Carlo techniques and the basic scheme we outlined above can be refined in many different ways so as to ensure more accurate estimates with improved convergence rates; the reader who wants to discover more is encouraged to consult Glasserman (2004).

Essential Mathematics for Market Risk Management
by Simon Hubbert

23 Alternative Approaches to VaR

In this short chapter we revisit the Normal VaR framework of Chapter 10 and demonstrate how two fairly simple modifications can (potentially) deliver more accurate VaR estimates. Firstly, we make the assumption that the portfolio loss random variable has the t-distribution; this simple idea allows us to account for the fat-tailed nature of losses and also delivers closed-form solutions. Secondly, we propose a correction to the normal distribution so that both the skewness and the kurtosis of the true distribution are accounted for.

23.1 THE t-DISTRIBUTED ASSUMPTION

In order to overcome the deficiencies of the Normal VaR framework we propose that an alternative distribution be used, one that is better placed to capture the properties of financial loss data. To set the scene we will assume that μ and σ denote mean and volatility of $\mathcal{L}_t$ and, in addition, we let $F(x)$ and $f(x)$ denote, respectively, the distribution and density functions of the standardized loss

$$\mathcal{Z}_t = \frac{\mathcal{L}_t - \mu}{\sigma}. \tag{23.1}$$

Under the assumption that F is continuous and strictly increasing, we know that the daily VaR at confidence level α is given by

$$\text{VaR}_\alpha = \mu + \sigma F^{-1}(\alpha). \tag{23.2}$$

We see from the above formula that the crucial tool needed to deliver VaR_α is the ability to calculate the α-quantile $F^{-1}(\alpha)$.

We have already discovered that setting F to be the standard normal distribution is inadequate because it fails to capture the true tail behaviour. It is widely regarded that out of the many candidates for F the t-distribution is the most suitable since it is flexible enough to capture a range of tail decay behaviour. We have already encountered the t-distribution in Section 12.5 and, from there, we recall that the thickness of its tail (and hence the value of its kurtosis coefficient) varies with the number of degrees of freedom, see Figure 23.1.

In order to pursue this idea further we shall assume that $\mu \in \mathbb{R}$ and $\sigma > 0$ are two constants chosen so that

$$\frac{\mathcal{L}_t - \mu}{\sigma} \text{ has the standard } t\text{-distribution with } n \text{ degrees of freedom.}$$

We recall from Section 11.5 that if X is t-distributed random variable with $n > 2$ degrees of freedom then its first two moments are

$$\mathbb{E}[X] = 0 \quad \text{and} \quad \mathbb{E}[X^2] = \frac{n}{n-2}$$

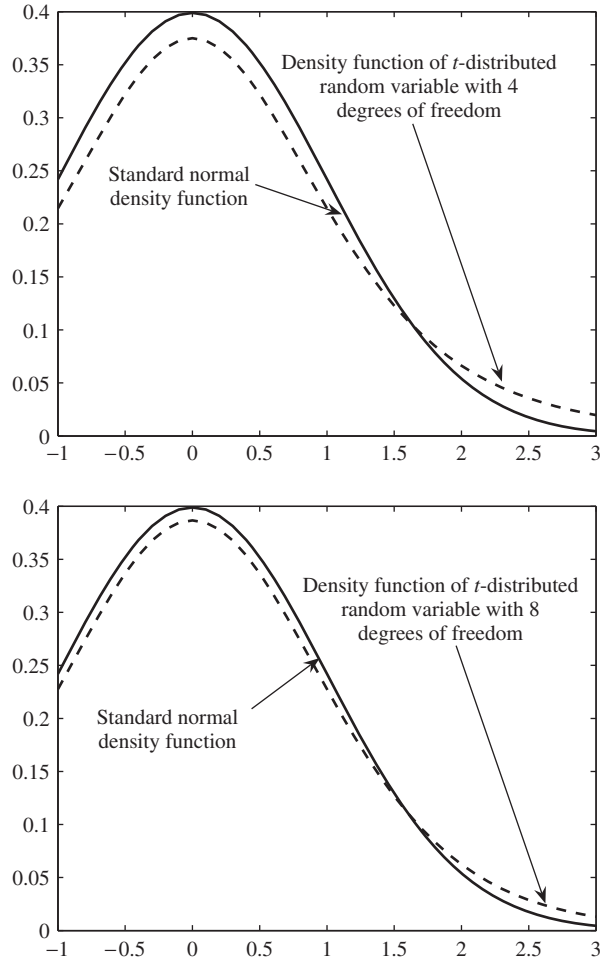


Figure 23.1 Comparison of normal and t-distributed density functions.

Thus, as a result, we can deduce that the mean of $\mathcal{L}_t$ is simply the constant μ and its volatility is given by $\sigma\sqrt{n/n-2}$.

We can now use (23.2) to deduce that the t-distributed VaR for a confidence level α is given by

$$\text{VaR}_\alpha(\mathcal{L}_t) = \mu + \sigma\sqrt{\frac{n}{n-2}}t_n^{-1}(\alpha) \tag{23.3}$$

where, using (12.15) and (12.16), $t_n^{-1}(\alpha)$ is the number that satisfies

$$\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\frac{1}{\sqrt{n\pi}}\int_{-\infty}^{t_n^{-1}(\alpha)}\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}dx = \alpha. \tag{23.4}$$

We remark that, as with the normal case, the value of $t_n^{-1}(\alpha)$ is readily available from statistical tables and/or most mathematical software packages.

A glance at (23.3) reveals that the VaR estimate depends upon n, the number of degrees of freedom we take to fix the t-distribution. One way to select this value would be to employ the maximum likelihood method, i.e., based upon a time series of historical losses $\{\mathcal{L}_{t-\tau}\}_{\tau=1}^{T}$ we search for the value of n that maximizes the likelihood function

$$L(n) = \prod_{\tau=1}^{T} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \frac{1}{\sqrt{n\pi}} \left(1 + \frac{1}{n}\left(\frac{\mathcal{L}_{t-\tau} - \mu}{\sigma}\right)^2\right)^{-\left(\frac{n+1}{2}\right)},$$

or equivalently, and perhaps more conveniently, the value of $n > 2$ that maximizes the log-likelihood function

$$LL(n) = T\left[\log\left(\Gamma\left(\frac{n+1}{2}\right)\right) - \log\left(\Gamma\left(\frac{n}{2}\right)\right) - \frac{1}{2}\log(\pi n)\right] - \frac{n+1}{2}\sum_{\tau=1}^{T}\left(1 + \frac{1}{n}\left(\frac{\mathcal{L}_{t-\tau} - \mu}{\sigma}\right)^2\right).$$

The maximum likelihood estimate is then found by allowing a suitable numerical optimization algorithm to search for the maximum of the above function.

As an alternative to the maximum likelihood estimation there is a much simpler approach (see also Christoffersen, 2003) which takes advantage of the fact that closed-form expressions are available for the moments of the t-distribution as a function of n. One of our main concerns is to ensure that the t-distribution we select should faithfully capture the tail of the loss distribution. In view of this, we focus attention on the kurtosis coefficient which, we recall from Section 11.5, is given by the formula

$$\mathcal{K}(X) = 3\left(\frac{n-2}{n-4}\right), \text{ where } X \text{ is } t\text{-distributed with } n \text{ degrees of freedom.}$$

Using this fact the practitioner can then use the historical time series $\{\mathcal{L}_{t-\tau}\}_{\tau=1}^{T}$ to find an estimate $\widehat{\mathcal{K}}$ for the kurtosis coefficient. We then solve

$$\widehat{\mathcal{K}} = 3\left(\frac{n-2}{n-4}\right),$$

i.e., we set

$$n = \frac{4\widehat{\mathcal{K}} - 6}{\widehat{\mathcal{K}} - 3}.$$

In order to develop the t-distribution framework a little further we now tackle the problem of calculating TVaR. We recall from (9.12) that the TVaR at confidence level α is given by

$$\text{TVaR}_\alpha(\mathcal{L}_t) = \frac{1}{1-\alpha}\int_\alpha^1 \text{VaR}_u(\mathcal{L}_t)du.$$

Now, substituting (23.3) in the above formula and integrating we find that

$$\mathrm{TVaR}_\alpha(\mathcal{L}_t) = \mu + \frac{\sigma}{1-\alpha}\sqrt{\frac{n}{n-2}}\int_\alpha^1 t_n^{-1}(u)du.$$

We can simplify the expression further by making the substitution

$$x = t_n^{-1}(u) \Rightarrow u = t_n(x) \Rightarrow du = p_n(x)dx,$$

where

$$p_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\frac{1}{\sqrt{n\pi}}\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}$$

is the density function of a t-distributed random variable with n degrees of freedom, see (12.15). The substitution causes the integration limits to change as follows, $\alpha \mapsto t_n^{-1}(\alpha)$ and $1 \mapsto \infty$, and we are left with

$$\mathrm{TVaR}_\alpha(\mathcal{L}_t) = \mu + \frac{\sigma}{1-\alpha}\frac{1}{\sqrt{\pi(n-2)}}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}\int_{t_n^{-1}(\alpha)}^\infty x\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}dx.$$

The above integral simplifies once more by employing the following substitution:

$$y = 1 + \frac{x^2}{n} \Rightarrow dy = \frac{2x}{n}dx,$$

and so we see that

$$\begin{aligned}\int_{t_n^{-1}(\alpha)}^\infty x\left(1+\frac{x^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}dx &= \frac{n}{2}\int_{1+\frac{(t_n^{-1}(\alpha))^2}{n}}^\infty y^{-\left(\frac{n+1}{2}\right)}dy \\ &= \frac{n}{2}\left[\frac{y^{-\left(\frac{n+1}{2}\right)}y}{\frac{1-n}{2}}\right]_{y=1+\frac{(t_n^{-1}(\alpha))^2}{n}}^{y\to\infty} \\ &= \left(1+\frac{(t_n^{-1}(\alpha))^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}\left(\frac{n+(t_n^{-1}(\alpha))^2}{n-1}\right).\end{aligned}$$

This calculation enables us to deduce that TVaR under the daily loss random variable is distributed as a t-distributed random variable, given by:

$$\mathrm{TVaR}_\alpha(\mathcal{L}_t) = \mu + \frac{\sigma}{1-\alpha}\sqrt{\frac{n}{n-2}}p_n(t_n^{-1}(\alpha))\left(\frac{n+(t_n^{-1}(\alpha))^2}{n-1}\right). \tag{23.5}$$

The t-distribution framework for risk measurement can be viewed as an improvement over the normal case because it accounts directly for the possibility of fat tails in the loss distribution. It is popular with practitioners due to the fact that there exist closed-form

expressions for both VaR_α and TVaR_α. However, one perceived drawback is the fact that the t-distribution, like the normal case, is symmetric and thus all risk calculations ignore the very real possibility that the true loss distribution may be skewed.

23.2 CORRECTIONS TO THE NORMAL ASSUMPTION

For our next approach we simply investigate whether we can quantify how much the standardized loss random variable (23.1) deviates from a standard normal random variable. Our plan of attack for this problem is to attempt to deliver an expansion of $f = F'$, the true density of (23.1), that has the following form:

$$f(x) = a_0\phi(x) + \frac{a_1}{1!}\phi'(x) + \frac{a_2}{2!}\phi''(x) + \cdots + \frac{a_n}{n!}\phi^{(n)}(x) + \cdots, \tag{23.6}$$

where ϕ denotes the standard normal density function.

Thus, our mission is to provide a formula for the expansion coefficients $(a_n)_{n=0}^{\infty}$ and also to shed light upon the conditions needed for convergence of (23.6) to be guaranteed for all x. The approach that we take is inspired by Cramér's treatment of the same problem, see Cramér (1966) Section 17.6.

In order to kick-start our analysis we will require a remarkable result from the branch of mathematics which deals with the decomposition of functions. In order to set the scene, we provide the following definition:

Definition 23.1. *Let $F : \mathbb{R} \to [0, 1]$ denote a distribution function whose moments are all finite. We say that $\widehat{x}$ is a point of increase for F if*

$$F(\widehat{x} - h) < F(\widehat{x} + h) \quad \text{for all } h > 0.$$

The following key result can be found in Szego (1959), Section 2.2.

Theorem 23.2. *Let $F : \mathbb{R} \to [0, 1]$ denote a distribution function whose moments are all finite. If the set of all points of increase of F is infinite then there exists a sequence of polynomials $(p_n(x))_{n=0}^{\infty}$ uniquely determined by the following conditions:*

- *p_n is of degree n and the coefficient of x^n is positive.*
- *The system $(p_n)_{n=0}^{\infty}$ is orthonormal with respect to F, i.e., the polynomials satisfy the orthogonality conditions*

$$\int_{\mathbb{R}} p_m(x)p_n(x)dF(x) = \begin{cases} 1 & \text{if } m = n, \\ 0 & \text{if } m \neq n. \end{cases}$$

In view of the above result we say that $(p_n)_{n=0}^{\infty}$ is the sequence of orthogonal polynomials associated with the distribution F.

To see why this result may be useful we consider the case of the standard normal distribution Φ, for which we know that the corresponding density is given by

$$\phi(x) = \frac{d}{dx}\Phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{23.7}$$

In addition, we provide the following definition:

Definition 23.3. *The Hermite polynomials denoted by $H_0, H_1, \ldots$ are defined by*

$$\frac{d^n}{dx^n}\exp\left(-\frac{x^2}{2}\right) = (-1)^n H_n(x)\exp\left(-\frac{x^2}{2}\right) \quad n = 0, 1, 2, \ldots \tag{23.8}$$

We remark that each $H_n(x)$ is a polynomial of degree n, and using (23.8), the first few examples are given by

$$\begin{aligned} H_0(x) &= 1, \\ H_1(x) &= x, \\ H_2(x) &= x^2 - 1, \\ H_3(x) &= x^3 - 3x, \\ H_4(x) &= x^4 - 6x^2 + 3. \end{aligned} \tag{23.9}$$

It can be shown that the sequence of orthogonal polynomials associated with Φ is closely related to the Hermite polynomials, indeed, by repeated integration by parts one can demonstrate that

$$\begin{aligned} \int_{\mathbb{R}} H_m(x)H_n(x)d\Phi(x) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} H_m(x)H_n(x)\exp\left(-\frac{x^2}{2}\right)dx \\ &= \begin{cases} n! & \text{if } m = n, \\ 0 & \text{if } m \neq n, \end{cases} \end{aligned} \tag{23.10}$$

and hence we can deduce that

$$\left(h_n(x) := \frac{1}{\sqrt{n!}}H_n(x)\right)_{n=0}^{\infty}$$

is the sequence of orthogonal polynomials associated with Φ. We notice that, according to (23.8) and (23.7), we have

$$\phi^{(n)}(x) = (-1)^n H_n(x)\phi(x)$$

and so (23.6) can be written as

$$f(x) = a_0\phi(x) - \frac{a_1}{1!}H_1(x)\phi(x) + \frac{a_2}{2!}H_2(x)\phi(x) + \cdots + \frac{a_n}{n!}(-1)^n H_n(x)\phi(x) + \cdots$$

Now, multiplying the series by $H_n(x)$ and integrating we see that (23.10) allows us to conclude that the expansion coefficients are given by

$$a_n = (-1)^n \int_{\mathbb{R}} H_n(x)f(x)dx \quad n = 0, 1, \ldots \tag{23.11}$$

Given that f is the density function of a zero-mean, unit-variance random variable we can use (23.9) to deduce that

$$a_0 = 1 \quad \text{and} \quad a_1 = a_2 = 0.$$

Furthermore, if we let μ_k denote the kth moment of the standardized loss $\mathcal{Z}_t$ then we can also compute

$$\begin{aligned}
a_3 &= -\mu_3 = -\mathcal{S} \quad \text{(the skewness coefficient of } \mathcal{L}_t\text{)},\\
a_4 &= \mu_4 - 3 = \mathcal{K} - 3 \quad \text{(the excess kurtosis coefficient of } \mathcal{L}_t\text{)},\\
a_5 &= \mu_5 + 10\mu_3,\\
a_6 &= \mu_6 - 15\mu_4 + 30.
\end{aligned}$$

Continuing in this way it is clear that a typical expansion coefficient a_k can be expressed as a certain linear combination of $\mathcal{Z}_t$'s moments of order $\leq k$. Furthermore, one can use these expressions in (23.6) to show that

$$f(x) = \phi(x) + r(x),$$

where the remainder is given by

$$\begin{aligned}
r(x) = & \underbrace{-\frac{1}{3!}\mathcal{S}\phi^{(3)}(x)}_{\text{first approximation}}\\
& + \underbrace{\frac{1}{4!}(\mathcal{K}-3)\phi^{(4)}(x) + \frac{10}{6!}\mathcal{S}^2\phi^{(6)}(x)}_{\text{first correction} \Rightarrow \text{improves accuracy}}\\
& \underbrace{-\frac{1}{5!}(\mu_5 - 10\mathcal{S})\phi^{(5)}(x) - \frac{35}{7!}\mathcal{S}(\mathcal{K}-3)\phi^{(7)}(x) - \frac{280}{9!}\mathcal{S}^3\phi^{(9)}(x) + \cdots}_{\text{second correction} \Rightarrow \text{improves accuracy further}}
\end{aligned}$$

We remark that each term of the expansion affects the accuracy of the expansion in a different way; it is not necessarily the case that the addition of one extra term will create an improved approximation. The above equation is displayed with the intention of demonstrating how the accuracy of the expansion improves as a new line of terms is added to the expansion. The reader who is interested in understanding this in more detail is advised to consult Cramér (1966), Section 17.7.

The expansions that we have developed in this section are better known as Edgeworth expansions, named after their discoverer Francis Edgeworth (1905). The expansion that is most commonly used in practice is as follows:

$$f(x) \approx \phi(x) - \frac{1}{3!}\mathcal{S}\phi^{(3)}(x) + \frac{1}{4!}(\mathcal{K}-3)\phi^{(4)}(x) + \frac{10}{6!}\mathcal{S}^2\phi^{(6)}(x),$$

which is popular as it corrects for deviations in both skewness and kurtosis.

We remark that we can integrate this to provide the following approximation for the distribution:

$$F(x) \approx \Phi(x) - \frac{1}{3!}\mathcal{S}\Phi^{(3)}(x) + \frac{1}{4!}(\mathcal{K}-3)\Phi^{(4)}(x) + \frac{10}{6!}\mathcal{S}^2\Phi^{(6)}(x).$$

Edgeworth expansions have found applications in many areas of finance, they are particularly popular in option pricing models, see Section 11.4 of Jondeau, Poon and Rockinger (2007). From a risk management perspective however, it is clear from (23.2) that VaR calculations could possibly be enhanced if we had access to an Edgeworth expansion for F^{-1}, the inverse of the standardized loss distribution. A review of the statistics literature shows that this problem was first investigated in the late 1930s by Edmund Cornish, a young Australian statistician, in collaboration with Sir Ronald Fisher, the famous English geneticist and statistician. Indeed, in 1937 they successfully demonstrated how an Edgeworth-type expansion could be used to deliver an approximation to the α-quantile $F^{-1}(\alpha)$ of a distribution which is known to be close to the standard normal case, see Cornish and Fisher (1937). The theoretical development of their approach is extremely technical and beyond the scope of this book; however, the interested reader may wish to consult Kendall and Stuart (1956). As an application of their pioneering work we quote the most commonly used version of the Cornish–Fisher theorem:

Theorem 23.4. *Let X denote a zero-mean, unit-variance random variable whose distribution function F is approximately equal to the standard normal distribution Φ. For a confidence level $\alpha \in (0,1)$, let us define*

$$x_\alpha = F^{-1}(\alpha) \quad \text{and} \quad z_\alpha = \Phi^{-1}(\alpha).$$

The following approximation then holds:

$$\begin{aligned} x_\alpha &\approx z_\alpha + \frac{\mathcal{S}}{3!}H_2(z_\alpha) - \frac{\mathcal{S}^2}{4!}\left(\frac{4}{3}H_3(z_\alpha) + \frac{2}{3}H_1(z_\alpha)\right) + \frac{\mathcal{K}-3}{4!}H_3(z_\alpha) \\ &= z_\alpha + \frac{\mathcal{S}}{6}(z_\alpha^2 - 1) - \frac{\mathcal{S}^2}{36}z_\alpha(2z_\alpha^2 - 5) + \frac{\mathcal{K}-3}{24}z_\alpha(z_\alpha^2 - 3), \end{aligned} \tag{23.12}$$

where $\mathcal{S}$ and $\mathcal{K}$ denote the skewness and kurtosis of X.

The above approximation (23.12) is commonly called the fourth-order Cornish–Fisher expansion since it allows for a correction for skewness and kurtosis, i.e., the third and fourth moments respectively. We remark that it is possible to develop the expansion much further in order to account for deviations in the higher-order moments. We can now employ the result directly to provide the following modified formula for VaR:

$$\text{VaR}_\alpha^{\text{mod}} = \mu + \sigma[z_\alpha + \text{ correction}], \tag{23.13}$$

where

$$\text{correction} = \frac{\mathcal{S}}{6}(z_\alpha^2 - 1) - \frac{\mathcal{S}^2}{36}(2z_\alpha^3 - 5z_\alpha) + \frac{\mathcal{K}-3}{24}(z_\alpha^3 - 3z_\alpha). \tag{23.14}$$

The modified Cornish–Fisher formula for VaR_α takes into account the fact that the underlying loss distribution is likely to be skewed and fat-tailed. The skewness and kurtosis coefficients can easily be estimated from historical time series data and so it is a fairly straightforward task to calculate $\text{VaR}_\alpha^{\text{mod}}$.

We close this section with a word of caution. The modified Value at Risk estimate (23.13) is designed for portfolios whose daily loss random variable does not differ too greatly from a normal random variable. In such cases practical experience shows that $\text{VaR}_\alpha^{\text{mod}}$ is far superior to the unreliable Normal VaR estimates. However, on the other hand, practical experience also shows that if the underlying distribution F is extremely fat-tailed then the theoretical assumption that F is close to the standard normal breaks down and in these cases one finds that $\text{VaR}_\alpha^{\text{mod}}$ is extremely unreliable. A nice discussion of the pitfalls of using $\text{VaR}_\alpha^{\text{mod}}$ can be found in Alexander (2008b).

Essential Mathematics for Market Risk Management
by Simon Hubbert

24 Backtesting

In 1996 the Basel committee on banking supervision announced their new internal model approach for calculating regulatory risk capital (RRC). We think of RRC as the buffer fund of cash and equity that a financial institution must maintain in order to survive a potential loss resulting from extreme market events. The committee proposed that the level of RRC be determined via a formula involving the recent history of VaR estimates calculated by the bank's own internal (i.e., in-house) model. Specifically, the committee insisted that each financial institution compute their VaR estimates for a 10-day holding period and at the 99% confidence level. This approach is universally adhered to and thus financial institutions across the globe employ their chosen risk model to compute the required regulatory VaR figure, denoted by $\text{VaR}_{0.99}^{(10)}$, and then, based upon these values, they must ensure that their risk capital is at least that defined by the following formula:

$$\text{RRC}(t) = \max\left(\text{VaR}_{0.99}^{(10)}(t-1), \frac{P}{60}\sum_{\tau=1}^{60}\text{VaR}_{0.99}^{(10)}(t-\tau)\right), \tag{24.1}$$

where P is a penalty factor which is initially set to 3 but can be increased by the regulator if the institution's VaR model is found to be sub-standard.

As we know, there is no one-size-fits-all approach for calculating VaR but rather a whole host of different methodologies of varying levels of sophistication. Thus, it is the responsibility of the risk manager to implement the most appropriate one. Indeed, the selection of the final VaR model is a crucial decision and one that must strike a balance between the following factors:

- A VaR model that tends to overestimate risk will, according to (24.1), cause the bank to set aside a much larger sum of risk capital than is needed.
- On the other hand, a VaR model that tends to underestimate risk will be penalized by the regulator. Specifically, the value of P in (24.1) will be increased to ensure that the bank sets aside a sufficient amount of risk capital. This penalty factor will only be reduced when the bank can demonstrate that it has made improvements to its risk model.

In view of this it is clear that the risk manager requires some means of backtesting the implemented VaR model in order to judge its validity; we shall devote this chapter to developing a scientific framework to address this problem.

24.1 QUANTIFYING THE PERFORMANCE OF VaR

To motivate our discussion we shall assume that at time $t-1$ our chosen VaR model delivers the estimate $\text{VaR}_\alpha(t)$ for the daily Value at Risk for date t. The success of this estimate will be established at the end of date t when we can compare $\text{VaR}_\alpha(t)$ against $\mathcal{L}_t$, the realized

loss for that day. In view of this we define an indicator random variable by

$$I_t = \begin{cases} 1 & \text{if } \mathcal{L}_t > \text{VaR}_\alpha(t), \\ 0 & \text{otherwise.} \end{cases}$$

We let p denote the probability of a failure, i.e.,

$$p = \mathbb{P}[I_t = 1] \Rightarrow \mathbb{P}[I_t = 0] = 1 - p.$$

In mathematical terminology we say that I_t is an example of a Bernoulli trial; the name given to a random experiment for which there are only two possible outcomes: success and failure. The event $\{I_t = 1\}$ (a failure) represents a VaR exception and thus, for a perfect risk model, we would expect that p (the probability of an exception) should equal $1 - \alpha$.

In order to judge the performance of the chosen model we make the additional assumption that $(I_{t+\tau})_{\tau \geq 0}$ is an independent sequence. We then turn to the past history of the model and examine the past T days' worth of VaR estimates. Specifically, we compute the number of exceptions by evaluating

$$T_{\text{ex}} = \sum_{\tau=1}^{T} I_{t-\tau}, \tag{24.2}$$

and consequently, the proportion of exceptions is given by

$$\pi_{\text{ex}} = \frac{1}{T} \sum_{\tau=1}^{T} I_{t-\tau} = \frac{T_{\text{ex}}}{T}. \tag{24.3}$$

Now, if our model is fit for purpose we would expect that $\pi_{\text{ex}} \approx 1 - \alpha$. In fact, the Basel regulatory committee use this idea when setting the value of the penalty parameter P in (24.1). Specifically, a sample of 250 VaR estimates at the 99% confidence level is required. In this setting the regulators calculate π_{ex} (24.3) which, all being well, should be approximately 0.01. The penalty parameter is determined by the amount by which π_{ex} deviates from 0.01, see Table 24.1.

24.2 TESTING THE PROPORTION OF VaR EXCEPTIONS

In order to establish a concrete statistical test we follow the approach of Christoffersen (1998), see also Christoffersen (2003), Chapter 8. We begin with the simple observation that the likelihood of a VaR success/failure is summarized by $p^{I_t}(1-p)^{1-I_t}$. Now, given

Table 24.1 Penalty parameter values determined by regulatory backtesting

π_{ex}	≤ 0.016	0.02	0.024	0.028	0.032	0.036	≥ 0.04
P	3	3.4	3.5	3.65	3.75	3.85	4

that we assume the indicator events occur independently of each other, we can use historical data from the past T days to construct the associated likelihood function

$$\begin{aligned} L(p) &= \prod_{\tau=1}^{T} p^{I_{t-\tau}} (1-p)^{1-I_{t-\tau}} \\ &= p^{\sum_{\tau=1}^{T} I_{t-\tau}} (1-p)^{T-\sum_{\tau=1}^{T} I_{t-\tau}} \\ &= p^{T_{\text{ex}}} (1-p)^{T-T_{\text{ex}}}, \end{aligned}$$

and, hence, also the associated log-likelihood function

$$\begin{aligned} LL(p) &= \log\left(p^{T_{\text{ex}}} (1-p)^{T-T_{\text{ex}}}\right) \\ &= T_{\text{ex}} \log(p) + (T - T_{\text{ex}}) \log(1-p). \end{aligned} \tag{24.4}$$

If we let p^* denote the value of p that maximizes (24.4), then its value must satisfy the first-order condition

$$\frac{d}{dp} LL(p)\Big|_{p=p^*} = \frac{T_{ex}}{p^*} - \frac{T - T_{ex}}{1-p^*} = 0. \tag{24.5}$$

One can easily show that (24.5) holds when $p^* = T_{\text{ex}}/T$, i.e., the maximum likelihood estimate for the probability of a VaR exception is the same as the sample estimate (24.3).

We are now in a position to develop a statistical test that is designed to validate (or otherwise) the null hypothesis

$$H_0 : p = 1 - \alpha \quad \text{against} \quad H_1 : p \neq 1 - \alpha.$$

To do this we employ the so-called generalized likelihood ratio (18.17). We recall that the construction of this ratio involves the calculation of the following two quantities:

- The maximum value of the likelihood function when the unknown parameter is allowed to vary over the full parameter space Θ.
  - In our case the parameter space is the interval $[0, 1]$ and the maximum value of the likelihood function is

$$L_{\text{full}}(p^*) = \left(\frac{T_{\text{ex}}}{T}\right)^{T_{\text{ex}}} \left(1 - \frac{T_{\text{ex}}}{T}\right)^{T-T_{\text{ex}}}. \tag{24.6}$$

- The maximum value of the likelihood function when the unknown parameter is restricted to some subspace Θ_{res} of Θ.
  - In our case the restricted parameter space is the single point $\{1 - \alpha\}$ (the true probability of a VaR_α exception) and the corresponding likelihood function is

$$L_{\text{res}}(1-\alpha) = (1-\alpha)^{T_{\text{ex}}} \alpha^{T-T_{\text{ex}}}.$$

In view of the above, the generalized likelihood ratio for the proportion of VaR exceptions based on T observations is given by

$$\boldsymbol{\lambda}_{\text{prop}}(T) = \frac{L_{\text{res}}(1-\alpha)}{L_{\text{full}}(p^*)} = \frac{(1-\alpha)^{T_{\text{ex}}}\alpha^{T-T_{\text{ex}}}}{\left(\frac{T_{\text{ex}}}{T}\right)^{T_{\text{ex}}}\left(1-\frac{T_{\text{ex}}}{T}\right)^{T-T_{\text{ex}}}}.$$

We note that, by definition, $\boldsymbol{\lambda}_{\text{prop}}(T)$ can never exceed unity. Indeed, the smaller its value the less likely the null hypothesis is supported by the data. We can be more scientific here since, under certain regularity conditions, we can appeal to Proposition 18.2, which tells us that the following log ratio test statistic:

$$LR_{\text{prop}} = -2\log(\boldsymbol{\lambda}_{\text{prop}}(T)) = -2\log\left(\frac{(1-\alpha)^{T_{\text{ex}}}\alpha^{T-T_{\text{ex}}}}{\left(\frac{T_{\text{ex}}}{T}\right)^{T_{\text{ex}}}\left(1-\frac{T_{\text{ex}}}{T}\right)^{T-T_{\text{ex}}}}\right) \tag{24.7}$$

converges to a chi-squared random variable with one degree of freedom as $T \to \infty$. Using this information we test the validity of the null hypothesis. Specifically, we choose a significance level $l \in (0, 1)$ for the test and proceed as follows:

- Using statistical tables or a computer package, calculate $z_1(l)$, the $100 \cdot l\%$ quantile of the χ_1^2, i.e., the value that satisfies $\mathbb{P}[Z_1 \leq z_1(l)] = l$, where Z_1 is a χ^2 random variable with one degree of freedom.
- For a large enough sample we can assume that LR_{prop} is approximately a χ^2 random variable with one degree of freedom, and thus

$$\mathbb{P}\Big[LR_{\text{prop}} \leq z_1(l)\Big] \approx l.$$

- We then follow the decision rule

$$\begin{aligned} &\text{accept } H_0, \text{ i.e., accept the model if } LR_{\text{prop}} \leq z_1(l) \\ &\text{otherwise the model is rejected.} \end{aligned} \tag{24.8}$$

- For further evidence we compute the p-value of the test:

$$p = 1 - \chi_1^2(LR_{\text{prop}}).$$

  If the p-value is small, specifically if $p < l$, then this can be used as evidence to support the rejection of the model.

For instance, if a significance level of 5% is chosen, i.e., $l = 0.05$, then the critical value of the χ_1^2 distribution is $z_1(0.05) = 3.84$. Thus, if the LR_{prop} test value is larger than 3.84 and if the p-value is less than 0.05 then we should consider rejecting the VaR model at the 5% level.

A word of caution: the practitioner should not lose sight of the fact that the derivation of the LR test statistic is based upon asymptotic theory, i.e., the distributional result holds in the limit as $T \to \infty$. In practice this means that, for more confident results, the test should only be applied when the sample size T is large. For an excellent account of the properties of the test and further examples, see Kupiec (1995).

24.3 TESTING THE INDEPENDENCE OF VaR EXCEPTIONS

A drawback of the above LR test is that although it captures the proportion of VaR exceptions it gives no consideration to their timing. The timing of VaR exceptions is important because, as we know, financial losses exhibit volatility clustering and, in periods of high volatility, the chance of suffering a large loss is increased. A good VaR model will take the clustering phenomenon into account and, as such, we would not expect to find VaR exceptions clustered together. It is therefore important that a good VaR model should treat VaR exceptions as independent events. In the late 1990s Christoffersen tackled this issue and proposed an enhancement of the LR test statistic to incorporate a test for the independence of VaR exceptions (Christoffersen, 1998). The idea behind his approach is to consider a set-up where the probability of an exception today is dependent upon whether or not an exception occurred yesterday. To set the scene we define

$$p_{\text{ee}} = \mathbb{P}[I_{t-\tau} = 1 \text{ given } I_{t-\tau-1} = 1] \text{ (exception, given previous exception)},$$
$$p_{\text{en}} = \mathbb{P}[I_{t-\tau} = 0 \text{ given } I_{t-\tau-1} = 1] \text{ (exception given previous non-exception)}.$$

Furthermore, using the laws of conditional probability, we have

$$p_{\text{ne}} = \mathbb{P}[I_{t-\tau} = 1 \text{ given } I_{t-\tau-1} = 0] = 1 - p_{\text{ee}},$$
$$p_{\text{nn}} = \mathbb{P}[I_{t-\tau} = 0 \text{ given } I_{t-\tau-1} = 0] = 1 - p_{\text{en}}.$$

We now examine a time series of past losses and previous VaR estimates to fix the following numbers:

$$T_{\text{ee}} = \text{occurrences of exception followed by exception};$$
$$T_{\text{en}} = \text{occurrences of exception followed by non-exception};$$
$$T_{\text{ne}} = \text{occurrences of non-exception followed by exception};$$
$$T_{\text{nn}} = \text{occurrences of non-exception followed by non-exception}.$$

The corresponding likelihood function for these observations is then given by

$$L(p_{\text{ee}}, p_{\text{en}}) = (1 - p_{\text{en}})^{T_{\text{nn}}} p_{\text{en}}^{T_{\text{en}}} (1 - p_{\text{ee}})^{T_{\text{ne}}} p_{\text{ee}}^{T_{\text{ee}}},$$

and the log-likelihood version is

$$LL(p_{\text{en}}, p_{\text{ee}}) = T_{\text{nn}} \log(1 - p_{\text{en}}) + T_{\text{en}} \log(p_{\text{en}}) + T_{\text{ne}} \log(1 - p_{\text{ee}}) + T_{\text{ee}} \log(p_{\text{ee}}).$$

The maximum likelihood estimates, p_{ee}^* and p_{en}^* say, are found by solving the first-order conditions, i.e.,

$$\frac{\partial LL}{\partial p_{\text{ee}}} = -\frac{T_{\text{ne}}}{1 - p_{\text{ee}}^*} + \frac{T_{\text{ee}}}{p_{\text{ee}}^*} = 0 \quad \Rightarrow \quad p_{\text{ee}}^* = \frac{T_{\text{ee}}}{T_{ne} + T_{\text{ee}}} \tag{24.9}$$

and

$$\frac{\partial LL}{\partial p_{\text{en}}} = -\frac{T_{\text{nn}}}{1 - p^*_{\text{en}}} + \frac{T_{\text{en}}}{p^*_{\text{en}}} = 0 \quad \Rightarrow \quad p^*_{\text{en}} = \frac{T_{\text{en}}}{T_{\text{en}} + T_{\text{nn}}}. \tag{24.10}$$

In this framework we can deduce that the unconstrained maximum value of the likelihood function is given by

$$L_{\text{full}}(p^*_{\text{en}}, p^*_{\text{ee}}) = \left(1 - p^*_{\text{en}}\right)^{T_{\text{nn}}} \left(p^*_{\text{en}}\right)^{T_{\text{en}}} \left(1 - p^*_{\text{ee}}\right)^{T_{\text{ne}}} \left(p^*_{\text{ee}}\right)^{T_{\text{ee}}}.$$

If we now enforce the restriction that the sequence of exceptions is independent, i.e., that $p_{\text{ne}} = p_{\text{ee}} = p$, then the likelihood function collapses to the more familiar form

$$p^{T_{\text{ex}}}(1 - p)^{T - T_{\text{ex}}}$$

where T_{ex}, as given by (24.2), is just the number of VaR exceptions over the period. In this setting we know from (24.5) that this quantity is maximized at $p^* = T_{\text{ex}}/T$ and so, the restricted likelihood function is given by

$$L_{\text{res}}(p^*) = \left(\frac{T_{\text{ex}}}{T}\right)^{T_{\text{ex}}} \left(1 - \frac{T_{\text{ex}}}{T}\right)^{T - T_{\text{ex}}}.$$

Developing in the same fashion as in the previous section, we can deduce that the generalized likelihood ratio for the independence of VaR exceptions is given by

$$\begin{aligned} \boldsymbol{\lambda}_{\text{ind}}(T) &= \frac{L_{\text{res}}(p^*)}{L_{\text{full}}(p^*_{\text{en}}, p^*_{\text{ee}})} \\ &= \frac{\left(\frac{T_{\text{ex}}}{T}\right)^{T_{\text{ex}}} \left(1 - \frac{T_{\text{ex}}}{T}\right)^{T - T_{\text{ex}}}}{\left(1 - p^*_{\text{en}}\right)^{T_{\text{nn}}} \left(p^*_{\text{en}}\right)^{T_{\text{en}}} \left(1 - p^*_{\text{ee}}\right)^{T_{\text{ne}}} \left(p^*_{\text{ee}}\right)^{T_{\text{ee}}}}, \end{aligned} \tag{24.11}$$

where p^*_{en} and p^*_{ee} are given by (24.10) and (24.9) respectively. Using the same arguments as before one can deduce that the log ratio test statistic

$$LR_{\text{ind}}(T) = -2 \log\left(\boldsymbol{\lambda}_{\text{ind}}(T)\right) \tag{24.12}$$

converges to a chi-squared random variable with one degree of freedom as $T \to \infty$. The value of $LR_{\text{ind}}(T)$ can then be used to validate the independence assumption of VaR violations in the usual way. In particular, this test is able to reject VaR models which exhibit clustering of VaR violations.

Our development so far has focused upon two important features that any proposed VaR model should exhibit:

- The proportion of VaR_α exceptions taken over the past T days should remain close to the expected value of $1 - \alpha$.
  - This property can be examined using the test statistic $LR_{\text{prop}}(T)$ (24.7) provided T is sufficiently large.

- The model should not produce clusters of VaR exceptions, they should be considered as independent events.
  - This property can be examined using the test statistic $LR_{\text{ind}}(T)$ (24.12) again, provided T is sufficiently large.

We now present a simple but helpful observation. We note that the numerator of $\boldsymbol{\lambda}_{\text{prop}}(T)$ is the same as the denominator of $\boldsymbol{\lambda}_{\text{ind}}(T)$, thus we can multiply these two quantities together to yield a new likelihood ratio

$$\begin{aligned}\boldsymbol{\lambda}_{\text{joint}}(T) &= \boldsymbol{\lambda}_{\text{prop}}(T)\boldsymbol{\lambda}_{\text{ind}}(T) \\ &= \frac{(1-\alpha)^{T_{\text{ex}}}\alpha^{T-T_{\text{ex}}}}{\left(1-p_{\text{en}}^{*}\right)^{T_{\text{nn}}}\left(p_{\text{en}}^{*}\right)^{T_{\text{en}}}\left(1-p_{\text{ee}}^{*}\right)^{T_{\text{ne}}}\left(p_{\text{ee}}^{*}\right)^{T_{\text{ee}}}}.\end{aligned} \tag{24.13}$$

As a consequence, we can employ the additive property of the logarithmic function to deduce that

$$\begin{aligned}-2\log\left(\boldsymbol{\lambda}_{\text{joint}}(T)\right) &= -2\log\left(\boldsymbol{\lambda}_{\text{prop}}(T)\boldsymbol{\lambda}_{\text{ind}}(T)\right) \\ &= -2\log\left(\boldsymbol{\lambda}_{\text{prop}}(T)\right) - 2\log\left(\boldsymbol{\lambda}_{\text{ind}}(T)\right) \\ &= LR_{\text{prop}}(T) + LR_{\text{ind}}(T).\end{aligned}$$

This development enables us to define the new test statistic

$$\begin{aligned}LR_{\text{joint}}(T) &= -2\log\left(\boldsymbol{\lambda}_{\text{joint}}(T)\right) \\ &= LR_{\text{prop}}(T) + LR_{\text{ind}}(T)\end{aligned} \tag{24.14}$$

which, by definition, converges to a chi-squared random variable with two degrees of freedom as $T \to \infty$. Thus, we can use $LR_{\text{joint}}(T)$ to jointly test that the proportion of VaR exceptions is within range and that they are independent; this is Christoffersen's framework for backtesting and is summarized as follows:

- Using a large sample of observed losses and VaR estimates, employ Christoffersen's joint test statistic $LR_{\text{joint}}(T)$ to validate the performance of the VaR model.
  - If the model passes the test there is nothing more to do.
  - If the model fails the test then one should determine whether this is because it inaccurately approximates the proportion of exceptions and/or if the exceptions tend to be clustered. This can be done by separately evaluating and testing $LR_{\text{prop}}(T)$ and $LR_{\text{ind}}(T)$ respectively.

References

Abramowitz, M. and Stegun, I.A. (1964) *Handbook of Mathematical Functions*, Dover Publications, New York.

Acerbi, C. and Tasche, D. (2002) On the coherence of expected shortfall, *Journal of Banking and Finance* **26**(7) 1487–1503.

Alexander, C. (2008a) *Market Risk Analysis, Volume II, Practical Financial Econometrics*, John Wiley, Chichester.

Alexander, C. (2008b) *Market Risk Analysis, Volume IV, Value-at-Risk Models*, John Wiley, Chichester.

Artzner, P., Delbaen, F., Eber, J.-M. and Heath, D. (1999) Coherent measures of risk, *Mathematical Finance* **9** 203–228.

Balkema, J. and de Hahn, L. (1974) Residual life at great age, *Annals of Probability* **2**.

Baxter, B.J.C. (2010) *Numerical Analysis Lecture Notes*, http://www.cato.tzo.com/brad/M2N1/.

Berger, M.A. (1992) *An Introduction to Probability and Stochastic Processes*, Springer Texts in Statistics, Springer, Berlin.

Bielecki, T.R. and Rutkowski, M. (2010) *Credit Risk: Modelling, Valuation and Hedging*, Springer Finance, Berlin.

Black, F. and Scholes, M. (1973) The pricing of options and corporate liabilities, *Journal of Political Economy* **81**, 637–654.

Bollerslev, T. (1986) Generalized autoregressive conditional heteroscedasticity, *Journal of Econometrics* **31**, 307–327.

Box, G.E.P. and Muller, M.E. (1958) A note on the generation of random normal deviates, *Annals of Mathematical Statistics* **29**(2), 610–611.

Britten-Jones, M. and Schaeffer, S.M. (1999) Nonlinear value at risk, *European Finance Review* **2** 161–187.

Cornish, E.A. and Fisher, R.A. (1937) Moments and cumulants in the specification of distributions, *Revue de l'Institut International de Statistique 5*, **4** 307–320.

Christoffersen, P.F. (1998) Evaluating interval forecasts, *International Economic Review* **39** 841–862.

Christoffersen, P.F. (2003) *Elements of Financial Risk Management*, Academic Press, New York.

Cox, D.R. and Hinkley, D.V. (1979) *Theoretical Statistics*, Chapman and Hall, London.

Cramér, H. (1966) *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ.

Dowd, K. (2002) *An Introduction to Market Risk Measurement*, John Wiley, Chichester.

Edgeworth, F.Y. (1905) The law of error, *Proceedings of the Cambridge Philosophical Society* **20** 36–66.

Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997) *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.

Engle, R.F. (1982) Auto-regressive conditional heteroskedasticity with estimates of variance of United Kingdom inflation, *Econometrica* **50** 987–1007.

Engle, R.F. and Ng, V. (1993) Measuring and testing the impact of news on volatility, *Journal of Finance* **48**(5) 1749–1778.

Fisher, R.A. and Tippet, L.H.C. (1928) Limiting forms of the frequency distribution of the largest or smallest number of the sample. *Proceedings of the Cambridge Philosophical Society* **24** 180–190.

Gil-Pelaez, J. (1951) Note on the inversion theorem, *Biometrika* **38** 481–482.

Gill, P.E., Murray, W. and Wright, M.H. (1982) *Practical Optimization*, Academic Press, New York.

Glasserman, P. (2004) *Monte Carlo Methods in Financial Engineering*, Springer Science, Berlin.

Glosten, L.R., Jagannathan, R. and Runkle, D.E. (1993) On the relation between the expected value and the volatility of the nominal excess return on stocks, *Journal of Finance* **48**(5) 1779–1801.

Gnedenko, B.V. (1941) Limit theorems for the maximal term of a variational series, *Comptes Rendus de l'Acad Sc l'USRR* **32**.

Gouriéroux, C. (1997) *ARCH Models and Financial Applications*, Springer Series in Statistics, Springer, Berlin.

Higham, D.J. (2004) *An Introduction to Financial Option Valuation*, Cambridge University Press, Cambridge.

Huand, C.F. and Litzenberger, R. (1988) *Foundations for Financial Economics*, North Holland, Amsterdam.

Hull, J.C. (2007) *Risk Management and Financial Institutions*, Pearson Prentice Hall, New Jersey.

Imhof, J.P. (1961) Computing the distribution of quadratric forms in normal variables, *Biometrika* **48** 419–426.

Jarque, C.K. and A.K. Bera, (1980) Efficiency tests for normality, homoskedasticity and serial independence of regression residuals, *Economic Letters* **6** 255–259.

Jondeau, E., Poon, S. and Rockinger, M. (2007) *Financial Modeling Under Non-Gaussian Distributions*, Springer Finance (Financial Engineering). Springer, Berlin

Jorion, P. (2006) *Value at Risk*, McGraw-Hill, New York.

Joshi, M.S. (2005) *The Concepts and Practice of Mathematical Finance*, Cambridge University Press, Cambridge.

Kendall, M.G. and Stuart, A. (1956) *The Advanced Theory of Statistics, Volume 1*, Charles Griffin and Co. Ltd.

Kupiec, P. (1995) Techniques for verifying the accuracy of risk measurement models, *Journal of Derivatives* **2** 173–184.

Mandelbrot, B.B. (1963) The variation of certain speculative prices, *Journal of Business* **XXXVI** 675–702.

Markowitz, H.M. (1952) Portfolio selection, *Journal of Finance* **7** 77–91.

McNeil, A.J., Frey, R. and Embrechts, P. (2005) *Quantitative Risk Management*, Princeton University Press, Princeton, NJ.

Merton, R.C. (1973) Theory of rational option pricing, *The Bell Journal of Econometrics and Management Science* **4**, 141–183.

Mina, J. and Ulmer, A. (1999) *Delta-Gamma: Four ways*, RiskMetrics Publication.

Moix, P-Y. (2001) *The Measurement of Market Risk*, LNEMS **504** Springer, Berlin.

Nash, S.G. and Sofer, A. (1996) *Linear and Nonlinear Programming*, McGraw-Hill, New York.

Neftci, S.N. (1996) *An Introduction to the Mathematics of Financial Derivatives*, Academic Press, New York.

Nelson, D.B. (1991) Conditional heteroskedasticity in asset returns: a new approach, *Econometrica* **59**, 347–370.

Picklands, J. (1975) Statistical inference using extreme order statistics, *Annals of Statistics* **3** 119–131.

Poon, S.-H. (2005) *A Practical Guide to Forecasting Financial Market Volatility*, John Wiley, Chichester.

Schönbucher, P.J. (2003) *Credit Derivatives Pricing Models*, John Wiley, Chichester.

Sharpe, W.F. (1964) Capital asset prices: a theory of market equilibrium under conditions of risk, *Journal of Finance* **19**(3) 425–442.

Szegö, G. (1959) *Orthogonal Polynomials*, American Mathematical Society, New York.

Taylor, S.J. (2007). *Asset Price Dynamics, Volatility and Prediction*, Princeton University Press, Princeton, NJ.

Wilmott, S., Howison, S. and Dewynne, J. (1995) *The Mathematics of Financial Derivatives – A Student Introduction*, Cambridge University Press, Cambridge.

Index