1
Risk in Perspective

In this chapter we provide a non-mathematical discussion of various issues that form the background to the rest of the book. In Section 1.1 we begin with the nature of risk itself and discuss how risk relates to randomness; in the financial context (which includes insurance) we summarize the main kinds of risks encountered and explain what it means to measure and manage such risks.

A brief history of financial risk management and the development of financial regulation is given in Section 1.2, while Section 1.3 contains a summary of the regulatory framework in the financial and insurance industries.

In Section 1.4 we take a step back and attempt to address the fundamental question of why we might want to measure and manage risk at all. Finally, in Section 1.5 we turn to quantitative risk management (QRM) explicitly and set out our own views concerning the nature of this discipline and the challenge it poses. This section in particular should give more insight into our choice of methodological topics in the rest of the book.

1.1 Risk

The *Concise Oxford English Dictionary* defines risk as “hazard, a chance of bad consequences, loss or exposure to mischance”. In a discussion with students taking a course on financial risk management, ingredients that are typically discussed are events, decisions, consequences and uncertainty. It is mostly only the downside of risk that is mentioned, rarely a possible upside, i.e. the potential for a gain. While for many people risk has largely negative connotations, it may also represent an opportunity. Much of the financial industry would not exist were it not for the presence of financial risk and the opportunities afforded to companies that are able to create products and services that offer more financial certainty to their clients.

For financial risks no single one-sentence definition of risk is entirely satisfactory. Depending on context, one might arrive at notions such as “any event or action that may adversely affect an organization’s ability to achieve its objectives and execute its strategies” or, alternatively, “the quantifiable likelihood of loss or less-than-expected returns”.

1.1.1 Risk and Randomness

Regardless of context, risk strongly relates to uncertainty, and hence to the notion of randomness. Randomness has eluded a clear, workable definition for many centuries;

it was not until 1933 that the Russian mathematician A. N. Kolmogorov gave an axiomatic definition of randomness and probability (see Kolmogorov 1933). This definition and its accompanying theory provide the language for the majority of the literature on risk, including this book.

Our reliance on probability may seem unsatisfactorily narrow to some. It bypasses several of the current debates on risk and uncertainty (Frank Knight), the writings on probabilistic thinking within economics (John Maynard Keynes), the unpredictability of unprecedented financial shocks, often referred to as Black Swans (Nassim Taleb), or even the more political expression of the known, the unknown and the unknowable (Donald Rumsfeld); see the Notes and Comments section for more explanation. Although these debates are interesting and important, at some point clear definitions and arguments are called for and this is where mathematics as a language enters. The formalism of Kolmogorov, while not the only possible approach, is a tried-and-tested framework for mathematical reasoning about risk.

In Kolmogorov's language a probabilistic model is described by a triplet $(\Omega, \mathcal{F}, P)$. An element ω of Ω represents a realization of an experiment, in economics often referred to as a state of nature. The statement "the probability that an event A occurs" is denoted (and in Kolmogorov's axiomatic system defined) as $P(A)$, where A is an element of $\mathcal{F}$, the set of all events. P denotes the probability measure. For the less mathematically trained reader it suffices to accept that Kolmogorov's system translates our intuition about randomness into a concise, axiomatic language and clear rules.

Consider the following examples: an investor who holds stock in a particular company; an insurance company that has sold an insurance policy; an individual who decides to convert a fixed-rate mortgage into a variable one. All of these situations have something important in common: the investor holds today an asset with an uncertain future value. This is very clear in the case of the stock. For the insurance company, the policy sold may or may not be triggered by the underlying event covered. In the case of a mortgage, our decision today to enter into this refinancing agreement will change (for better or for worse) the future repayments. So randomness plays a crucial role in the valuation of current products held by the investor, the insurance company and the home owner.

To model these situations a mathematician would now define the value of a risky position X to be a function on the probability space $(\Omega, \mathcal{F}, P)$; this function is called a *random variable*. We leave for the moment the range of X (i.e. its possible values) unspecified. Most of the modelling of a risky position X concerns its *distribution function* $F_X(x) = P(X \leqslant x)$: the probability that by the end of the period under consideration the value of the risk X is less than or equal to a given number x. Several risky positions would then be denoted by a random vector $(X_1, \dots, X_d)$, also written in bold face as $\boldsymbol{X}$; time can be introduced, leading to the notion of random (or so-called stochastic) processes, usually written $(\boldsymbol{X}_t)$. Throughout this book we will encounter many such processes, which serve as essential building blocks in the mathematical description of risk.

We therefore expect the reader to be at ease with basic notation, terminology and results from elementary *probability and statistics*, the branch of mathematics dealing with *stochastic* models and their application to the real world. The word "stochastic" is derived from the Greek "stochazesthai", the art of guessing, or "stochastikos", meaning skilled at aiming ("stochos" being a target). In discussing stochastic methods for risk management we hope to emphasize the skill aspect rather than the guesswork.

1.1.2 Financial Risk

In this book we discuss risk in the context of finance and insurance (although many of the tools introduced are applicable well beyond this context). We start by giving a brief overview of the main risk types encountered in the financial industry.

The best-known type of risk is probably *market risk*: the risk of a change in the value of a financial position or portfolio due to changes in the value of the underlying components on which that portfolio depends, such as stock and bond prices, exchange rates, commodity prices, etc. The next important category is *credit risk*: the risk of not receiving promised repayments on outstanding investments such as loans and bonds, because of the "default" of the borrower. A further risk category is *operational risk*: the risk of losses resulting from inadequate or failed internal processes, people and systems, or from external events.

The three risk categories of market, credit and operational risk are the main ones we study in this book, but they do not form an exhaustive list of the full range of possible risks affecting a financial institution, nor are their boundaries always clearly defined. For example, when a corporate bond falls in value this is market risk, but the fall in value is often associated with a deterioration in the credit quality of the issuer, which is related to credit risk. The ideal way forward for a successful handling of financial risk is a *holistic* approach, i.e. an integrated approach taking all types of risk and their interactions into account.

Other important notions of risk are *model risk* and *liquidity risk*. The former is the risk associated with using a misspecified (inappropriate) model for measuring risk. Think, for instance, of using the Black–Scholes model for pricing an exotic option in circumstances where the basic Black–Scholes model assumptions on the underlying securities (such as the assumption of normally distributed returns) are violated. It may be argued that model risk is always present to some degree.

When we talk about liquidity risk we are generally referring to price or market liquidity risk, which can be broadly defined as the risk stemming from the lack of marketability of an investment that cannot be bought or sold quickly enough to prevent or minimize a loss. Liquidity can be thought of as "oxygen for a healthy market"; a market requires it to function properly but most of the time we are not aware of its presence. Its absence, however, is recognized immediately, with often disastrous consequences.

In banking, there is also the concept of *funding liquidity risk*, which refers to the ease with which institutions can raise funding to make payments and meet withdrawals as they arise. The management of funding liquidity risk tends to be a specialist activity of bank treasuries (see, for example, Choudhry 2012) rather than trading-desk risk managers and is not a subject of this book. However, funding liquidity and market liquidity can interact profoundly in periods of financial stress. Firms that have problems obtaining funding may sell assets in fire sales to raise cash, and this in turn can contribute to market illiquidity, depressing prices, distorting the valuation of assets on balance sheets and, in turn, making funding even more difficult to obtain; this phenomenon has been described as a liquidity spiral (Brunnermeier and Pedersen 2009).

In insurance, a further risk category is *underwriting risk*: the risk inherent in insurance policies sold. Examples of risk factors that play a role here are changing patterns of natural catastrophes, changes in demographic tables underlying (long-dated) life products, political or legal interventions, or customer behaviour (such as lapsation).

1.1.3 Measurement and Management

Much of this book is concerned with techniques for the statistical measurement of risk, an activity which is part of the process of managing risk, as we attempt to clarify in this section.

Risk measurement. Suppose we hold a portfolio consisting of d underlying investments with respective weights $w_1, \dots, w_d$, so that the change in value of the portfolio over a given holding period (the so-called profit and loss, or P&L) can be written as $X = \sum_{i=1}^{d} w_i X_i$, where X_i denotes the change in value of the ith investment. Measuring the risk of this portfolio essentially consists of determining its distribution function $F_X(x) = P(X \leqslant x)$, or functionals describing this distribution function such as its mean, variance or 99th percentile.

In order to achieve this, we need a properly calibrated *joint* model for the underlying random vector of investments $(X_1, \dots, X_d)$, so statistical methodology has an important role to play in risk measurement; based on historical observations and given a specific model, a statistical estimate of the distribution of the change in value of a position, or one of its functionals, is calculated. In Chapter 2 we develop a detailed framework framework for risk measurement. As we shall see—and this is indeed a main theme throughout the book—this is by no means an easy task with a unique solution.

It should be clear from the outset that good risk measurement is essential. Increasingly, the clients of financial institutions demand objective and detailed information on the products that they buy, and firms can face legal action when this information is found wanting. For any product sold, a proper quantification of the underlying risks needs to be explicitly made, allowing the client to decide whether or not the product on offer corresponds to his or her risk appetite; the 2007–9 crisis saw numerous violations of this basic principle. For more discussion of the importance of the quantitative approach to risk, see Section 1.5.

Risk management. In a very general answer to the question of what risk management is about, Kloman (1990) writes:

> To many analysts, politicians, and academics it is the management of environmental and nuclear risks, those technology-generated macrorisks that appear to threaten our existence. To bankers and financial officers it is the sophisticated use of such techniques as currency hedging and interest-rate swaps. To insurance buyers or sellers it is coordination of insurable risks and the reduction of insurance costs. To hospital administrators it may mean "quality assurance". To safety professionals it is reducing accidents and injuries. In summary, risk management is *a discipline for living with the possibility that future events may cause adverse effects*.

The last phrase in particular (the emphasis is ours) captures the general essence of risk management: it is about ensuring *resilience* to future events. For a financial institution one can perhaps go further. A financial firm's attitude to risk is not passive and defensive; a bank or insurer actively and willingly takes on risk, because it seeks a return and this does not come without risk. Indeed, risk management can be seen as the core competence of an insurance company or a bank. By using its expertise, market position and capital structure, a financial institution can manage risks by repackaging or bundling them and transferring them to markets in customized ways.

The management of risk at financial institutions involves a range of tasks. To begin with, an enterprise needs to determine the capital it should hold to absorb losses, both for regulatory and economic capital purposes. It also needs to manage the risk on its books. This involves ensuring that portfolios are well diversified and optimizing portfolios according to risk–return considerations. The risk profile of the portfolio can be altered by hedging exposures to certain risks, such as interest-rate or foreign-exchange risk, using derivatives. Alternatively, some risks can be repackaged and sold to investors in a process known as securitization; this has been applied to both insurance risks (weather derivatives and longevity derivatives) and credit risks (mortgage-backed securities, collateralized debt obligations). Firms that use derivatives need to manage their derivatives books, which involves the tasks of pricing, hedging and managing collateral for such trades. Finally, financial institutions need to manage their counterparty credit risk exposures to important trading partners; these arise from bilateral, over-the-counter derivatives trades, but they are also present, for example, in reinsurance treaties.

We also note that the discipline of risk management is very much the core competence of an actuary. Indeed, the Institute and Faculty of Actuaries has used the following definition of the actuarial profession:

> Actuaries are respected professionals whose innovative approach to making business successful is matched by a responsibility to the public interest. Actuaries identify solutions to financial problems. They manage assets and liabilities by analysing past events, assessing the present risk involved and modelling what could happen in the future.

Actuarial organizations around the world have collaborated to create the Chartered Enterprise Risk Actuary qualification to show their commitment to establishing best practice in risk management.

1.2 A Brief History of Risk Management

In this section we treat the historical development of risk management by sketching some of the innovations and some of the events that have shaped modern risk management for the financial industry. We also describe the more recent development of regulation in the industry, which has, to some extent, been a process of reaction to a series of incidents and crises.

1.2.1 From Babylon to Wall Street

Although risk management has been described as "one of the most important innovations of the 20th century" by Steinherr (1998), and most of the story we tell is relatively modern, some concepts that are used in modern risk management, and in derivatives in particular, have been around for longer. In our selective account we stress the example of financial derivatives as these have played a role in many of the events that have shaped modern regulation and increased the complexity of the risk-management challenge.

The ancient world to the twentieth century. A derivative is a financial instrument derived from an underlying asset, such as an option, future or swap. For example, a European call option with strike K and maturity T gives the holder the right, but not the obligation, to obtain from the seller at maturity the underlying security for a price K; a European put option gives the holder the right to dispose of the underlying at a price K.

Dunbar (2000) interprets a passage in the Code of Hammurabi from Babylon of 1800 BC as being early evidence of the use of the option concept to provide financial cover in the event of crop failure. A very explicit mention of options appears in Amsterdam towards the end of the seventeenth century and is beautifully narrated by Joseph de la Vega in his 1688 *Confusión de Confusiones*, a discussion between a lawyer, a trader and a philosopher observing the activity on the Beurs of Amsterdam. Their discussion contains what we now recognize as European call and put options and a description of their use for investment as well as for risk management—it even includes the notion of short selling. In an excellent recent translation (de la Vega 1996) we read:

> If I may explain "opsies" [further, I would say that] through the payment of the premiums, one hands over values in order to safeguard one's stock or to obtain a profit. One uses them as sails for a happy voyage during a beneficent conjuncture and as an anchor of security in a storm.

After this, de la Vega continues with some explicit examples that would not be out of place in any modern finance course on the topic.

Financial derivatives in general, and options in particular, are not so new. Moreover, they appear here as instruments to manage risk, "anchors of security in a storm", rather than as dangerous instruments of speculation, the "wild beasts of finance" (Steinherr 1998), that many believe them to be.

Academic innovation in the twentieth century. While the use of risk-management ideas such as derivatives can be traced further back, it was not until the late twentieth century that a theory of valuation for derivatives was developed. This can be seen as perhaps the most important milestone in an age of academic developments in the general area of quantifying and managing financial risk.

Before the 1950s, the desirability of an investment was mainly equated to its return. In his groundbreaking publication of 1952, Harry Markowitz laid the foundation of the theory of portfolio selection by mapping the desirability of an investment onto a risk–return diagram, where risk was measured using standard deviation (see Markowitz 1952, 1959). Through the notion of an *efficient frontier* the portfolio manager could optimize the return for a given risk level. The following decades saw explosive growth in risk-management methodology, including such ideas as the Sharpe ratio, the Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT). Numerous extensions and refinements that are now taught in any MBA course on finance followed.

The famous Black–Scholes–Merton formula for the price of a European call option appeared in 1973 (see Black and Scholes 1973). The importance of this formula was underscored in 1997 when the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel was awarded to Robert Merton and Myron Scholes (Fischer Black had died some years earlier) "for a new method to determine the value of derivatives".

In the final two decades of the century the mathematical finance literature developed rapidly, and many ideas found their way into practice. Notable contributions include the pioneering papers by Harrison and Kreps (1979) and Harrison and Pliska (1981) clarifying the links between no-arbitrage pricing and martingale theory. A further example is the work on the term structure of interest rates by Heath, Jarrow and Morton (1992). These and other papers elaborated the mathematical foundations of financial mathematics. Textbooks on stochastic integration and Itô calculus became part of the so-called quant's essential reading and were, for a while, as likely to be seen in the hands of a young investment banker as the *Financial Times*.

Growth of markets in the twentieth century. The methodology developed for the rational pricing and hedging of financial derivatives changed finance. The "wizards of Wall Street" (i.e. the mathematical specialists conversant in the new methodology) have had a significant impact on the development of financial markets over the last few decades. Not only did the new option-pricing formula work, it transformed the market. When the Chicago Options Exchange first opened in 1973, fewer than a thousand options were traded on the first day. By 1995, over a million options were changing hands each day, with current nominal values outstanding in the derivatives markets in the tens of trillions. So great was the role played by the Black–Scholes–Merton formula in the growth of the new options market that, when the American stock market crashed in 1987, the influential business magazine *Forbes* attributed the blame squarely to that one formula. Scholes himself has said that it was not so much the formula that was to blame, but rather that market traders had not become sufficiently sophisticated in using it.

Along with academic innovation, developments in information technology (IT) also helped lay the foundations for an explosive growth in the volume of new risk-management and investment products. This development was further aided by worldwide deregulation in the 1980s. Important additional factors contributing to an increased demand for risk-management skills and products were the oil crises of the 1970s and the 1970 abolition of the Bretton Woods system of fixed exchange rates. Both energy prices and foreign exchange risk became highly volatile risk factors and customers required products to hedge them. The 1933 Glass–Steagall Act—passed in the US in the aftermath of the 1929 Depression to prohibit commercial banks from underwriting insurance and most kinds of securities—indirectly paved the way for the emergence of investment banks, hungry for new business. Glass–Steagall was replaced in 1999 by the Financial Services Act, which repealed many of the former's key provisions, although the 2010 Dodd–Frank Act, passed in the aftermath of the 2007–9 financial crisis, appears to mark an end to the trend of deregulation.

Disasters of the 1990s. In January 1992 the president of the New York Federal Reserve, E. Gerald Corrigan, speaking at the Annual Mid-Winter Meeting of the New York State Bankers Association, said:

> You had all better take a very, very hard look at off-balance-sheet activities. The growth and complexity of [these] activities and the nature of the credit settlement risk they entail should give us cause for concern.... I hope this sounds like a warning, because it is. Off-balance-sheet activities [i.e. derivatives] have a role, but they must be managed and controlled carefully and they must be understood by top management as well as by traders and rocket scientists.

Corrigan was referring to the growing volume of derivatives in banks' trading books and the fact that, in many cases, these did not appear as assets or liabilities on the balance sheet. His words proved prescient.

On 26 February 1995 Barings Bank was forced into administration. A loss of £700 million ruined the oldest merchant banking group in the UK (established in 1761). Besides numerous operational errors (violating every qualitative guideline in the risk-management handbook), the final straw leading to the downfall of Barings was a so-called straddle position on the Nikkei held by the bank's Singapore-based trader Nick Leeson. A straddle is a short position in a call and a put with the same strike—such a position allows for a gain if the underlying (in this case the Nikkei index) does not move too far up or down. There is, however, considerable loss potential if the index moves down (or up) by a large amount, and this is precisely what happened when the Kobe earthquake occurred.

Three years later, Long-Term Capital Management (LTCM) became another prominent casualty of losses due to derivatives trading when it required a $3.5 billion payout to prevent collapse, a case made all the more piquant by the fact that

Myron Scholes and Robert Merton were principals at the hedge fund. Referring to the Black–Scholes formula, an article in the *Observer* newspaper asked: “Is this really the key to future wealth? Win big, lose bigger.”

There were other important cases in this era, leading to a widespread discussion of the need for increased regulation, including Metallgesellschaft in 1993 (speculation on oil prices using derivatives) and Orange County in 1994 (speculation on interest rates using derivatives).

In the life insurance industry, Equitable Life, the world’s oldest mutual insurer, provided a case study of what can happen when the liabilities arising from insurance products with embedded options are not properly hedged. Prior to 1988, Equitable Life had sold pension products that offered the option of a guaranteed annuity rate at maturity of the policy. The guarantee rate of 7% had been set in the 1970s when inflation and annuity rates were high, but in 1993 the current annuity rate fell below the guarantee rate and policyholders exercised their options. Equitable Life had not been hedging the option and it quickly became evident that they were faced with an enormous increase in their liabilities; the Penrose Report (finally published in March 2004) concluded that Equitable Life was underfunded by around £4.5 billion by 2001. It was the policyholders who suffered when the company reneged on their pension promises, although many of the company’s actions were later ruled unlawful and some compensation from the public purse was agreed. However, this case provides a good illustration of the need to regulate the capital adequacy of insurers to protect policyholders.

The turn of the century. The end of the twentieth century proved to be a pivotal moment for the financial system worldwide. From a value of around 1000 in 1996, the Nasdaq index quintupled to a maximum value of 5408.62 on 10 March 2000 (which remains unsurpassed as this book goes to press). The era 1996–2000 is now known as the dot-com bubble because many of the firms that contributed to the rise in the Nasdaq belonged to the new internet sector.

In a speech before the American Enterprise Institute on 5 December 1996, Alan Greenspan, chairman of the Federal Reserve from 1987 to 2006, said, “But how do we know when *irrational exuberance* has unduly escalated assets, which then become subject to prolonged contractions as they have in Japan over the past decade?” The term irrational exuberance seemed to perfectly describe the times. The Dow Jones Industrial Average was also on a historic climb, breaking through the 10 000 barrier on 29 March 1999, and prompting books with titles like *Dow 40 000: Strategies for Profiting from the Greatest Bull Market in History*. It took four years for the bubble to burst, but from its March 2000 maximum the Nasdaq plummeted to half of its value within a year and tested the 1000 barrier in late 2002. Equity indices fell worldwide, although markets recovered and began to surge ahead again from 2004.

The dot-com bubble was in many respects a conventional asset bubble, but it was also during this period that the seeds of the next financial crisis were being sown. Financial engineers had discovered the magic of securitization: the bundling and repackaging of many risks into securities with defined risk profiles that could be sold to potential investors. While the idea of transferring so-called tranches of a pool of risks to other risk bearers was well known to the insurance world, it was now being applied on a massive scale to credit-risky assets, such as mortgages, bonds, credit card debt and even student loans (see Section 12.1.1 for a description of the tranching concept).

In the US, the subprime lending boom to borrowers with low credit ratings fuelled the supply of assets to securitize and a market was created in mortgage-backed securities (MBSs). These in turn belonged to the larger pool of assets that were available to be transformed into collateralized debt obligations (CDOs). The banks originating these credit derivative products had found a profitable business turning poor credit risks into securities. The volume of credit derivatives ballooned over a very short period; the CDO market accounted for almost $3 trillion in nominal terms by 2008 but this was dwarfed by the nominal value of the credit default swap (CDS) market, which stood at about $30 trillion.

Credit default swaps, another variety of credit derivative, were originally used as instruments for hedging large corporate bond exposures, but they were now increasingly being used by investors to speculate on the changing credit outlook of companies by adopting so-called naked positions (see Section 10.1.4 for more explanation). Although the actual economic value of CDS and CDO markets was actually smaller (when the netting of cash flows is considered), these are still huge figures when compared with world gross domestic product (GDP), which was of the order of $60 trillion at that time.

The consensus was that all this activity was a good thing. Consider the following remarks made by the then chairman of the Federal Reserve, Alan Greenspan, before the Council on Foreign Relations in Washington DC on 19 November 2002 (Greenspan 2002):

> More recently, instruments . . . such as credit default swaps, collateralized debt obligations and credit-linked notes have been developed and their use has grown rapidly in recent years. The result? Improved credit risk management together with more and better risk-management tools appear to have significantly reduced loan concentrations in telecommunications and, indeed, other areas and the associated stress on banks and other financial institutions. . . . It is noteworthy that payouts in the still relatively small but rapidly growing market in credit derivatives have been proceeding smoothly for the most part. Obviously this market is still too new to have been tested in a widespread down-cycle for credit, but, to date, it appears to have functioned well.

As late as April 2006 the International Monetary Fund (IMF) wrote in its Global Financial Stability Report that:

> There is a growing recognition that the dispersion of credit risk by banks to a broader and more diverse group of investors, rather than warehousing such risks on their balance sheets, has helped to make the banking and overall financial system more resilient. . . . The improved

> resilience may be seen in fewer bank failures and more consistent credit provision. Consequently, the commercial banks, a core system of the financial system, may be less vulnerable today to credit or economic shocks.

It has to be said that the same IMF report also warned about possible vulnerabilities, and the potential for market disruption, if these credit instruments were not fully understood.

One of the problems was that not all of the risk from CDOs was being dispersed to outside investors as the IMF envisaged. As reported in Acharya et al. (2009), large banks were holding on to a lot of it themselves:

> These large, complex financial institutions ignored their own business models of securitization and chose not to transfer credit risk to other investors. Instead they employed securitization to manufacture and retain tail risk that was systemic in nature and inadequately capitalized.... Starting in 2006, the CDO group at UBS noticed that their risk-management systems treated AAA securities as essentially risk-free even though they yielded a premium (the proverbial free lunch). So they decided to hold onto them rather than sell them! After holding less than $5 billion of them in 02/06, the CDO desk was warehousing a staggering $50 billion in 09/07.... Similarly, by late summer of 2007, Citigroup had accumulated over $55 billion of AAA-rated CDOs.

On the eve of the crisis many in the financial industry seemed unconcerned. AIG, the US insurance giant, had become heavily involved in underwriting MBS and CDO risk by selling CDS protection through its AIG Financial Products arm. In August 2007 the chief executive officer of AIG Financial Products is quoted as saying:

> It is hard for us, without being flippant, to even see a scenario within any kind of realm of reason that would see us losing one dollar in any of these transactions.

The financial crisis of 2007–9. After a peak in early 2006, US house prices began to decline in 2006 and 2007. Subprime mortgage holders, experiencing difficulties in refinancing their loans at higher interest rates, defaulted on their payments in increasing numbers. Starting in late 2007 this led to a rapid reassessment of the riskiness of securitizations and to losses in the value of CDO securities. Banks were forced into a series of dramatic *write-downs* of the value of these assets on their balance sheets, and the severity of the impending crisis became apparent.

Reflecting on the crisis in his article "It doesn't take Nostradamus" in the 2008 issue of *Economists' Voice*, Nobel laureate Joseph E. Stiglitz recalled the views he expressed in 1992 on securitization and the housing market:

> The question is, has the growth of securitization been a result of more efficient transaction technologies or an unfounded reduction in concern about the importance of screening loan applicants? It is perhaps too

> early to tell, but we should at least entertain the possibility that it is the latter rather than the former.

He also wrote:

> At the very least, the banks have demonstrated ignorance of two very basic aspects of risk: (a) the importance of correlation ... [and] (b) the possibility of price declines.

These "basic aspects of risk", which would appear to belong in a Banking 101 class, plunged the world's economy into its most serious crisis since the late 1920s. Salient events included the demise of such illustrious names as Bear Stearns (which collapsed and was sold to JPMorgan Chase in March 2008) and Lehman Brothers (which filed for Chapter 11 bankruptcy on 15 September 2008). The latter event in particular led to worldwide panic. As markets tumbled and liquidity vanished it was clear that many banks were on the point of collapse. Governments had to bail them out by injecting capital or by acquiring their distressed assets in arrangements such as the US Troubled Asset Relief Program.

AIG, which had effectively been insuring the default risk in securitized products by selling CDS protection, got into difficulty when many of the underlying securities defaulted; the company that could not foresee itself "losing one dollar in any of these transactions" required an emergency loan facility of $85 billion from the Federal Reserve Bank of New York on 16 September 2008. In the view of George Soros (2009), CDSs were "instruments of destruction" that should be outlawed:

> Some derivatives ought not to be allowed to be traded at all. I have in mind credit default swaps. The more I've heard about them, the more I've realised they're truly toxic.

Much has been written about these events, and this chapter's Notes and Comments section contains a number of references. One strand of the commentary that is relevant for this book is the apportioning of a part of the blame to mathematicians (or financial engineers); the failure of valuation models for complex securitized products made them an easy target. Perhaps the most publicized attack came in a blog by Felix Salmon (*Wired Magazine*, 23 February 2009) under the telling title "Recipe for disaster: the formula that killed Wall Street". The formula in question was the *Gauss copula*, and its application to credit risk was attributed to David Li. Inspired by what he had learned on an actuarial degree, Li proposed that a tool for modelling dependent lifetimes in life insurance could be used to model correlated default times in bond portfolios, thus providing a framework for the valuation and risk management of CDOs, as we describe in Chapter 12.

While an obscure formula with a strange name was a gift for bloggers and newspaper headline writers, even serious regulators joined in the chorus of criticism of mathematics. The Turner Review of the global banking crisis (Lord Turner 2009) has a section entitled "Misplaced reliance on sophisticated mathematics" (see Section 1.3.3 for more on this theme). But this reliance on mathematics was only one factor in the crisis, and certainly not the most important. Mathematicians had also warned well beforehand that the world of securitization was being built on shaky model foundations that were difficult to calibrate (see, for example, Frey, McNeil and Nyfeler 2001). It was also abundantly clear that political shortsightedness, the greed of market participants and the slow reaction of regulators had all contributed in very large measure to the scale of the eventual calamity.

Recent developments and concerns. New threats to the financial system emerge all the time. The financial crisis of 2007–9 led to recession and sovereign debt crises. After the wave of bank bailouts, concerns about the solvency of banks were transformed into concerns about the abilities of countries to service their own debts. For a while doubts were cast on the viability of the eurozone, as it seemed that countries might elect to, or be forced to, exit the single currency.

On the more technical side, the world of high-frequency trading has raised concerns among regulators, triggered by such events as the Flash Crash of 6 May 2010. In this episode, due to "computer trading gone wild", the Dow Jones lost around 1000 points in a couple of minutes, only to be rapidly corrected. High-frequency trading is a form of algorithmic trading in which trades are executed by computers according to algorithms in fractions of a second. One notable casualty of algorithmic trading was Knight Capital, which lost $460 million due to trading errors on 1 August 2012. Going forward, it is clear that vigilance is required concerning the risks arising from the deployment of new technologies and their systemic implications.

Indeed, *systemic risk* is an ongoing concern to which we have been sensitized by the financial crisis. This is the risk of the collapse of the entire financial system due to the propagation of financial stress through a network of participants. When Lehman Brothers failed there was a moment when it seemed possible that there could be a catastrophic cascade of defaults of banks and other firms. The interbank lending market had become dysfunctional, asset prices had plummeted and the market for any form of debt was highly illiquid. Moreover, the complex chains of relationships in the CDS markets, in which the same credit-risky assets were referenced in a large volume of bilateral payment agreements, led to the fear that the default of a further large player could cause other banks to topple like dominoes.

The concerted efforts of many governments were successful in forestalling the Armageddon scenario. However, since the crisis, research into financial networks and their embedded systemic risks has been an important research topic. These networks are complex, and as well as banks and insurance companies they contain members of a "shadow banking system" of hedge funds and structured investment vehicles, which are largely unregulated. One important theme is the identification of so-called systemically important financial institutions (SIFI) whose failure might cause a systemic crisis.

1.2.2 The Road to Regulation

There is no doubt that regulation goes back a long way, at least to the time of the Venetian banks and the early insurance enterprises sprouting in London's coffee shops in the eighteenth century. In those days there was more reliance on self-regulation or local regulation, but rules were there. However, the key developments that led to the present prudential regulatory framework in financial services are a very much more recent story.

The main aim of modern prudential regulation has been to ensure that financial institutions have enough capital to withstand financial shocks and remain solvent. Robert Jenkins, a member of the Financial Policy Committee of the Bank of England, was quoted in the *Independent* on 27 April 2012 as saying:

> Capital is there to absorb losses from risks we understand and risks we may not understand. Evidence suggests that neither risk-takers nor their regulators fully understand the risks that banks sometimes take. That's why banks need an appropriate level of loss absorbing equity.

Much of the regulatory drive originated from the Basel Committee of Banking Supervision. This committee was established by the central-bank governors of the Group of Ten at the end of 1974. The Group of Ten is made up of (oddly) eleven industrial countries that consult and cooperate on economic, monetary and financial matters. The Basel Committee does not possess any formal supranational supervising authority, and hence its conclusions do not have legal force. Rather, it formulates broad supervisory standards and guidelines and recommends statements of best practice in the expectation that individual authorities will take steps to implement them through detailed arrangements—statutory or otherwise—that are best suited to their own national system. The summary below is brief. Interested readers can consult, for example, Tarullo (2008) for further details, and should also see this chapter's Notes and Comments section.

The first Basel Accord. The first Basel Accord on Banking Supervision (Basel I, from 1988) took an important step towards an international minimum capital standard. Its main emphasis was on credit risk, by then clearly the most important source of risk in the banking industry. In hindsight, however, Basel I took an approach that was fairly coarse and measured risk in an insufficiently differentiated way. In measuring credit risk, claims were divided into three crude categories according to whether the counterparties were governments, regulated banks or others. For instance, the risk capital charge for a loan to a corporate borrower was five times higher than for a loan to an Organisation for Economic Co-operation and Development (OECD) bank. The risk weighting for all corporate borrowers was identical, independent of their credit rating. The treatment of derivatives was also considered unsatisfactory.

The birth of VaR. In 1993 the G-30 (an influential international body consisting of senior representatives of the private and public sectors and academia) published a seminal report addressing, for the first time, so-called off-balance-sheet products, like derivatives, in a systematic way. Around the same time, the banking industry clearly saw the need for proper measurement of the risks stemming from these new products. At JPMorgan, for instance, the famous Weatherstone 4.15 report asked for a one-day, one-page summary of the bank's market risk to be delivered to the chief executive officer in the late afternoon (hence "4.15"). Value-at-risk (VaR) as a market risk measure was born and the JPMorgan methodology, which became known as RiskMetrics, set an industry-wide standard.

In a highly dynamic world with round-the-clock market activity, the need for instant market valuation of trading positions (known as *marking-to-market*) became a necessity. Moreover, in markets where so many positions (both long and short) were written on the same underlyings, managing risks based on simple aggregation of nominal positions became unsatisfactory. Banks pushed to be allowed to consider *netting* effects, i.e. the compensation of long versus short positions on the same underlying.

In 1996 an important amendment to Basel I prescribed a so-called *standardized* model for market risk, but at the same time allowed the bigger (more sophisticated) banks to opt for an *internal* VaR-based model (i.e. a model developed in house). Legal implementation was to be achieved by the year 2000. The coarseness problem for credit risk remained unresolved and banks continued to claim that they were not given enough incentives to diversify credit portfolios and that the regulatory capital rules currently in place were far too risk insensitive. Because of overcharging on the regulatory capital side of certain credit positions, banks started shifting business away from certain market segments that they perceived as offering a less attractive risk–return profile.

The second Basel Accord. By 2001 a consultative process for a new Basel Accord (Basel II) had been initiated; the basic document was published in June 2004. An important aspect was the establishment of the three-pillar system of regulation: Pillar 1 concerns the quantification of regulatory capital; Pillar 2 imposes regulatory oversight of the modelling process, including risks not considered in Pillar 1; and Pillar 3 defines a comprehensive set of disclosure requirements.

Under Pillar 1 the main theme of Basel II was credit risk, where the aim was to allow banks to use a finer, more risk-sensitive approach to assessing the risk of their credit portfolios. Banks could opt for an *internal-ratings-based* approach, which permitted the use of internal or external credit-rating systems wherever appropriate.

The second important theme of Basel II at the level of Pillar 1 was the consideration of operational risk as a new risk class. A basic premise of Basel II was that the overall size of regulatory capital throughout the industry should stay unchanged under the new rules. Since the new rules for credit risk were likely to reduce the credit risk charge, this opened the door for operational risk, defined as the risk of losses resulting from inadequate or failed internal processes, people and systems or from external events; this definition included legal risk but excluded reputational and strategic risk.

Mainly due to the financial crisis of 2007–9, implementation of the Basel II guidelines across the globe met with delays and was rather spread out in time. Various further amendments and additions to the content of the original 2004 document were made. One important criticism of Basel II that emerged from the crisis was that it was inherently *procyclical*, in that it forced firms to take action to increase their capital ratios at exactly the wrong point in the business cycle, when their actions had a negative impact on the availability of liquidity and made the situation worse (see Section 1.3.3 for more discussion on this).

Basel 2.5. One clear lesson from the crisis was that modern products like CDOs had opened up opportunities for regulatory arbitrage by transferring credit risk from the capital-intensive banking book (or loan book) to the less-capitalized trading book. Some enhancements to Basel II were proposed in 2009 with the aim of addressing the build-up of risk in the trading book that was evident during the crisis. These enhancements, which have come to be known as Basel 2.5, include a *stressed VaR* charge, based on calculating VaR from data for a twelve-month period of market turmoil, and the so-called *incremental risk charge*, which seeks to capture some of the default risk in trading book positions; there were also specific new rules for certain securitizations.

The third Basel Accord. In view of the failure of the Basel rules to prevent the 2007–9 crisis, the recognized deficiencies of Basel II mentioned above, and the clamour from the public and from politicians for regulatory action to make banks and the banking system safer, it is no surprise that attention quickly shifted to Basel III.

In 2011 a series of measures was proposed that would extend Basel II (and 2.5) in five main areas:

(1) measures to increase the quality and amount of bank capital by changing the definition of key capital ratios and allowing countercyclical adjustments to these ratios in crises;

(2) a strengthening of the framework for counterparty credit risk in derivatives trading, with incentives to use central counterparties (exchanges);

(3) the introduction of a leverage ratio to prevent excessive leverage;

(4) the introduction of various ratios that ensure that banks have sufficient funding liquidity;

(5) measures to force systemically important banks to have even higher capacity to absorb losses.

Most of the new rules will be phased in progressively, with a target end date of 2019, although individual countries may impose stricter guidelines with respect to both schedule and content.

Parallel developments in insurance regulation. The insurance industry worldwide has also been subject to increasing risk regulation in recent times. However, here the story is more fragmented and there has been much less international coordination of efforts. The major exception has been the development of the Solvency II framework in the European Union, a process described in more detail below. As the most detailed and model intensive of the regulatory frameworks proposed, it serves as our main reference point for insurance regulation in this book. The development of the Solvency II framework is overseen by the European Insurance and Occupational Pensions Authority (EIOPA; formerly the Committee of European Insurance and Occupational Pensions Supervisors (CEIOPS)), but the implementation in individual countries is a matter for national regulators, e.g. the Prudential Regulatory Authority in the UK.

In the US, insurance regulation has traditionally been a matter for state governments. The National Association of Insurance Commissioners (NAIC) provides support to insurance regulators from the individual states, and helps to promote the development of accepted regulatory standards and best practices; it is up to the individual states whether these are passed into law, and if so in what form. In the early 1990s the NAIC promoted the concept of risk-based capital for insurance companies as a response to a number of insolvencies in the preceding years; the NAIC describes risk-based capital as "a method of measuring the minimum amount of capital appropriate for a reporting entity to support its overall business operations in consideration of its size and profile". The method, which is a rules-based approach rather than a model-based approach, has become the main plank of insurance regulation in the US.

Federal encroachment on insurance supervision has generally been resisted, although this may change due to a number of measures enacted after the 2007–9 crisis in the wide-ranging 2010 Dodd–Frank Act. These include the creation of both the Federal Insurance Office, to "monitor all aspects of the insurance sector", and the Financial Stability Oversight Council, which is "charged with identifying risks to the financial stability of the United States" wherever they may arise in the world of financial services.

The International Association of Insurance Supervisors has been working to foster some degree of international convergence in the processes for regulating the capital adequacy of insurers. They have promoted the idea of the Own Risk and Solvency Assessment (ORSA). This has been incorporated into the Solvency II framework and has also been embraced by the NAIC in the US.

There are also ongoing initiatives that aim to bring about convergence of banking and insurance regulation, particularly with respect to financial conglomerates engaged in both banking and insurance business. The Joint Forum on Financial Conglomerates was established in early 1996 under the aegis of the Basel Committee, the International Association of Insurance Supervisors and the International Organization of Securities Commissions to take forward this work.

From Solvency I to Solvency II. Mirroring the progress in the banking sector, Solvency II is the latest stage in a process of regulatory evolution from simple and crude rules to a more risk-sensitive treatment of the capital requirements of insurance companies.

The first European Union non-life and life directives on solvency margins appeared around 1970. The solvency margin was defined as an extra capital buffer against unforeseen events such as higher than expected claims levels or unfavourable investment results. However, there were differences in the way that regulation was applied across Europe and there was a desire for more harmonization of regulation and mutual recognition.

Solvency I, which came into force in 2004, is a rather coarse rules-based framework calling for companies to have a minimum guarantee fund (minimal capital) of €3 million, and a solvency margin consisting of 16–18% of non-life premiums together with 4% of the technical provisions for life. This has led to a single robust system that is easy to understand and inexpensive to monitor. However, on the negative side, it is mainly volume based, not explicitly risk based; issues like guarantees, embedded options and the proper matching of assets and liabilities are largely neglected in many countries.

To address these shortcomings, Solvency II was initiated in 2001 with the publication of the influential Sharma Report. While the Solvency II directive was adopted by the Council of the European Union and the European Parliament in November 2009, implementation of the framework is not expected until 1 January 2016. The process of refinement of the framework is managed by EIOPA, and one of the features of this process has been a series of quantitative impact studies in which companies have effectively tried out aspects of the proposals and information has been gathered with respect to the impact and practicability of the new regulations.

The goal of the Solvency II process is that the new framework should strengthen the capital adequacy regime by reducing the possibilities of consumer loss or market disruption in insurance; Solvency II therefore has both policyholder-protection and financial-stability motives. Moreover, it is also an aim that the harmonization of regulation in Europe should promote deeper integration of the European Union insurance market and the increased competitiveness of European insurers. A high-level description of the Solvency II framework is given in Section 1.3.2.

The Swiss Solvency Test (SST). Special mention should be made of Switzerland, which has already developed and implemented its own principles-based risk capital regulation for the insurance industry. The SST has been in force since 1 January 2011. It follows similar principles to Solvency II but differs in some details of its treatment of different types of risk; it also places more emphasis on the development of internal models. The implementation of the SST falls under the remit of the Swiss Financial Markets Supervisory Authority, a body formed in 2007 from the merger of the banking and insurance supervisors, which has statutory authority over banks, insurers, stock exchanges, collective investment schemes and other entities.

1.3 The Regulatory Framework

This section describes in more detail the framework that has emerged from the Basel process and the European Union solvency process.

1.3.1 The Basel Framework

As indicated in Section 1.2.2, the Basel framework should be regarded as the product of an evolutionary process. As this book goes to press, the Basel II and Basel 2.5 proposals have been implemented in many developed countries (with some variations in detail), while the proposals of Basel III are still being debated and refined. We sketch the framework as currently implemented, before indicating some of the proposed changes and additions to the framework in Basel III.

The three-pillar concept. A key feature of the Basel framework is the three-pillar concept, as is apparent from the following statement summarizing the Basel philosophy, which accompanied the original Basel II publication (Basel Committee on Banking Supervision 2004):

> The Basel II Framework sets out the details for adopting more risk-sensitive minimum capital requirements [Pillar 1] for banking organizations. The new framework reinforces these risk-sensitive requirements by laying out principles for banks to assess the adequacy of their capital and for supervisors to review such assessments to ensure banks have adequate capital to support their risks [Pillar 2]. It also seeks to strengthen market discipline by enhancing transparency in banks' financial reporting [Pillar 3]. The text that has been released today reflects the results of extensive consultations with supervisors and bankers worldwide. It will serve as the basis for national rule-making and approval processes to continue and for banking organizations to complete their preparations for the new Framework's implementation.

Under *Pillar 1*, banks are required to calculate a *minimum capital charge*, referred to as regulatory capital. There are separate Pillar 1 capital charges for credit risk in the banking book, market risk in the trading book and operational risk, which are considered to be the main quantifiable risks. Most banks use internal models based on VaR methodology to compute the capital charge for market risk. For credit risk and operational risk banks may choose between several approaches of increasing risk sensitivity and complexity, some details of which are discussed below.

Pillar 2 recognizes that any quantitative approach to risk management should be embedded in a properly functioning corporate governance structure. Best-practice risk management imposes constraints on the organization of the institution, i.e. the board of directors, management, employees, and internal and external audit processes. In particular, the board of directors assumes the ultimate responsibility for oversight of the risk landscape and the formulation of the company's risk appetite. Through Pillar 2, also referred to as the *supervisory review process*, local regulators review the various checks and balances that have been put in place. Under Pillar 2, residual quantifiable risks that are not included in Pillar 1, such as interest-rate risk in the banking book, must be considered and *stress tests* of a bank's capital adequacy must be performed. The aim is to ensure that the bank holds capital in line with its true economic loss potential, a concept known as *economic capital*.

Finally, in order to fulfil its promise that increased regulation will increase transparency and diminish systemic risk, clear reporting guidelines on the risks carried by financial institutions are called for. *Pillar 3* seeks to establish *market discipline* through a better public disclosure of risk measures and other information relevant to risk management. In particular, banks will have to offer greater insight into the adequacy of their capitalization.

Credit and market risk; the banking and trading books. Historically, banking activities have been organized around the banking book and the trading book, a split that reflects different accounting practices for different kinds of assets. The banking book contains assets that are *held to maturity*, such as loans; these are typically valued at book value, based on the original cost of the asset. The trading book contains assets and instruments that are *available to trade*; these are generally valued by *marking-to-market* (i.e. using quoted market prices). From a regulatory point of view, credit risk is mainly identified with the banking book and market risk is mainly identified with the trading book.

We have already noted that there are problems with this simple dichotomy and that the Basel 2.5 rules were introduced (partly) to account for the neglect of credit risk (default and rating-migration risk) in the trading book. There are also forms of market risk in the banking book, such as interest-rate risk and foreign-exchange risk. However, the Basel framework continues to observe the distinction between banking book and trading book and we will describe the capital charges in terms of the two books. It is clear that the distinction is somewhat arbitrary and rests on the concept of "available to trade". Moreover, there can be incentives to "switch" or move instruments from one book to the other (particularly from the banking book to the trading book) to benefit from a more favourable capital treatment. This is acknowledged by the Basel Committee in its background discussion of the "Fundamental review of the trading book: a revised market risk framework" (Basel Committee on Banking Supervision 2013a):

> The Committee believes that the definition of the regulatory boundary between the trading book and the banking book has been a source of weakness in the design of the current regime. A key determinant of the boundary has been banks' self-determined intent to trade.... Coupled with large differences in capital requirements against similar types of risk on either side of the boundary, the overall capital framework proved susceptible to arbitrage before and during the crisis.... To reduce the incentives for arbitrage, the Committee is seeking a less permeable boundary with strict limits on switching between books and measures to prevent "capital benefit" in instances where switching is permitted.

The capital charge for the banking book. The credit risk of the banking book portfolio is assessed as the sum of *risk-weighted assets*: that is, the sum of notional exposures weighted by a coefficient reflecting the creditworthiness of the counterparty (the risk weight). To calculate risk weights, banks use either the *standardized* approach or one of the more advanced *internal-ratings-based* (IRB) approaches. The choice of method depends on the size and complexity of the bank, with the larger, international banks having to go for IRB approaches. The capital charge is determined as a fraction of the sum of risk-weighted assets in the portfolio. This fraction, known as the capital ratio, was 8% under Basel II but is already being increased ahead of the planned implementation of Basel III in 2019.

The standardized approach refers to a system that has been in place since Basel I, whereby the risk weights are prescribed by the regulator according to the nature and creditworthiness of the counterparty. For example, there are risk weights for retail loans secured on property (mortgages) and for unsecured retail loans (such as credit cards and overdrafts); there are also different risk weights for corporate and government bonds with different ratings.

Under the more advanced IRB approaches, banks may dispense with the system of fixed risk weights provided by the regulator. Instead, they may make an *internal* assessment of the riskiness of a credit exposure, expressing this in terms of an estimated annualized *probability of default* and an estimated *loss given default*, which are used as inputs in the calculation of risk-weighted assets. The total sum of risk-weighted assets is calculated using formulas specified by the Basel Committee; the formulas also take into account the fact that there is likely to be positive correlation (sometimes called systematic risk) between the credit risks in the portfolio. The use of internally estimated probabilities of default and losses given default allows for increased risk sensitivity in the IRB capital charges compared with the standardized approach. It should be noted, however, that the IRB approaches do not permit fully internal models of credit risk in the banking book; they only permit internal estimation of inputs to a model that has been specified by the regulator.

The capital charge for the trading book. For market risk in the trading book there is also the option of a standardized approach based on a system of risk weights and specific capital charges for different kinds of instrument. However, most major banks elect to use an *internal VaR model approach*, as permitted by the 1996 amendment to Basel I. In Sections 2.2 and 9.2 of this book we give a detailed description of the VaR approach to trading book risk measurement. The approach is based on the estimation of a P&L distribution for a ten-day holding period and the estimation of a particular percentile of this distribution: the 99th percentile of the losses.

A ten-day VaR at 99% of $20 million therefore means that it is estimated that our market portfolio will incur a loss of $20 million *or more* with probability 1% by the end of a ten-day holding period, if the composition remains fixed over this period. The conversion of VaR numbers into an actual capital charge is accomplished by a formula that we discuss in Section 2.3.3.

The VaR calculation is the main component of risk quantification for the trading book, but the 2009 Basel 2.5 revision added further elements (see Basel Committee on Banking Supervision 2012, p. 10), including the following.

Stressed VaR: banks are required to carry out a VaR calculation essentially using the standard VaR methodology but calibrating their models to a historical twelve-month period of significant financial stress.

Incremental risk charge: Since default and rating-migration risk are not generally considered in the standard VaR calculation, banks must calculate an additional charge based on an estimate of the 99.9th percentile of the one-year distribution of losses due to defaults and rating changes. In making this calculation they may use internal models for credit risk (in contrast to the banking book) but must also take into account the market liquidity of credit-risky instruments.

Securitizations: exposures to securitizations in the trading book are subject to a series of new capital charges that bring them more into line with equivalent exposures in the banking book.

The capital charge for operational risk. There are also options of increasing sophistication for assessing operational risk. Under the *basic-indicator* and *standardized* approaches, banks may calculate their operational risk charge using simple formulas based on gross annual income. Under the *advanced measurement approach*, banks may develop internal models. Basel is not prescriptive about the form of these models provided they capture the tail risk of extreme events; most such models are based on historical loss data (internal and external to the firm) and use techniques that are drawn from the actuarial modelling of general insurance losses. We provide more detail in Chapter 13.

New elements of Basel III. Under Basel III there will be a number of significant changes and additions to the Basel framework. While the detail of the new rules may change before final implementation in 2019, the main developments are now clear.

- Banks will need to hold both *more capital* and *better-quality capital* as a function of the risks taken. The "better quality" is achieved though a more restrictive definition of eligible capital (through more stringent definitions of Tier 1 and Tier 2 capital and the phasing out of Tier 3 capital); see Section 2.1.3 for more explanation of capital tiers. The "more" comes from the addition (on top of the minimum ratio of 8%) of a capital conservation buffer of 2.5% of risk-weighted assets, for building up capital in good times to absorb losses under stress, and a countercyclical buffer within the range 0–2.5%, in order to enhance the shock resilience of banks and limit expansion in periods of excessive credit growth. This leads to a total (Tier 1 plus Tier 2) ratio of up to 13%, compared with Basel II's 8%. There will be a gradual phasing in of all these new ratios, with a target date for full implementation of 1 January 2019.

- A *leverage ratio* will be imposed to put a floor under the build-up of excessive leverage in the banking system. Leverage will essentially be measured through the ratio of Tier 1 capital to total assets. A minimum ratio of 3% is currently being tested but the precise definitions may well change as a result of testing experience and bank lobbying. The leverage limit will restrain the size of bank assets, regardless of their riskiness.

- The risk coverage of the system of capital charges is being extended, in particular to include a charge for *counterparty credit risk*. When counterparty credit risk is taken into account in the valuation of over-the-counter derivatives contract, the default-risk-free value has to be adjusted by an amount known as the credit value adjustment (CVA); see Section 17.2 for more explanation. There will now be a charge for changes in CVA.

- Banks will become subject to *liquidity rules*; this is a completely new direction for the Basel framework, which has previously been concerned only with capital adequacy. A *liquidity coverage ratio* will be introduced to ensure that banks have enough highly liquid assets to withstand a period of net cash outflow lasting thirty days. A *net stable funding ratio* will ensure that sufficient funding is available in order to cover long-term commitments (exceeding one year).

It should also be mentioned that under an ongoing review of the trading book, the principle of risk quantification may change from one based on VaR (a percentile) to one based on *expected shortfall* (ES). For a given holding period, the ES at the 99% level, say, is the expected loss given that the loss is higher than the VaR at the 99% level over the same period. ES is a severity measure that always dominates the frequency measure VaR and gives information about the expected size of tail losses; it is also a measure with superior aggregation properties to VaR, as discussed in Section 2.3.5 and Chapter 8 (particularly Sections 8.1 and 8.4.4).

1.3.2 The Solvency II Framework

Below we give an outline of the Solvency II framework, which will come into force in the countries of the European Union on or before 1 January 2016.

Main features. In common with the Basel Accords, Solvency II adopts a three-pillar system, where the first pillar requires the quantification of regulatory capital requirements, the second pillar is concerned with governance and supervision, and the third pillar requires the disclosure of information to the public to improve market discipline by making it easier to compare the risk profiles of companies.

Under Pillar 1, a company calculates its *solvency capital requirement*, which is the amount of capital it should have to ensure that the probability of insolvency over a one-year period is no more than 0.5%—this is often referred to as a confidence level of 99.5%. The company also calculates a smaller *minimum capital requirement*, which is the minimum capital it should have to continue operating without supervisory intervention.

To calculate the capital requirements, companies may use either an *internal model* or a simpler *standard formula* approach. In either case the intention is that a *total balance sheet* approach is taken in which all risks and their interactions are considered. The insurer should have *own funds* (a surplus of assets over liabilities) that exceed both the solvency capital requirement and the minimum capital requirement. The assets and liabilities of the firm should be valued in a *market-consistent* manner.

The supervisory review of the company takes place under Pillar 2. The company must demonstrate that it has a risk-management system in place and that this system is integrated into decision-making processes, including the setting of risk appetite by the company's board, and the formulation of risk limits for different business units. An internal model must pass the "use test": it must be an integral part of the risk-management system and be actively used in the running of the firm. Moreover, a firm must undertake an ORSA as described below.

Market-consistent valuation. In Solvency II the valuation must be carried out according to market-consistent principles. Where possible it should be based on actual *market values*, in a process known as *marking-to-market*. In a Solvency II glossary provided by the Comité Européen des Assurances and the Groupe Consultatif in 2007, market value is defined as:

> The amount for which an asset could be exchanged or a liability settled, between knowledgeable, willing parties in an arm's length transaction, based on observable prices within an active, deep and liquid market which is available to and generally used by the entity.

The concept of market value is related to the concept of *fair value* in accounting, and the principles adopted in Solvency II valuation have been influenced by International Financial Reporting Standards (IFRS) accounting standards. When no relevant market values exist (or when they do not meet the quality criteria described by the concept of an "active, deep and liquid market"), then market-consistent valuation requires the use of models that are calibrated, as far as possible, to be consistent with financial market information, a process known as *marking-to-model*; we discuss these ideas in more detail in Section 2.2.2.

The market-consistent valuation of the liabilities of an insurer is possible when the cash flows paid to policyholders can be fully replicated by the cash flows generated by the so-called matching assets that are held for that purpose; the value of the liability is then given by the value of the *replicating portfolio* of matching assets. However, it is seldom the case that liabilities can be fully replicated and hedged; mortality risk is a good example of a risk factor that is difficult to hedge.

The valuation of the unhedgeable part of a firm's liabilities is carried out by computing the sum of a *best estimate* of these liabilities (basically an expected value) plus an extra *risk margin* to cover some of the uncertainty in the value of the liability. The idea of the risk margin is that a third party would not be willing to take over the unhedgeable liability for a price set at the best estimate but would have to be further compensated for absorbing the additional uncertainty about the true value of the liability.

Standard formula approach. Under this approach an insurer calculates capital charges for different kinds of risk within a series of *modules*. There are modules, for example, for market risk, counterparty default risk, life underwriting risk, non-life underwriting risk and health insurance risk. The risk charges arising from these modules are aggregated to obtain the solvency capital requirement using a formula that involves a set of prescribed correlation parameters (see Section 8.4.2).

Within each module, the approach drills down to fundamental risk factors; for example, within the market-risk module, there are sub-modules relating to interest-rate risk, equity risk, credit-spread risk and other typical market-risk factors. Capital charges are calculated with respect to each risk factor by considering the effect of a series of defined stress scenarios on the value of net assets (assets minus liabilities). The stress scenarios are intended to represent 1-in-200-year events (i.e. events with an annual probability of 0.5%).

The capital charges for each risk factor are aggregated to obtain the module risk charge using a similar kind of formula to the one used at the highest level. Once again, a set of correlations expresses the regulatory view of dependencies between the effects of the fundamental risk factors. The details are complex and run to many pages, but the approach is simple and highly prescriptive.

Internal-model approach. Under this approach firms can develop an internal model for the financial and underwriting risk factors that affect their business; they may then seek regulatory approval to use this model in place of the standard formula. The model often takes the form of a so-called *economic scenario generator* in which risk-factor scenarios for a one-year period are randomly generated and applied to the assets and liabilities to determine the solvency capital requirement. Economic scenario generators vary greatly in their detail, ranging from simple distributional models to more sophisticated dynamic models in discrete or continuous time.

ORSA. In a 2008 Issues Paper produced by CEIOPS, the ORSA is described as follows:

> The entirety of the processes and procedures employed to identify, assess, monitor, manage, and report the short and long term risks a (re)insurance undertaking faces or may face and to determine the own funds necessary to ensure that the undertaking's overall solvency needs are met at all times.

The concept of an ORSA is not unique to Solvency II and a useful alternative definition has been provided by the NAIC in the US on its website:

> In essence, an ORSA is an internal process undertaken by an insurer or insurance group to assess the adequacy of its risk management and current and prospective solvency positions under normal and severe stress scenarios. An ORSA will require insurers to analyze all reasonably foreseeable and relevant material risks (i.e., underwriting, credit, market, operational, liquidity risks, etc.) that could have an impact on an insurer's ability to meet its policyholder obligations.

The Pillar 2 ORSA is distinguished from the Pillar 1 capital calculations in a number of ways. First, the definition makes clear that the ORSA refers to a process, or set of processes, and not simply an exercise in regulatory compliance. Second, each firm's ORSA is its *own* process and is likely to be *unique*, since it is not bound by a common set of rules. In contrast, the standard-formula approach to Pillar 1 is clearly a uniform process for all companies; moreover, firms that seek internal-model approval for Pillar 1 are subject to very similar constraints.

Finally, the ORSA goes beyond the one-year time horizon (which is a limitation of Pillar 1) and forces firms to assess solvency over their business planning horizon, which can mean many years for typical long-term business lines, such as life insurance.

1.3.3 Criticism of Regulatory Frameworks

The benefits arising from the regulation of financial services are not generally in doubt. Customer-protection acts, responsible corporate governance, fair and comparable accounting rules, transparent information on risk, capital and solvency for shareholders and clients are all viewed as positive developments.

Very few would argue the extreme position that the prudential regulatory frameworks we have discussed are not needed; in general, after a crisis, the demand (at least from the public and politicians) is for more regulation. Nevertheless, there are aspects of the regulatory frameworks that have elicited criticism, as we now discuss.

Cost and complexity. The cost factor of setting up a well-functioning risk-management system compliant with the present regulatory framework is significant, especially (in relative terms) for smaller institutions. On 27 March 2013, the *Financial Times* quoted Andrew Bailey (head of the Prudential Regulatory Authority in the UK) as saying that Solvency II compliance was set to cost UK companies at least £3 billion, a "frankly indefensible" amount. Related to the issue of cost is the belief that regulation, in its attempt to become more risk sensitive, is becoming too complex; this theme is taken up by the Basel Committee in their 2013 discussion paper entitled "The regulatory framework: balancing risk sensitivity, simplicity and comparability" (Basel Committee on Banking Supervision 2013b).

Endogenous risk. In general terms, this refers to the risk that is generated within a system and amplified by the system due to feedback effects. Regulation, a feature of the system, may be one of the channels by which shocks are amplified.

Regulation can lead to *risk-management herding*, whereby institutions following similar (perhaps VaR-based) rules may all be "running for the same exit" in times of crisis, consequently destabilizing an already precarious situation even further. This herding phenomenon has been suggested in connection with the 1987 stock market crash and the events surrounding the 1998 LTCM crisis (Daníelsson et al. 2001b).

An even more compelling example was observed during the 2007–9 crisis; to comply with regulatory capital ratios in a market where asset values were falling and risks increasing, firms adjusted their balance sheets by selling assets, causing further asset value falls and vanishing market liquidity. This led to criticism of the inherently *procyclical* nature of the Basel II regulation, whereby capital requirements may rise in times of stress and fall in times of expansion; the Basel III proposals attempt to address this issue with a countercyclical capital buffer.

Consequences of fair-value accounting and market-consistent valuation. The issue of procyclicality is also related to the widespread use of fair-value accounting and market-consistent valuation, which are at the heart of both the Basel rules for the trading book and the Solvency II framework. The fact that capital requirements are so closely coupled to volatile financial markets has been another focus of criticism.

An example of this is the debate around the valuation of insurance liabilities in periods of market stress. A credit crisis, of the kind experienced in 2007–9, can impact the high-quality corporate bonds that insurance companies hold on the asset side of their balance sheets. The relative value of corporate bonds compared with safe government bonds can fall sharply as investors demand more compensation for taking on both the credit risk and, in particular, the liquidity risk of corporate bonds.

The effect for insurers is that the value of their assets falls relative to the value of their liabilities, since the latter are valued by comparing cash flows with safe government bonds. At a particular point in time, an insurer may appear to have insufficient capital to meet solvency capital requirements. However, if an insurer has matched its asset and liability cash flows and can continue to meet its contractual obligations to policyholders, the apparent depletion of capital may not be a problem; insurance is a long-term business and the insurer has no short-term need to sell assets or offload liabilities, so a loss of capital need not be realized unless some of the bonds actually default.

Regulation that paints an unflattering picture of an insurer's solvency position is not popular with regulated firms. Firms have argued that they should be able to value liabilities at a lower level, by comparing the cash flows not with expensive government bonds but instead with the corporate bonds that are actually used as matching assets, making allowance only for the credit risk in corporate bonds. This has given rise to the idea of discounting with an extra *illiquidity premium*, or *matching premium*, above a risk-free rate. There has been much debate about this issue between those who feel that such proposals undermine market-consistent valuation and those who believe that strict adherence to market-consistent valuation overstates risk and has potential systemic consequences (see, for example, Wüthrich 2011).

Limits to quantification. Further criticism has been levelled at the highly quantitative nature of regulation and the extensive use of mathematical and statistical methods. The section on "Misplaced reliance on sophisticated mathematics" in the Turner Review of the global banking crisis (Lord Turner 2009) states that:

> The very complexity of the mathematics used to measure and manage risk, moreover, made it increasingly difficult for top management and boards to assess and exercise judgement over the risk being taken. Mathematical sophistication ended up not containing risk, but providing false assurances that other prima facie indicators of increasing risk (e.g. rapid credit extension and balance sheet growth) could be safely ignored.

This idea that regulation can lead to overconfidence in the quality of statistical risk measures is related to the view that the essentially backward-looking nature of estimates derived from historical data is a weakness. The use of conventional VaR-based methods has been likened to driving a car while looking in the rear-view mirror, the idea being that this is of limited use in preparing for the shocks that lie ahead.

The extension of the quantitative approach to operational risk has been controversial. Whereas everyone agrees that risks such as people risk (e.g. incompetence, fraud), process risk (e.g. model, transaction and operational control risk), technology risk (e.g. system failure, programming error) and legal risk are important, there is much disagreement on the extent to which these risks can be measured.

Limits to the efficacy of regulation. Finally, there is some debate about whether or not tighter regulation can ever prevent the occurrence of crises like that of 2007–9. The sceptical views of central bankers and regulatory figures were reported in the *Economist* in an article entitled "The inevitability of instability" (25 January 2014) (see also Prates 2013). The article suggests that "rules are constantly overtaken by financial innovation" and refers to the economist J. K. Galbraith (1993), who wrote:

> All financial innovation involves, in one form or another, the creation of debt secured in greater or lesser adequacy by real assets.... All crises have involved debt that, in one fashion or another, has become dangerously out of scale in relation to the underlying means of payment.

Tightening up the capital treatment of securitizations may prevent a recurrence of the events surrounding the 2007–9 crisis, but, according to the sceptical view, it will not prevent different forms of debt-fuelled crisis in the future.

1.4 Why Manage Financial Risk?

An important issue that we have barely touched upon is the reason for investing in risk management in the first place. This question can be addressed from various perspectives, including those of the customer of a financial institution, its shareholders, its management, its board of directors, regulators, politicians, or the general public; each of these stakeholders may have a different view. In the selective account we give here, we focus on two viewpoints: that of society as a whole, and that of the shareholders (owners) of a firm.

1.4.1 A Societal View

Modern society relies on the smooth functioning of banking and insurance systems, and it has a collective interest in the stability of such systems. The regulatory process that has given us the Basel and Solvency II frameworks was initially motivated by the desire to prevent the insolvency of individual institutions, thus protecting customers and policyholders; this is sometimes referred to as a *microprudential* approach. However, the reduction of systemic risk—the danger that problems in a single financial institution may spill over and, in extreme situations, disrupt the normal functioning of the entire financial system—has become an important secondary focus, particularly since the 2007–9 crisis. Regulation therefore now also takes a *macroprudential* perspective.

Most members of society would probably agree that protection of customers against the failure of an individual firm is an important aim, and there would be widespread agreement that the promotion of financial stability is vital. However, it is not always clear that the two aims are well aligned. While there are clearly situations where the failure of one company may lead to spillover effects that result in a systemic crisis, there may also be situations where the long-term interests of financial stability are better served by allowing a company to fail: it may provide a lesson in the importance of better risk management for other companies. This issue is clearly related to the *systemic importance* of the company in question: in other words, to its size and the extent of its connectivity to other firms. But the recognition that there may be firms that are too important or are *too big to fail* creates a *moral hazard*, since the management of such a firm may take more risk in the knowledge that the company would be bailed out in a crisis. Of course, it may be the case that in some countries some institutions are also *too big to save*.

The 2007–9 crisis provided a case study that brought many of these issues to the fore. As we noted in our account of the crisis in Section 1.2, it was initially believed that the growth in securitization was dispersing credit risk throughout the system and was beneficial to financial stability. But the warehousing of vast amounts of inadequately capitalized credit risk (in the form of CDOs) in trading books, combined with the interconnectedness of banks through derivatives and interbank lending activities, meant that quite the opposite was true. The extent of the systemic risk that had been accumulating became apparent when Lehman Brothers filed for bankruptcy on 15 September 2008 and governments intervened to bail out the banks.

It was the following phase of the crisis during which society suffered. The world economy went into recession, households defaulted on their debts, and savings and pensions were hit hard. The crisis moved “from Wall Street to Main Street”. Naturally, this led to resentment as banking remained a highly rewarded profession and it seemed that the government-sponsored bailouts had allowed banks “to privatize their gains and socialize their losses”.

There has been much debate since the crisis on whether the US government could have intervened to save Lehman, as it did for other firms such as AIG. In the *Financial Times* on 14 September 2009, the historian Niall Ferguson wrote:

> Like the executed British admiral in Voltaire’s famous phrase, Lehman had to die *pour encourager les autres*—to convince the other banks that they needed injections of public capital, and to convince the legislature to approve them. Not everything in history is inevitable; contingencies abound. Sometimes it is therefore right to say “if only”. But an imagined rescue of Lehman Brothers is the wrong counterfactual. The right one goes like this. If only Lehman’s failure and the passage of TARP had been followed—not immediately, but after six months—by a clear statement to the surviving banks that none of them was henceforth too big to fail, then we might actually have learnt something from this crisis.

While it is difficult to speak with authority for “society”, the following conclusions do not seem unreasonable. The interests of society are served by enforcing the discipline of risk management in financial firms, through the use of regulation. Better risk management can reduce the risk of company failure and protect customers and policyholders who stand in a very unequal financial relationship with large firms. However, the regulation employed must be designed with care and should not promote herding, procyclical behaviour or other forms of endogenous risk that could result in a systemic crisis with far worse implications for society than the failure of a single firm. Individual firms need to be allowed to fail on occasion, provided customers can be shielded from the worst consequences through appropriate compensation schemes. A system that allows firms to become too big to fail creates moral hazard and should be avoided.

1.4.2 The Shareholder's View

It is widely believed that proper financial risk management can increase the value of a corporation and hence shareholder value. In fact, this is the main reason why corporations that are not subject to regulation by financial supervisory authorities engage in risk-management activities. Understanding the relationship between shareholder value and financial risk management also has important implications for the design of risk-management systems. Questions to be answered include the following.

- When does risk management increase the value of a firm, and which risks should be managed?
- How should risk-management concerns factor into investment policy and capital budgeting?

There is a rather extensive corporate-finance literature on the issue of "corporate risk management and shareholder value". We briefly discuss some of the main arguments. In this way we hope to alert the reader to the fact that there is more to risk management than the mainly technical questions related to the implementation of risk-management strategies dealt with in the core of this book.

The first thing to note is that from a corporate-finance perspective it is by no means obvious that in a world with perfect capital markets risk management enhances shareholder value: while *individual* investors are typically risk averse and should therefore manage the risk in their portfolios, it is not clear that risk management or risk reduction at the *corporate level*, such as hedging a foreign-currency exposure or holding a certain amount of risk capital, increases the value of a corporation. The rationale for this (at first surprising) observation is simple: if investors have access to perfect capital markets, they can do the risk-management transactions they deem necessary via their own trading and diversification. The following statement from the chief investment officer of an insurance company exemplifies this line of reasoning: "If our shareholders believe that our investment portfolio is too risky, they should short futures on major stock market indices."

The potential irrelevance of corporate risk management for the value of a corporation is an immediate consequence of the famous *Modigliani–Miller Theorem* (Modigliani and Miller 1958). This result, which marks the beginning of modern corporate-finance theory, states that, in an ideal world without taxes, bankruptcy costs and informational asymmetries, and with frictionless and arbitrage-free capital markets, the financial structure of a firm, and hence also its risk-management decisions, are irrelevant when assessing the firm's value. Hence, in order to find reasons for corporate risk management, one has to "turn the Modigliani–Miller Theorem upside down" and identify situations where risk management enhances the value of a firm by deviating from the unrealistically strong assumptions of the theorem. This leads to the following rationales for risk management.

- Risk management can reduce *tax costs*. Under a typical tax regime the amount of tax to be paid by a corporation is a *convex* function of its profits; by reducing the variability in a firm's cash flow, risk management can therefore lead to a higher expected after-tax profit.
- Risk management can be beneficial, since a company may (and usually will) have better access to capital markets than individual investors.
- Risk management can increase firm value in the presence of *bankruptcy costs*, as it makes bankruptcy less likely.
- Risk management can reduce the impact of *costly external financing* on the firm value, as it facilitates the achievement of optimal investment.

The last two points merit a more detailed discussion. Bankruptcy costs consist of direct bankruptcy costs, such as the cost of lawsuits, and the more important indirect bankruptcy costs. The latter may include liquidation costs, which can be substantial in the case of intangibles like research and development and knowhow. This is why high research and development spending appears to be positively correlated with the use of risk-management techniques. Moreover, increased likelihood of bankruptcy often has a negative effect on key employees, management and customer relations, in particular in areas where a client wants a long-term business relationship. For instance, few customers would want to enter into a life insurance contract with an insurance company that is known to be close to bankruptcy. On a related note, banks that are close to bankruptcy might be faced with the unpalatable prospect of a bank run, where depositors try to withdraw their money simultaneously. A further discussion of these issues is given in Altman (1993).

It is a "stylized fact" of corporate finance that for a corporation, external funds are more costly to obtain than internal funds, an observation which is usually attributed to problems of asymmetric information between the management of a corporation and bond and equity investors. For instance, raising external capital from outsiders by issuing new shares might be costly if the new investors, who have incomplete information about the economic prospects of a firm, interpret the share issue as a sign that the firm is overvalued. This can generate a rationale for risk management for the following reason: without risk management the increased variability of a company's cash flow will be translated either into an increased variability of the funds that need to be raised externally or to an increased variability in the amount of investment. With increasing marginal costs of raising external capital and decreasing marginal profits from new investment, we are left with a decrease in (expected) profits. Proper risk management, which amounts to a smoothing of the cash flow generated by a corporation, can therefore be beneficial. For references to the literature see Notes and Comments below.

1.5 Quantitative Risk Management

The aim of this chapter has been to place QRM in a larger historical, regulatory and even societal framework, since a study of QRM without a discussion of its proper setting and motivation makes little sense. In the remainder of the book we adopt a somewhat narrower view and treat QRM as a quantitative science that uses the language of mathematics in general, and of probability and statistics in particular.

In this section we discuss the relevance of the Q in QRM, describe the quantitative modelling challenge that we have attempted to meet in this book, and end with thoughts on where QRM may lead in the future.

1.5.1 The Q in QRM

In Section 1.2.1 we discussed the view that the use of advanced mathematical modelling and valuation techniques has been a contributory factor in financial crises, particularly those attributed to derivative products, such as CDOs in the 2007–9 crisis. We have also referred to criticism of the quantitative, statistical emphasis of the modern regulatory framework in Section 1.3.3. These arguments must be taken seriously, but we believe that it is neither possible nor desirable to remove the quantitative element from risk management.

Mathematics and statistics provide us with a suitable language and appropriate concepts for describing financial risk. This is clear for complex financial products such as derivatives, which cannot be valued and handled without mathematical models. But the need for quantitative modelling also arises for simpler products, such as a book of mortgages for retail clients. The main risk in managing such a book is the occurrence of disproportionately many defaults: a risk that is directly related to the dependence between defaults (see Chapter 11 for details). In order to describe this dependence, we need mathematical concepts from multivariate statistics, such as correlations or copulas; if we want to carry out a simulation study of the behaviour of the portfolio under different economic scenarios, we need a mathematical model that describes the joint distribution of default events; if the portfolio is large, we will also need advanced simulation techniques to generate the relevant scenarios efficiently.

Moreover, mathematical and statistical methods can do better than they did in the 2007–9 crisis. In fact, providing concepts, techniques and tools that address some of the weaker points of current methodology is a main theme of our text and we come back to this point in the next section.

There is a view that, instead of using mathematical models, there is more to be learned about risk management through a *qualitative* analysis of historical case studies and the formulation of narratives. What is often overlooked by the non-specialist is that mathematical models are themselves nothing more than narratives, albeit narratives couched in a precise symbolic language. Addressing the question "What is mathematics?", Gale and Shapley (1962) wrote: "Any argument which is carried out with sufficient precision is mathematical." Lloyd Shapley went on to win the 2012 Nobel Memorial Prize in Economic Science.

It is certainly true that mathematical methods can be misused. Mathematicians are very well aware that a mathematical result has not only a conclusion but, equally importantly, certain conditions under which it holds. Statisticians are well aware that inductive reasoning on the basis of models relies on the assumption that these conditions hold in the real world. This is especially true in economics, which as a social science is concerned with phenomena that are not easily described by clear mathematical or physical laws. By starting with questionable assumptions, models can be used (or manipulated) to deliver bad answers. In a talk on 20 March 2009, the economist Roger Guesnerie said, "For this crisis, mathematicians are innocent . . . and this in both meanings of the word." The implication is that quantitative risk managers must become more worldly about the ways in which models are used. But equally, the regulatory system needs to be more vigilant about the ways in which models can be gamed and the institutional pressures that can circumvent the best intentions of prudent quantitative risk managers.

We are firmly of the opinion—an opinion that has only been reinforced by our study of financial crises—that the Q in QRM is an essential part of the process. We reject the idea that the Q is part of the problem, and we believe that it remains (if applied correctly and honestly) a part of the solution to managing risk. In summary, we strongly agree with Shreve (2008), who said:

> Don't blame the quants. Hire good ones instead and listen to them.

1.5.2 The Nature of the Challenge

When we began this book project we set ourselves the task of defining a new discipline of QRM. Our approach to this task has had two main strands. On the one hand, we have attempted to put current practice onto a firmer mathematical footing, where, for example, concepts like P&L distributions, risk factors, risk measures, capital allocation and risk aggregation are given formal definitions and a consistent notation. In doing this we have been guided by the consideration of what topics should form the core of a course on QRM for a wide audience of students interested in risk-management issues; nonetheless, the list is far from complete and will continue to evolve as the discipline matures. On the other hand, the second strand of our endeavour has been to put together material on techniques and tools that go beyond current practice and address some of the deficiencies that have been repeatedly raised by critics. In the following paragraphs we elaborate on some of these issues.

Extremes matter. A very important challenge in QRM, and one that makes it particularly interesting as a field for probability and statistics, is the need to address unexpected, abnormal or extreme outcomes, rather than the expected, normal or average outcomes that are the focus of many classical applications. This is in tune with the regulatory view expressed by Alan Greenspan in 1995 at the Joint Central Bank Research Conference:

> From the point of view of the risk manager, inappropriate use of the normal distribution can lead to an understatement of risk, which must be

> balanced against the significant advantage of simplification. From the central bank's corner, the consequences are even more serious because we often need to concentrate on the left tail of the distribution in formulating lender-of-last-resort policies. Improving the characterization of the distribution of extreme values is of paramount importance.

While the quote is older, the same concern about underestimation of extremes is raised in a passage in the Turner Review (Lord Turner 2009):

> Price movements during the crisis have often been of a size whose probability was calculated by models (even using longer-term inputs) to be almost infinitesimally small. This suggests that the models systematically underestimated the chances of small probability high impact events.... It is possible that financial market movements are inherently characterized by fat-tail distributions. VaR models need to be buttressed by the application of stress test techniques which consider the impact of extreme movements beyond those which the model suggests are at all probable.

Much space in our book is devoted to models for financial risk factors that go beyond the normal (or Gaussian) model and attempt to capture the related phenomena of heavy or fat tails, excess volatility and extreme values.

The interdependence and concentration of risks. A further important challenge is presented by the multivariate nature of risk. Whether we look at market risk or credit risk, or overall enterprise-wide risk, we are generally interested in some form of aggregate risk that depends on high-dimensional vectors of underlying risk factors, such as individual asset values in market risk or credit spreads and counterparty default indicators in credit risk.

A particular concern in our multivariate modelling is the phenomenon of dependence between extreme outcomes, when many risk factors move against us simultaneously. In connection with the LTCM case (see Section 1.2.1) we find the following quote in *Business Week* (September 1998):

> Extreme, synchronized rises and falls in financial markets occur infrequently but they do occur. The problem with the models is that they did not assign a high enough chance of occurrence to the scenario in which many things go wrong at the same time—the "perfect storm" scenario.

In a perfect storm scenario the risk manager discovers that portfolio diversification arguments break down and there is much more of a concentration of risk than had been imagined. This was very much the case with the 2007–9 crisis: when borrowing rates rose, bond markets fell sharply, liquidity disappeared and many other asset classes declined in value, with only a few exceptions (such as precious metals and agricultural land), a perfect storm was created.

We have mentioned (see Section 1.2.1) the notorious role of the Gauss copula in the 2007–9 financial crisis. An April 2009 article in the *Economist*, with the title

"In defence of the Gaussian copula", evokes the environment at the time of the securitization boom:

> By 2001, correlation was a big deal. A new fervour was gripping Wall Street—one almost as revolutionary as that which had struck when the Black–Scholes model brought about the explosion in stock options and derivatives in the early 1980s. This was structured finance, the culmination of two decades of quants on Wall Street.... The problem, however, was correlation. The one thing any off-balance-sheet securitisation could not properly capture was the interrelatedness of all the hundreds of thousands of different mortgage loans they owned.

The Gauss copula appeared to solve this problem by offering a model for the correlated times of default of the loans or other credit-risky assets; the perils of this approach later became clear. In fact, the Gauss copula is not an example of the use of oversophisticated mathematics; it is a relatively simple model that is difficult to calibrate reliably to available market information. The modelling of dependent credit risks, and the issue of model risk in that context, is a subject we look at in some detail in our treatment of credit risk.

The problem of scale. A further challenge in QRM is the typical scale of the portfolios under consideration; in the most general case, a portfolio may represent the entire position in risky assets of a financial institution. Calibration of detailed multivariate models for all risk factors is an almost impossible task, and any sensible strategy must involve dimension reduction; that is to say, the identification of key risk drivers and a concentration on modelling the main features of the overall risk landscape.

In short, we are forced to adopt a fairly broad-brush approach. Where we use econometric tools, such as models for financial return series, we are content with relatively simple descriptions of individual series that capture the main phenomenon of volatility, and which can be used in a parsimonious multivariate factor model. Similarly, in the context of portfolio credit risk, we are more concerned with finding suitable models for the default dependence of counterparties than with accurately describing the mechanism for the default of an individual, since it is our belief that the former is at least as important as the latter in determining the risk of a large diversified portfolio.

Interdisciplinarity. Another aspect of the challenge of QRM is the fact that ideas and techniques from several existing quantitative disciplines are drawn together. When one considers the ideal education for a quantitative risk manager of the future, then a combined quantitative skill set should undoubtedly include concepts, techniques and tools from such fields as mathematical finance, statistics, financial econometrics, financial economics and actuarial mathematics. Our choice of topics is strongly guided by a firm belief that the inclusion of modern statistical and econometric techniques and a well-chosen subset of actuarial methodology are essential for the establishment of best-practice QRM. QRM is certainly not just about financial mathematics and derivative pricing, important though these may be.

Communication and education. Of course, the quantitative risk manager operates in an environment where additional non-quantitative skills are equally important. Communication is certainly an important skill: risk professionals, by the definition of their duties, will have to interact with colleagues with diverse training and backgrounds, at all levels of their organization. Moreover, a quantitative risk manager has to familiarize him or herself quickly with all-important market practice and institutional details. A certain degree of humility will also be required to recognize the role of QRM in a much larger picture.

A lesson from the 2007–9 crisis is that improved education in QRM is essential; from the front office to the back office to the boardroom, the users of models and their outputs need to be better trained to understand model assumptions and limitations. This task of educating users is part of the role of a quantitative risk manager, who should ideally have (or develop) the pedagogical skills to explain methods and conclusions to audiences at different levels of mathematical sophistication.

1.5.3 QRM Beyond Finance

The use of QRM technology is not restricted to the financial services industry, and similar developments have taken place, or are taking place, in other sectors of industry. Some of the earliest applications of QRM are to be found in the manufacturing industry, where similar concepts and tools exist under names like reliability or total quality control. Industrial companies have long recognized the risks associated with bringing faulty products to the market. The car manufacturing industry in Japan, in particular, was an early driving force in this respect.

More recently, QRM techniques have been adopted in the transport and energy industries, to name but two. In the case of energy, there are obvious similarities with financial markets: electrical power is traded on energy exchanges; derivatives contracts are used to hedge future price uncertainty; companies optimize investment portfolios combining energy products with financial products; some Basel methodology can be applied to modelling risk in the energy sector. However, there are also important dissimilarities due to the specific nature of the industry; most importantly, there are the issues of the cost of storage and transport of electricity as an underlying commodity, and the necessity of modelling physical networks including the constraints imposed by the existence of national boundaries and quasi-monopolies.

There are also markets for environmental emission allowances. For example, the Chicago Climate Futures Exchange offers futures contracts on sulphur dioxide emissions. These are traded by industrial companies producing the pollutant in their manufacturing process, and they force such companies to consider the cost of pollution as a further risk in their risk landscape.

A natural consequence of the evolution of QRM thinking in different industries is an interest in the transfer of risks between industries; this process is known as alternative risk transfer. To date the best examples of risk transfer are between the insurance and banking industries, as illustrated by the establishment of catastrophe futures by the Chicago Board of Trade in 1992. These came about in the wake of Hurricane Andrew, which caused $20 billion of insured losses on the East Coast of the US. While this was a considerable event for the insurance industry in relation to overall reinsurance capacity, it represented only a drop in the ocean compared with the daily volumes traded worldwide on financial exchanges. This led to the recognition that losses could be covered in future by the issuance of appropriately structured bonds with coupon streams and principal repayments dependent on the occurrence or non-occurrence of well-defined natural catastrophe events, such as storms and earthquakes.

A speculative view of where these developments may lead is given by Shiller (2003), who argues that the proliferation of risk-management thinking coupled with the technological sophistication of the twenty-first century will allow any agent in society, from a company to a country to an individual, to apply QRM methodology to the risks they face. In the case of an individual this may be the risk of unemployment, depreciation in the housing market or investment in the education of children.

Notes and Comments

The language of probability and statistics plays a fundamental role throughout this book, and readers are expected to have a good knowledge of these subjects. At the elementary level, Rice (1995) gives a good first introduction to both. More advanced texts in probability and stochastic processes are Williams (1991), Resnick (1992) and Rogers and Williams (1994); the full depth of these texts is certainly not required for the understanding of this book, though they provide excellent reading material for more mathematically sophisticated readers who also have an interest in mathematical finance. Further recommended texts on statistical inference include Casella and Berger (2002), Bickel and Doksum (2001), Davison (2003) and Lindsey (1996).

In our discussion of risk and randomness in Section 1.1.1 we mentioned Knight (1921) and Keynes (1920), whose classic texts are very much worth revisiting. Knightian uncertainty refers to uncertainty that cannot be measured and is sometimes contrasted with risks that can be measured using probability. This relates to the more recent idea of a Black Swan event, a term popularized in Taleb (2007) but introduced in Taleb (2001). Black swans were believed to be imaginary creatures until the European exploration of Australia and the name is applied to unprecedented and unpredictable events that challenge conventional beliefs and models. Donald Rumsfeld, a former US Secretary of Defense, referred to “unknown unknowns” in a 2002 news briefing on the evidence for the presence of weapons of mass destruction in Iraq.

An excellent text on the history of risk and probability with financial applications in mind is Bernstein (1998). We also recommend Shiller (2012) for more on the societal context of financial risk management. A thought-provoking text addressing risk on Wall Street from a historical perspective is Brown (2012).

For the mathematical reader looking to acquire more knowledge about the relevant economics we recommend Mas-Colell, Whinston and Green (1995) for microeconomics, Campbell, Lo and MacKinlay (1997) or Gouriéroux and Jasiak (2001) for econometrics, and Brealey and Myers (2000) for corporate finance. From the vast literature on options, an entry-level text for the general reader is Hull (2014). At a more mathematical level we like Bingham and Kiesel (2004), Musiela and Rutkowski (1997), Shreve (2004a) and Shreve (2004b). One of the most readable texts on the basic notion of options is Cox and Rubinstein (1985). For a rather extensive list of the kind of animals to be found in the zoological garden of derivatives, see, for example, Haug (1998).

There are several texts on the spectacular losses that occurred as the result of speculative trading and the careless use of derivatives. For a historical overview of financial crises, see Reinhart and Rogoff (2009), as well as the much earlier Galbraith (1993) and Kindleberger (2000). Several texts exist on more recent crises; we list only a few. The LTCM case is well documented in Dunbar (2000), Lowenstein (2000) and Jorion (2000), the latter particularly focusing on the technical risk-measurement issues involved. Boyle and Boyle (2001) give a very readable account of the Orange County, Barings and LTCM stories (see also Jacque 2010). For the Equitable Life case see the original Penrose Report, published by the UK government (Lord Penrose 2004), or an interesting paper by Roberts (2012). Many books have emerged on the 2007–9 crisis; early warnings are well summarized, under Greenspan's memorable "irrational exuberance" phrase, in a pre-crisis book by Shiller (2000), and the post-mortem by the same author is also recommended (Shiller 2008).

An overview of options embedded in life insurance products is given in Dillmann (2002), guarantees are discussed in detail in Hardy (2003), and Briys and de Varenne (2001) contains an excellent account of risk-management issues facing the (life) insurance industry. For risk-management and valuation issues underlying life insurance, see Koller (2011) and Møller and Steffensen (2007). Market-consistent actuarial valuation is discussed in Wüthrich, Bühlmann and Furrer (2010).

The historical development of banking regulation is well described in Crouhy, Galai and Mark (2001) and Steinherr (1998). For details of the current rules and regulations coming from the Basel Committee, see its website at www.bis.org/bcbs. Besides copies of the various accords, one can also find useful working papers, publications and comments written by stakeholders on the various consultative packages. For Solvency II and the Swiss Solvency Test, many documents are to be found on the web. Comprehensive textbook accounts are Sandström (2006) and Sandström (2011), and a more technical treatment is found in Wüthrich and Merz (2013). The complexity of risk-management methodology in the wake of Basel II is critically addressed by Hawke (2003), from his perspective as US Comptroller of the Currency. Among the numerous texts written after the 2007–9 crisis, we found all of Rochet (2008), Shin (2010), Dewatripont, Rochet and Tirole (2010) and Bénéplanc and Rochet (2011) useful. For a discussion of issues related to the use of fair-value accounting during the financial crisis, see Ryan (2008).

For a very detailed overview of relevant practical issues underlying risk management, we again strongly recommend Crouhy, Galai and Mark (2001). A text stressing the use of VaR as a risk measure and containing several worked examples is Jorion (2007), whose author also has a useful teaching manual on the same subject

(Jorion 2002b). Insurance-related issues in risk management are nicely presented in Doherty (2000).

For a comprehensive discussion of the management of bank capital given regulatory constraints, see Matten (2000), Klaassen and van Eeghen (2009) and Admati and Hellwig (2013). Graham and Rogers (2002) contains a discussion of risk management and tax incentives. A formal account of the Modigliani–Miller Theorem and its implications can be found in many textbooks on corporate finance: a standard reference is Brealey and Myers (2000), and de Matos (2001) gives a more theoretical account from the perspective of modern financial economics. Both texts also discuss the implications of informational asymmetries between the various stakeholders in a corporation. Formal models looking at risk management from a corporate-finance angle are to be found in Froot and Stein (1998), Froot, Scharfstein and Stein (1993) and Stulz (1996, 2002). For a specific discussion on corporate-finance issues in insurance, see Froot (2007) and Hancock, Huber and Koch (2001).

There are several studies on the use of risk-management techniques for non-financial firms (see, for example, Bodnar, Hayt and Marston 1998; Geman 2005, 2009). Two references in the area of the reliability of industrial processes are Bedford and Cooke (2001) and Does, Roes and Trip (1999). Interesting edited volumes on alternative risk transfer are Shimpi (2001), Barrieu and Albertini (2009) and Kiesel, Scherer and Zagst (2010); a detailed study of model risk in the alternative risk transfer context is Schmock (1999). An area we have not mentioned so far in our discussion of QRM in the future is that of real options. A real option is the right, but not the obligation, to take an action (e.g. deferring, expanding, contracting or abandoning) at a predetermined cost called the exercise price. The right holds for a predetermined period of time—the life of the option. This definition is taken from Copeland and Antikarov (2001). Examples of real options discussed in the latter are the valuation of an internet project and of a pharmaceutical research and development project. A further useful reference is Brennan and Trigeorgis (1999).

A well-written critical view of the failings of the standard approach to risk management is given in Rebonato (2007). And finally, for an entertaining text on the biology of the much criticized "homo economicus", we like Coates (2012).

2

Basic Concepts in Risk Management

In this chapter we define or explain a number of fundamental concepts used in the measurement and management of financial risk. Beginning in Section 2.1 with the simplified balance sheet of a bank and an insurer, we discuss the risks faced by such firms, the nature of capital, and the need for a firm to have sufficient capital to withstand financial shocks and remain solvent.

In Section 2.2 we establish a mathematical framework for describing changes in the value of portfolios and deriving loss distributions. We provide a number of examples to show how this framework applies to different kinds of asset and liability portfolios. The examples are also used to discuss the meaning of value in more detail with reference to fair-value accounting and risk-neutral valuation.

Section 2.3 is devoted to the subject of using risk measures to determine risk or solvency capital. We present different quantitative approaches to measuring risk, with a particular focus on risk measures that are calculated from loss distributions, like value-at-risk and expected shortfall.

2.1 Risk Management for a Financial Firm

2.1.1 Assets, Liabilities and the Balance Sheet

A good way to understand the risks faced by a modern financial institution is to look at the stylized balance sheet of a typical bank or insurance company. A balance sheet is a financial statement showing *assets and liabilities*; roughly speaking, the assets describe the financial institution's investments, whereas liabilities refer to the way in which funds have been raised and the obligations that ensue from that fundraising.

A typical bank raises funds by taking in customer deposits, by issuing bonds and by borrowing from other banks or from central banks. Collectively these form the *debt capital* of the bank, which is invested in a number of ways. Most importantly, it is used for loans to retail, corporate and sovereign customers, invested in traded securities, lent out to other banks or invested in property or in other companies. A small fraction is also held as cash.

A typical insurance company sells insurance contracts, collecting premiums in return, and raises additional funds by issuing bonds. The liabilities of an insurance company thus consist of its obligations to policyholders, which take the form of a *technical reserve against future claims*, and its obligations to bondholders. The funds raised are then invested in traded securities, particularly bonds, as well as other assets such as real estate.

Table 2.1. The stylized balance sheet of a typical bank.

Bank ABC (31 December 2015)			
Assets		Liabilities	
Cash	£10M	Customer deposits	£80M
(and central bank balance)			
Securities	£50M	Bonds issued	
– bonds		– senior bond issues	£25M
– stocks		– subordinated bond issues	£15M
– derivatives		Short-term borrowing	£30M
Loans and mortgages	£100M	Reserves (for losses on loans)	£20M
– corporates			
– retail and smaller clients		*Debt (sum of above)*	£170M
– government			
Other assets	£20M		
– property			
– investments in companies		*Equity*	£30M
Short-term lending	£20M		
Total	£200M	Total	£200M

In both cases a small amount of extra funding stems from occasional *share issues*, which form the share capital of the bank or insurer. This form of funding is crucial as it entails no obligation towards outside parties.

These simplified banking and insurance *business models* are reflected in the stylized balance sheets shown in Tables 2.1 and 2.2. In these financial statements, assets and liabilities are valued on a given date. The position marked *equity* on the liability side of the balance sheet is the residual value defined in the balance sheet equation

$$\text{value of assets} = \text{value of liabilities} = \text{debt} + \text{equity}. \tag{2.1}$$

A company is *solvent* at a given point in time if the equity is nonnegative; otherwise it is insolvent. Insolvency should be distinguished from the notion of default, which occurs if a firm misses a payment to its debtholders or other creditors. In particular, an otherwise-solvent company can default because of *liquidity* problems, as discussed in more detail in the next section.

It should be noted that assigning values to the items on the balance sheet of a bank or insurance company is a non-trivial task. Broadly speaking, two different approaches can be distinguished. The practice of *fair-value accounting* attempts to value assets at the prices that would be received if they were sold and to value liabilities at the prices that would have to be paid if they were transferred to another party. Fair-value accounting is relatively straightforward for positions that are close to securities traded on liquid markets, since these are simply valued by (an estimate of) their market price. It is more challenging to apply fair-value principles to non-traded or illiquid assets and liabilities.

The more traditional practice of *amortized cost accounting* is still applied to many kinds of financial asset and liability. Under this practice the position is assigned a

Table 2.2. The stylized balance sheet of a typical insurer.

Insurer XYZ (31 December 2015)			
Assets		Liabilities	
Investments		Reserves for policies written	£80M
– bonds	£50M	(technical provisions)	
– stocks	£5M	Bonds issued	£10M
– real estate	£5M		
Investments for unit-linked contracts	£30M	*Debt (sum of above)*	£90M
Other assets	£10M		
– property			
		Equity	£10M
Total	£100M	Total	£100M

book value at its inception and this is carried forward over time. In some cases the value is progressively reduced or impaired to account for the aging of the position or the effect of adverse events. An example of assets valued at book value are the loans on the balance sheet of the bank. The book value would typically be an estimate of the present value (at the time the loans were made) of promised future interest and principal payments minus a provision for losses due to default.

In the European insurance industry the practice of *market-consistent valuation* has been promoted under the Solvency II framework. As described in Section 1.3.2, the rationale is very similar to that of fair-value accounting: namely, to value positions by "the amount for which an asset could be exchanged or a liability settled, between knowledgeable, willing parties in an arm's length transaction, based on observable prices within an active, deep and liquid market". However, there are some differences between market-consistent valuation and fair-value accounting for specific kinds of position. A European insurer will typically have two versions of the balance sheet in order to comply with accounting rules, on the one hand, and Solvency II rules for capital adequacy, on the other. The accounting balance sheet may mix fair-value and book-value approaches, but the Solvency II balance sheet will apply market-consistent principles throughout.

Overall, across the financial industry, there is a tendency for the accounting standard to move towards fair-value accounting, even if the financial crisis of 2007–9 demonstrated that this approach is not without problems during periods when trading activity and market liquidity suddenly vanish (see Section 1.3.3 for more discussion of this issue). Fair-value accounting for financial products will be discussed in more detail in Section 2.2.2.

2.1.2 Risks Faced by a Financial Firm

An obvious source of risk for a bank is a decrease in the value of its investments on the asset side of the balance sheet. This includes market risk, such as losses from securities trading, and credit risk. Another important risk is related to funding and so-called *maturity mismatch*: for a typical bank, large parts of the asset side consist of relatively illiquid, long-term investments such as loans or property, whereas a large part of the liabilities side consists of short-term obligations such as funds borrowed from money markets and most customer deposits. This may lead to problems when the cost of short-term refinancing increases due to rising short-term interest rates, because the banks may have difficulties selling long-term assets to raise funds. This can lead to the default of a bank that is technically solvent; in extreme cases there might even be a *bank run*, as was witnessed during the 2007–9 financial crisis. This clearly shows that risk is found on both sides of the balance sheet and that risk managers should not focus exclusively on the asset side.

The primary risk for an insurance company is clearly insolvency, i.e. the risk that the claims of policyholders cannot be met. This can happen due to adverse events affecting the asset side or the liability side of the balance sheet. On the asset side, the risks are similar to those for a bank. On the liability side, the main risk is that reserves are insufficient to cover future claim payments. It is important to bear in mind that the liabilities of a life insurer are of a long-term nature (due to the sale of products such as annuities) and are subject to many categories of risk including interest-rate risk, inflation risk and longevity risk, some of which also affect the asset side. An important aspect of the risk-management strategy of an insurance company is, therefore, to hedge parts of these risks by proper investment of the premium income (so-called liability-driven investment).

It should be clear from this discussion that a sound approach to risk management cannot look at one side of the balance sheet in isolation from the other.

2.1.3 Capital

There are many different notions of bank *capital*, and three broad concepts can be distinguished: *equity (or book) capital*, *regulatory capital* and *economic capital*. All of these notions of capital refer to items on the liability side of the balance sheet that entail no (or very limited) obligations to outside creditors and that can thus serve as a buffer against losses.

The equity capital can be read from the balance sheet according to the balance sheet equation in (2.1). It is therefore a measure of the value of the company to the shareholders. The balance sheet usually gives a more detailed breakdown of the equity capital by listing separate positions for *shareholder capital*, *retained earnings* and other items of lesser importance. Shareholder capital is the initial capital invested in the company by purchasers of equity. For companies financed by a single share issue, this is given by the numbers of shares issued multiplied by their price at the issuance date. Shareholder capital is therefore different from market capitalization, which is given by the number of shares issued multiplied by their current market price. Retained earnings are the accumulated earnings that have not been paid out in the form of dividends to shareholders; these can in principle be negative if the company has made losses.

Regulatory capital is the amount of capital that a company should have according to regulatory rules. For a bank, the rules are set out in the Basel framework, as described in more detail in Section 1.3.1. For European insurance companies, regulatory capital takes the form of a minimum capital requirement and a solvency capital requirement as set out in the Solvency II framework (see Section 1.3.2).

A regulatory capital framework generally specifies the amount of capital necessary for a financial institution to continue its operations, taking into account the size and the riskiness of its positions. Moreover, it specifies the quality of the capital and hence the form it should take on the balance sheet. In this context one usually distinguishes between different numbered capital *tiers*.

For example, in the Basel framework, Tier 1 capital is the sum of shareholder capital and retained earnings; in other words, the main constituents of the equity capital. This capital can act in full as a buffer against losses as there are no other claims on it. Tier 2 capital includes other positions of the balance sheet, in particular subordinated debt. Holders of this debt would effectively be the last to be paid before the shareholders in the event of the liquidation of the company, so subordinated debt can be viewed as an extra layer of protection for depositors and other senior debtholders. For illustration, the bank in Table 2.1 has Tier 1 capital of £30 million (assuming the equity capital consists of shareholder capital and retained earnings only) and Tier 2 capital of £45 million.

Economic capital is an estimate of the amount of capital that a financial institution needs in order to control the probability of becoming insolvent, typically over a one-year horizon. It is an internal assessment of risk capital that is guided by economic modelling principles. In particular, an economic capital framework attempts to take a holistic view that looks at assets and liabilities simultaneously, and works, where possible, with fair or market-consistent values of balance sheet items. Although, historically, regulatory capital frameworks have been based more on relatively simple rules and on book values for balance sheet items, there is increasing convergence between the economic and regulatory capital concepts, particularly in the insurance world, where Solvency II emphasizes market-consistent valuation of liabilities.

Note that the various notions of capital refer to the way in which a financial firm finances itself and not to the assets it invests in. In particular, capital requirements do not require the setting aside of funds that cannot be invested productively, e.g. by issuing new loans. There are other forms of financial regulation that refer to the asset side of the balance sheet and restrict the investment possibilities, such as obligatory cash reserves for banks and constraints on the proportion of insurance assets that may be invested in stocks.

Notes and Comments

A good introduction to the business of banking and the risks affecting banks is Choudhry (2012), while Thoyts (2010) provides a very readable overview of theory and practice in the insurance industry, with a focus on the UK. Readers wanting to go deeper into the subject of balance sheets have many financial accounting textbooks to choose from, a popular one being Elliott and Elliott (2013). A paper that gives more explanation of fair-value accounting and also discusses issues raised by the financial crisis is Ryan (2008).

Regulatory capital in the banking industry is covered in many of the documents produced by the Basel Committee, in particular the papers covering the Basel II and Basel III capital frameworks (Basel Committee on Banking Supervision 2006, 2011). For regulatory capital under Solvency II, see Sandström (2011). Textbook treatments of the management of bank capital given regulatory constraints are found in Matten (2000) and Klaassen and van Eeghen (2009), while Admati et al. (2013) provides a strong argument for capital regulation that ensures banks have a high level of equity capital. This issue is discussed at a slightly less technical level in the book by Admati and Hellwig (2013). A good explanation of the concept of economic capital may be found in the relevant entry in the *Encyclopedia of Quantitative Finance* (Rosen and Saunders 2010).

2.2 Modelling Value and Value Change

We have seen in Section 2.1.1 that an analysis of the risks faced by a financial institution requires us to consider the change in the value of its assets and liabilities. In Section 2.2.1 we set up a formal framework for modelling value and value change and illustrate this framework with stylized asset and liability portfolios. With the help of these examples we take a closer look at valuation methods in Section 2.2.2. Finally, in Section 2.2.3 we discuss the different approaches that are used to construct loss distributions for portfolios over given time horizons.

2.2.1 Mapping Risks

In our general mathematical model for describing financial risks we represent the uncertainty about future states of the world by a probability space $(\Omega, \mathcal{F}, P)$, which is the domain of all random variables (rvs) we introduce below.

We consider a given portfolio of assets and, in some cases, liabilities. At the simplest level, this could be a collection of stocks or bonds, a book of derivatives or a collection of risky loans. More generally, it could be a portfolio of life insurance contracts (liabilities) backed by investments in securities such as bonds, or even a financial institution's overall balance sheet. We denote the *value* of the portfolio at time t by V_t and assume that the rv V_t is known, or can be determined from information available, at time t. Of course, the valuation of many positions on a financial firm's balance sheet is a challenging task; we return to this issue in more detail in Section 2.2.2.

We consider a given risk-management time horizon Δt, which might be one day or ten days in market risk, or one year in credit, insurance or enterprise-wide risk management. To develop a simple formalism for talking about value, value change and the role of risk factors, we will make two simplifying assumptions:

- the portfolio composition remains fixed over the time horizon; and
- there are no intermediate payments of income during the time period.

While these assumptions may hold approximately for a one-day or ten-day horizon, they are unlikely to hold over one year, where items in the portfolio may mature and be replaced by other investments and where dividend or interest income may accumulate. In specific situations it would be possible to relax these assumptions, e.g. by specifying simple rebalancing rules for portfolios or by taking intermediate income into account.

Using a time-series notation (with time recorded in multiples of the time horizon Δt) we write the value of the portfolio at the end of the time period as V_{t+1} and the change in value of the portfolio as $\Delta V_{t+1} = V_{t+1} - V_t$. We define the *loss* to be $L_{t+1} := -\Delta V_{t+1}$, which is natural for short time intervals. For longer time intervals, on the other hand, this definition neglects the time value of money, and an alternative would be to define the loss to be $V_t - V_{t+1}/(1 + r_{t,1})$, where $r_{t,1}$ is the simple risk-free interest rate that applies between times t and $t + 1$; this measures the loss in units of money at time t. The rv L_{t+1} is typically random from the viewpoint of time t, and its distribution is termed the *loss distribution*. Practitioners in risk management are often concerned with the so-called P&L distribution. This is the distribution of the change in portfolio value ΔV_{t+1}. In this text we will often focus on L_{t+1} as this simplifies the application of many statistical methods and is in keeping with conventions in actuarial risk theory.

The value V_t is typically modelled as a function of time and a d-dimensional random vector $\boldsymbol{Z}_t = (Z_{t,1}, \dots, Z_{t,d})'$ of *risk factors*, i.e. we have the representation

$$V_t = f(t, \boldsymbol{Z}_t) \tag{2.2}$$

for some measurable function $f : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$. Risk factors are usually assumed to be observable, so the random vector $\boldsymbol{Z}_t$ takes some known realized value z_t at time t and the portfolio value V_t has realized value $f(t, z_t)$. The choice of the risk factors and of f is of course a modelling issue and depends on the portfolio at hand, on the data available and on the desired level of precision (see also Section 2.2.2). A representation of the portfolio value in the form (2.2) is termed a *mapping* of risks. Some examples of the mapping procedure are provided below.

We define the random vector of *risk-factor changes* over the time horizon to be $\boldsymbol{X}_{t+1} := \boldsymbol{Z}_{t+1} - \boldsymbol{Z}_t$. Assuming that the current time is t and using the mapping (2.2), the portfolio loss is given by

$$L_{t+1} = -(f(t+1, z_t + \boldsymbol{X}_{t+1}) - f(t, z_t)), \tag{2.3}$$

which shows that the loss distribution is determined by the distribution of the risk-factor change $\boldsymbol{X}_{t+1}$.

If f is differentiable, we may also use a first-order approximation L^{Δ}_{t+1} of the loss in (2.3) of the form

$$L^{\Delta}_{t+1} := -\biggl(f_t(t, z_t) + \sum_{i=1}^{d} f_{z_i}(t, z_t) X_{t+1,i}\biggr), \tag{2.4}$$

where the subscripts on f denote partial derivatives. The notation L^{Δ} stems from the standard *delta* terminology in the hedging of derivatives (see Example 2.2 below). The first-order approximation is convenient as it allows us to represent the loss as a *linear* function of the risk-factor changes. The quality of the approximation (2.4) is obviously best if the risk-factor changes are likely to be small (i.e. if we are measuring risk over a short horizon) and if the portfolio value is almost linear in the risk factors (i.e. if the function f has small second derivatives).

We now consider a number of examples from the areas of market, credit and insurance risk, illustrating how typical risk-management problems fit into this framework.

Example 2.1 (stock portfolio). Consider a fixed portfolio of d stocks and denote by λ_i the number of shares of stock i in the portfolio at time t. The price process of stock i is denoted by $(S_{t,i})_{t\in\mathbb{N}}$. Following standard practice in finance and risk management we use logarithmic prices as risk factors, i.e. we take $Z_{t,i} := \ln S_{t,i}$, $1 \leqslant i \leqslant d$, and we get $V_t = \sum_{i=1}^{d} \lambda_i \mathrm{e}^{Z_{t,i}}$. The risk-factor changes $X_{t+1,i} = \ln S_{t+1,i} - \ln S_{t,i}$ then correspond to the log-returns of the stocks in the portfolio. The portfolio loss from time t to $t+1$ is given by

$$L_{t+1} = -(V_{t+1} - V_t) = -\sum_{i=1}^{d} \lambda_i S_{t,i}(\mathrm{e}^{X_{t+1,i}} - 1),$$

and the linearized loss L_{t+1}^{Δ} is given by

$$L_{t+1}^{\Delta} = -\sum_{i=1}^{d} \lambda_i S_{t,i} X_{t+1,i} = -V_t \sum_{i=1}^{d} w_{t,i} X_{t+1,i}, \tag{2.5}$$

where the weight $w_{t,i} := (\lambda_i S_{t,i})/V_t$ gives the proportion of the portfolio value invested in stock i at time t. Given the mean vector and covariance matrix of the distribution of the risk-factor changes, it is very easy to compute the first two moments of the distribution of the linearized loss L^{Δ}. Suppose that the random vector $\boldsymbol{X}_{t+1}$ has a distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Using general rules for the mean and variance of linear combinations of a random vector (see also equations (6.7) and (6.8)), we immediately get

$$E(L_{t+1}^{\Delta}) = -V_t \boldsymbol{w}'\boldsymbol{\mu} \quad \text{and} \quad \operatorname{var}(L_{t+1}^{\Delta}) = V_t^2 \boldsymbol{w}'\Sigma\boldsymbol{w}. \tag{2.6}$$

Example 2.2 (European call option). We now consider a simple example of a portfolio of derivative securities: namely, a standard European call on a non-dividend-paying stock with maturity time T and exercise price K. We use the *Black–Scholes option-pricing formula* for the valuation of our portfolio. The value of a call option on a stock with price S at time t is given by

$$C^{\mathrm{BS}}(t, S; r, \sigma, K, T) := S\Phi(d_1) - K\mathrm{e}^{-r(T-t)}\Phi(d_2), \tag{2.7}$$

where Φ denotes the standard normal distribution function (df), r represents the continuously compounded risk-free interest rate, σ denotes the volatility of the underlying stock, and where

$$d_1 = \frac{\ln(S/K) + (r + \frac{1}{2}\sigma^2)(T-t)}{\sigma\sqrt{T-t}} \quad \text{and} \quad d_2 = d_1 - \sigma\sqrt{T-t}. \tag{2.8}$$

For notational simplicity we assume that the time to maturity of the option, $T - t$, is measured in units of the time horizon, and that the parameters r and σ are expressed in terms of those units; for example, if the time horizon is one day, then r and σ are the daily interest rate and volatility. This differs from standard market practice where time is measured in years and r and σ are expressed in annualized terms.

To map the portfolio at time t, let S_t denote the stock price at time t and let r_t and σ_t denote the values that a practitioner chooses to use at that time for the interest rate and volatility. The log-price of the stock ($\ln S_t$) is an obvious risk factor for changes in value of the portfolio. While in the Black–Scholes option-pricing model the interest rate and volatility are assumed to be constant, in real markets interest rates change constantly, as do the *implied volatilities* that practitioners tend to use as inputs for the volatility parameter. Hence, we take $\boldsymbol{Z}_t = (\ln S_t, r_t, \sigma_t)'$ as the vector of risk factors.

According to the Black–Scholes formula the value of the call option at time t equals $C^{\mathrm{BS}}(t, S_t; r_t, \sigma_t, K, T)$, which is of the form (2.2). The risk-factor changes are given by

$$\boldsymbol{X}_{t+1} = (\ln S_{t+1} - \ln S_t, r_{t+1} - r_t, \sigma_{t+1} - \sigma_t)',$$

and the linearized loss can be calculated to be

$$L^{\Delta}_{t+1} = -(C^{\mathrm{BS}}_t + C^{\mathrm{BS}}_S S_t X_{t+1,1} + C^{\mathrm{BS}}_r X_{t+1,2} + C^{\mathrm{BS}}_\sigma X_{t+1,3}), \tag{2.9}$$

where the subscripts denote partial derivatives of the Black–Scholes formula (2.7). Note that we have omitted the arguments of C^{BS} to simplify the notation. Note also that an S_t term appears because we take the equity risk factor to be the log-price of the stock rather than the price; applying the chain rule with $S = \mathrm{e}^{z_1}$ we have

$$C^{\mathrm{BS}}_{z_1} = C^{\mathrm{BS}}_S \frac{\mathrm{d}S}{\mathrm{d}z_1}\bigg|_{z_1 = \ln S_t} = C^{\mathrm{BS}}_S S_t.$$

In Section 9.1.2 and Example 9.1 we give more detail concerning the derivation of mapping formulas similar to (2.9) and pay more attention to the choice of timescale in the mapping function.

The derivatives of the Black–Scholes option-pricing function are often referred to as the *Greeks*: C^{BS}_S (the partial derivative with respect to the stock price S) is called the *delta* of the option; C^{BS}_t (the partial derivative with respect to time) is called the *theta* of the option; C^{BS}_r (the partial derivative with respect to the interest rate r) is called the *rho* of the option; and, in a slight abuse of the Greek language, C^{BS}_σ (the partial derivative with respect to volatility σ) is called the *vega* of the option. The Greeks play an important role in the risk management of derivative portfolios.

The reader should keep in mind that for portfolios containing derivatives, the linearized loss can be a rather poor approximation of the true loss, since the portfolio value is often a highly nonlinear function of the risk factors. This has led to the development of higher-order approximations such as the *delta–gamma approximation*, where first- and second-order derivatives are used (see Section 9.1.2).

Example 2.3 (stylized loan portfolio). In this example we show how losses from a portfolio of short-term loans fit into our general framework; a detailed discussion of models for loan portfolios will be presented in Chapter 11.

Following standard practice in credit risk management, the risk-management horizon Δt is taken to be one year. We consider a portfolio of loans to m different borrowers or obligors that have been made at time t and valued using a book-value approach. To keep the example simple we assume that all loans have to be repaid at time $t+1$. We denote the amount to be repaid by obligor i by k_i; this term comprises the interest payment at $t+1$ and the repayment of the loan principal. The *exposure* to obligor i is defined to be the present value of the promised interest and principal cash flows, and it is therefore given by $e_i = k_i/(1+r_{t,1})$.

In order to take the possibility of default into account we introduce a series of random variables $(Y_{t,i})_{t\in\mathbb{N}}$ that represent the default state of obligor i at t, and we let $Y_{t,i} = 1$ if obligor i has defaulted by time t, with $Y_{t,i} = 0$ otherwise. These variables are known as *default indicators*. For simplicity we assume that all obligors are in a non-default state at time t, so $Y_{t,i} = 0$ for all $1 \leqslant i \leqslant m$.

In keeping with valuation conventions, in practice we define the book value of a loan to be the exposure of the loan reduced by the discounted expected loss due to default; in this way the valuation includes a provision for default risk. We assume that in the case of a default of borrower i, the lender can recover an amount $(1-\delta_i)k_i$ at the maturity date $t+1$, where $\delta_i \in (0,1]$ describes the so-called *loss given default* of the loan, which is the percentage of the exposure that is lost in the event of default. Moreover, we denote by p_i the probability that obligor i defaults in the period $(t, t+1]$. In this introductory example we suppose that δ_i and p_i are known constants. In practice, p_i could be estimated using a credit scoring model (see Section 10.2 for more discussion). The discounted expected loss due to a default of obligor i is thus given by

$$\frac{1}{1+r_{t,1}}\delta_i p_i k_i = \delta_i p_i e_i.$$

The book value of loan i is therefore equal to $e_i(1-\delta_i p_i)$, the discounted expected pay-off of the loan. Note that in practice, one would make further provisions for administrative, refinancing and capital costs, but we ignore these issues for the sake of simplicity. Moreover, one should keep in mind that the book value is not an estimate for the fair value of the loan (an estimate for the amount for which the loan could be sold in a securitization deal); the latter is usually lower than the discounted expected pay-off of the loan, as investors demand a premium for bearing the default risk (see also our discussion of risk-neutral valuation in the next section). The book value of the loan portfolio at time t is thus given by

$$V_t = \sum_{i=1}^{m} e_i(1-\delta_i p_i).$$

The value of a loan to obligor i at the maturity date $t+1$ equals the size of the repayment and is therefore equal to k_i if $Y_{t+1,i} = 0$ (no default of obligor i) and equal to $(1-\delta_i)k_i$ if $Y_{t+1,i}=1$ (default of obligor i). Hence, V_{t+1}, the value of the portfolio at time $t+1$, equals

$$V_{t+1} = \sum_{i=1}^{m}((1-Y_{t+1,i})k_i + Y_{t+1,i}(1-\delta_i)k_i) = \sum_{i=1}^{m} k_i(1-\delta_i Y_{t+1,i}).$$

Since we use a relatively long risk-management horizon of one year, it is natural to discount V_{t+1} in computing the portfolio loss. Again using the fact that $e_i = k_i/(1+r_{t,1})$, we obtain

$$\begin{aligned} L_{t+1} = V_t - \frac{V_{t+1}}{1+r_{t,1}} &= \sum_{i=1}^{m} e_i(1-\delta_i p_i) - \sum_{i=1}^{m} e_i(1-\delta_i Y_{t+1,i}) \\ &= \sum_{i=1}^{m} \delta_i e_i Y_{t+1,i} - \sum_{i=1}^{m} \delta_i e_i p_i, \end{aligned}$$

which gives a simple formula for the portfolio loss involving exposures, default probabilities, losses given default and default indicators.

Finally, we explain how this example fits into the mapping framework given by (2.2) and (2.3). In this case the risk factors are the default indicator variables $\mathbf{Z}_t = (Y_{t,1},\dots,Y_{t,m})'$. If we write the mapping formula as

$$f(s,\mathbf{Z}_s) = \sum_{i=1}^{m}(1-Y_{s,i})\frac{k_i}{(1+r_{s,1})^{t+1-s}}(1-(t+1-s)\delta_i p_i) + \sum_{i=1}^{m} Y_{s,i}k_i(1-\delta_i),$$

we see that this gives the correct portfolio values at times $s=t$ and $s=t+1$. The issue of finding and calibrating a good model for the joint distribution of the risk-factor changes $\mathbf{Z}_{t+1}-\mathbf{Z}_t$ is taken up in Chapter 11.

Example 2.4 (insurance example). We consider a simple whole-life annuity product in which a policyholder (known as an annuitant) has purchased the right to receive a series of payments as long as he or she remains alive. Although realistic products would typically make monthly payments, we assume annual payments for simplicity and consider a risk-management horizon of one year.

At time t we assume that an insurer has a portfolio of n annuitants with current ages x_i, $i=1,\dots,n$. Annuitant i receives a fixed annual annuity payment κ_i, and the time of their death is represented by the random variable τ_i. The annuity payments are made in arrears at times $t+1, t+2, \dots$, a form of product known as a whole-life immediate annuity.

At time t there is uncertainty about the value of the cash flow to any individual annuitant stemming from the uncertainty about their time of death. The liability due to a single annuitant takes the form

$$\sum_{h=1}^{\infty} I_{\{\tau_i > t+h\}}\kappa_i D(t,t+h),$$

where $D(t,t+h)$ is a discount factor that gives the time-t value of one unit paid out at time $t+h$. Following standard discrete-time actuarial practice we set $D(t,t+h) =$

$(1+r_{t,h})^{-h}$, where $r_{t,h}$ is the h-year simple spot interest rate at time t. The expected present value of this liability at time t is given by

$$\sum_{h=1}^{\infty} q_t(x_i, h)\kappa_i \frac{1}{(1+r_{t,h})^h},$$

where we assume that $P(\tau_i > t + h) = q_t(x_i, h)$. In other words, the survival probability of annuitant i depends on only the current time t and the age x_i of the annuitant; $q_t(x, h)$ represents the probability that an individual aged x at time t will survive a further h years.

If n is sufficiently large, diversification arguments suggest that individual mortality risk (deviation of the variables $\tau_1, \dots, \tau_n$ from their expected values) may be neglected, and the overall portfolio liability may be represented by

$$B_t = \sum_{i=1}^{n}\sum_{h=1}^{\infty} q_t(x_i, h)\kappa_i \frac{1}{(1+r_{t,h})^h}.$$

Now consider the liability due to a single annuitant at time $t+1$, which is given by

$$\sum_{h=1}^{\infty} I_{\{\tau_i > t+1+h\}}\kappa_i \frac{1}{(1+r_{t+1,h})^h}.$$

We again use the large-portfolio diversification argument to replace $I_{\{\tau_i > t+1+h\}}$ by its expected value $q_t(x_i, h+1)$, and thus we approximate the portfolio liability at $t+1$ by

$$B_{t+1} = \sum_{i=1}^{n}\sum_{h=1}^{\infty} q_t(x_i, h+1)\kappa_i \frac{1}{(1+r_{t+1,h})^h}.$$

The lump-sum premium payments of the annuitants would typically be invested in a matching portfolio of bonds: that is, a portfolio chosen so that the cash flows from the bonds closely match the cash flows due to the policyholders. We assume that the investments have been made in (default-free) government bonds with d different maturities (all greater than or equal to one year) so that the asset value at time t is

$$A_t = \sum_{j=1}^{d} \frac{\lambda_j}{(1+r_{t,h_j})^{h_j}},$$

where h_j is the maturity of the jth bond and λ_j is the number of such bonds that have been purchased. The net asset value of the portfolio at time t is given by $V_t = A_t - B_t$.

This is a situation in which it would be natural to discount future asset and liability values back to time t, so that the loss (in units of time-t money) would be given by

$$L_{t+1} = -\left(\frac{A_{t+1}}{1+r_{t,1}} - A_t\right) + \left(\frac{B_{t+1}}{1+r_{t,1}} - B_t\right).$$

The risk factors in this example are the spot rates $\boldsymbol{Z}_t = (r_{t,1}, \dots, r_{t,m})'$, where m represents the maximum time horizon at which an annuity payment might have to be made. The mortality risk in the lifetime variables $\tau_1, \dots, \tau_m$ is eliminated from consideration by using the *life table* of fixed survival probabilities $\{q_t(x, h)\}$.

2.2.2 Valuation Methods

We now take a closer look at valuation principles in the light of the stylized examples of the previous section. While the loan portfolio of Example 2.3 would typically be valued using a book-value approach, as indicated, the stock portfolio (Example 2.1), the European call option (Example 2.2) and the asset-backed annuity portfolio (Example 2.4) would all be valued using a fair-value approach in practice. In this section we elaborate on the different methods used in fair-value accounting and explain how risk-neutral valuation may be understood as a special case of fair-value accounting.

We recall from Section 2.1.1 that the use of fair-value methodology for the assets and liabilities of an insurer is closely related to the concept of market-consistent valuation. The main practical difference is that the fair-value approach is applied to the accounting balance sheet for reporting purposes, whereas market-consistent valuation is applied to the Solvency II balance sheet for capital adequacy purposes. While there are differences in detail between the two rule books, it is sufficient for our purposes to view market-consistent valuation as a variant of fair-value accounting.

Fair-value accounting. In general terms, the fair value of an asset is an estimate of the price that would be received in selling the asset in a transaction on an active market. Similarly, the fair value of a liability is an estimate of the price that would have to be paid to transfer the liability to another party in a market-based transaction; this is sometimes referred to as the exit value.

Only a minority of balance sheet positions are traded directly in an active market. Accountants have therefore developed a three-stage hierarchy of fair-value accounting methods, extending fair-value accounting to non-traded items. This hierarchy, which is codified in the US as Financial Accounting Standard 157 and worldwide in the 2009 amendment to International Financial Reporting Standard 7, has the following levels.

Level 1: the fair value of an instrument is determined from quoted prices in an active market for the same instrument, without modification or repackaging.

Level 2: the fair value of an instrument is determined using quoted prices in active markets for similar (but not identical) instruments or by the use of valuation techniques, such as pricing models for derivatives, for which all significant inputs are based on observable market data.

Level 3: the fair value of an instrument is estimated using a valuation technique (pricing model) for which some key inputs are not observable market data (or otherwise publicly observable quantities).

In risk-management language these levels are sometimes described as mark-to-market, mark-to-model with objective inputs, and mark-to-model with subjective inputs.

The stock portfolio in Example 2.1 is a clear example of Level 1 valuation: the portfolio value is determined by simply looking up current market prices of the stocks.

Now consider the European call option of Example 2.2 and assume that the option is not traded on the market, perhaps because of a non-standard strike price or maturity, but that there is an otherwise active market for options on that stock. This would be an example of Level 2 valuation: a valuation technique (namely, the Black–Scholes option-pricing formula) is used to price the instrument. The inputs to the formula are the stock price, the interest rate and the implied volatility, which are market observables (since we assumed that there is an active market for options on the stock).

The insurance portfolio from Example 2.4 can be viewed as an example of Level 2 or Level 3 valuation, depending on the methods used to determine the input parameters. If the survival probabilities are determined from publicly available sources such as official life tables, the annuity example corresponds to Level 2 valuation since the other risk factors are essentially market observables, with the possible exception of long-term interest rates. If, on the other hand, proprietary data and methods are used to estimate the survival probabilities, then this would be Level 3 valuation.

Risk-neutral valuation. Risk-neutral valuation is a special case of fair-value accounting that is widely used in the pricing of financial products such as derivative securities. In risk-neutral pricing the values of financial instruments are computed as expected discounted values of future cash flows, where expectation is taken with respect to some probability measure Q, called a *risk-neutral pricing measure*. Q is an artificial measure that turns the discounted prices of traded securities into so-called martingales (fair bets), and it is also known as an *equivalent martingale measure*. Calibration procedures are used to ensure that prices obtained in this way are consistent with quoted market prices.

Hitherto, all our probabilities and expectations have been taken with respect to the *physical* or *real-world measure* P. In order to explain the concept of a risk-neutral measure Q and to illustrate the relationship between P and Q, we use a simple one-period model from the field of credit risk, which we refer to as the basic one-period default model. We consider a defaultable zero-coupon bond with maturity T equal to one year and make the following assumptions: the real-world default probability is $p = 1\%$; the recovery rate $1 - \delta$ (the proportion of the notional of the bond that is paid back in the case of a default) is deterministic and is equal to 60%; the risk-free simple interest rate equals 5%; the current ($t = 0$) price of the bond is $p_1(0, 1) = 0.941$; the price of the corresponding default-free bond is $p_0(0, 1) = (1.05)^{-1} = 0.952$. The price evolution of the bond is depicted in Figure 2.1.

The expected discounted value of the bond equals $(1.05)^{-1}(0.99 \cdot 1 + 0.01 \cdot 0.6) = 0.949 > p_1(0, 1)$. We see that in this example the price $p_1(0, 1)$ is smaller than the expected discounted value of the claim. This is the typical situation in real markets for corporate bonds, as investors demand a premium for bearing the default risk of the bond.

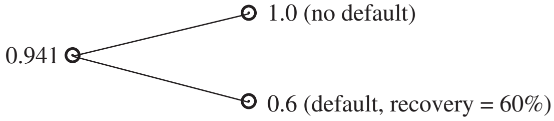

Figure 2.1. Evolution of the price $p_1(\cdot, 1)$ of a defaultable bond in the basic one-period default model; the probabilities of the upper and lower branches are 0.99 and 0.01, respectively.

In a one-period model, an equivalent martingale measure or risk-neutral measure is simply a new probability measure Q such that for every traded security the Q-expectation of the discounted pay-off equals the current price of the security, so that investing in this security becomes a fair bet. In more general situations (for example, in continuous-time models), the idea of a fair bet is formalized by the requirement that the discounted price process of a traded security is a so-called Q-martingale (hence the name martingale measure). In the basic one-period default model, Q is thus given in terms of an artificial default probability q such that

$$p_1(0, 1) = (1.05)^{-1}((1 - q) \cdot 1 + q \cdot 0.6).$$

Clearly, q is uniquely determined by this equation and we get that $q = 0.03$. Note that, in our example, q is bigger than the physical default probability $p = 0.01$; again, this is typical for real markets and reflects the risk premium demanded by buyers of defaultable bonds. The example also shows that different approaches are needed in order to determine the historical default probability p and the risk-neutral default probability q: the former is *estimated* from historical data such as the default history of firms of similar credit quality (see, for example, Sections 10.3.3 and 11.5), whereas q is *calibrated* to market prices of traded securities.

Under the risk-neutral pricing approach, the price of a security is computed as the (conditional) expected value of the discounted future cash flows, where expectation is taken with respect to the risk-neutral measure Q. Denoting the pay-off of the security at $t = 1$ by the rv H and the risk-free simple interest rate between time 0 and time 1 by $r_{0,1} \geqslant 0$, we obtain the following formula for the value V_0^H of the claim H at $t = 0$:

$$V_0^H = E^Q\left(\frac{H}{1 + r_{0,1}}\right). \tag{2.10}$$

For a specific example in the basic one-period default model, consider a *default put option* that pays one unit at $t = 1$ if the bond defaults and zero otherwise; the option can be thought of as a simplified version of a credit default swap. Using risk-neutral pricing, the value of the option at $t = 0$ is given by

$$V_0 = (1.05)^{-1}((1 - q) \cdot 0 + q \cdot 1) = (1.05)^{-1}0.03 = 0.0285.$$

In continuous-time models one usually uses continuous compounding, and (2.10) is therefore replaced by the slightly more general expression

$$V_t^H = E_t^Q(\mathrm{e}^{-r(T-t)}H), \quad t < T. \tag{2.11}$$

Here, T is the maturity date of the security and the subscript t on the expectation operator indicates that the expectation is taken with respect to the information available to investors at time t, as will be explained in more detail in Chapter 10.

Formulas (2.10) and (2.11) are known as *risk-neutral pricing rules*. Risk-neutral pricing applied to non-traded financial products is a typical example of Level 2 valuation: prices of traded securities are used to calibrate model parameters under the risk-neutral measure Q; this measure is then used to price the non-traded products. We give one example that underscores our use of the Black–Scholes pricing rule in Example 2.2.

Example 2.5 (European call option in Black–Scholes model). Consider again the European call option in Example 2.2 and suppose that options with our desired strike K and/or maturity time T are not traded, but that other options on the same stock are traded. We assume that under the real-world probability measure P the stock price (S_t) follows a geometric Brownian motion model (the so-called Black–Scholes model) with dynamics given by

$$\mathrm{d}S_t = \mu S_t \,\mathrm{d}t + \sigma S_t \,\mathrm{d}W_t$$

for constants $\mu \in \mathbb{R}$ (the drift) and $\sigma > 0$ (the volatility), and a standard Brownian motion (W_t). It is well known that there is an equivalent martingale measure Q under which the discounted stock price $(\mathrm{e}^{-rt} S_t)$ is a martingale; under Q, the stock price follows a geometric Brownian motion model with drift r and volatility σ. The European call option pay-off is $H = (S_T - K)^+$ and the risk-neutral valuation formula in (2.11) may be shown to take the form

$$V_t = E_t^Q(\mathrm{e}^{-r(T-t)}(S_T - K)^+) = C^{\mathrm{BS}}(t, S_t; r, \sigma, K, T), \quad t < T, \tag{2.12}$$

with C^{BS} as in Example 2.2. To assign a risk-neutral value to the call option at time t (knowing the current price of the stock S_t, the interest rate r and the option characteristics K and T), we need to calibrate the model parameter σ. As discussed above, we would typically use quoted prices $C^{\mathrm{BS}}(t, S_t; r, \sigma, K^*, T^*)$ for options on the stock with different characteristics to infer a value for σ and then plug the so-called implied volatility into (2.12).

There are two theoretical justifications for risk-neutral pricing. First, a standard result of mathematical finance (the so-called *first fundamental theorem of asset pricing*) states that a model for security prices is arbitrage free if and only if it admits at least one equivalent martingale measure Q. Hence, if a financial product is to be priced in accordance with no-arbitrage principles, its price must be given by the risk-neutral pricing formula for some risk-neutral measure Q. A second justification refers to hedging: in financial models it is often possible to replicate the pay-off of a financial product by trading in the assets, a practice known as *(dynamic) hedging*, and it is well known that in a frictionless market the cost of carrying out such a hedge is given by the risk-neutral pricing rule. Advantages and limitations of risk-neutral pricing will be discussed in more detail in Section 10.4.2.

2.2.3 Loss Distributions

Having mapped the risks of a portfolio, we now consider how to derive loss distributions with a view to using them in risk-management applications such as capital setting. Assuming the current time is t and recalling formula (2.3) for the loss over the time period $[t, t+1]$,

$$L_{t+1} = -\Delta V_{t+1} = -(f(t+1, z_t + \boldsymbol{X}_{t+1}) - f(t, z_t)),$$

we see that in order to determine the loss distribution (i.e. the distribution of L_{t+1}) we need to do two things: (i) specify a model for the risk-factor changes $\boldsymbol{X}_{t+1}$; and (ii) determine the distribution of the rv $f(t+1, z_t + \boldsymbol{X}_{t+1})$.

Note that effectively two kinds of model enter into this process. The models used in (i) are *projection models* used to forecast the behaviour of risk factors in the real world, and they are generally estimated from empirical data describing past risk-factor changes $(\boldsymbol{X}_s)_{s \leqslant t}$. Depending on the complexity of the positions involved, the mapping function f in (ii) will typically also embody *valuation models*; consider in this context the use of the Black–Scholes model to value a European call option, as described in Examples 2.2 and 2.5.

Broadly speaking, there are three kinds of method that can be used to address these challenges: an analytical method, a method based on the idea of historical simulation, or a simulation approach (also known as a Monte Carlo method).

Analytical method. In an analytical method we attempt to choose a model for $\boldsymbol{X}_{t+1}$ and a mapping function f in such a way that the distribution of L_{t+1} can be determined analytically. A prime example of this approach is the so-called variance–covariance method for market-risk management, which dates back to the early work of the RiskMetrics Group (JPMorgan 1996). In the variance–covariance method the risk-factor changes $\boldsymbol{X}_{t+1}$ are assumed to follow a multivariate normal distribution, denoted by $\boldsymbol{X}_{t+1} \sim N_d(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}$ is the mean vector and Σ the covariance (or variance–covariance) matrix of the distribution. This would follow, for example, from assuming that the risk factors $\boldsymbol{Z}_t$ evolve in continuous time according to a multivariate Brownian motion. The properties of the multivariate normal distribution are discussed in detail in Section 6.1.3.

We also assume that the linearized loss in terms of the risk factors is a sufficiently accurate approximation of the actual loss and simplify the problem by considering the distribution of L^{Δ}_{t+1} defined in (2.4). The linearized loss will have general structure

$$L^{\Delta}_{t+1} = -(c_t + \boldsymbol{b}_t' \boldsymbol{X}_{t+1}) \tag{2.13}$$

for some constant c_t and constant vector $\boldsymbol{b}_t$, which are known to us at time t. For a concrete example, consider the stock portfolio of Example 2.1, where the loss takes the form $L^{\Delta}_{t+1} = -v_t \boldsymbol{w}_t' \boldsymbol{X}_{t+1}$ and $\boldsymbol{w}_t$ is the vector of portfolio weights at time t.

An important property of the multivariate normal distribution is that a linear function (2.13) of $\boldsymbol{X}_{t+1}$ must have a univariate normal distribution. From general rules for calculating the mean and variance of linear combinations of a random vector we obtain that

$$L^{\Delta}_{t+1} \sim N(-c_t - \boldsymbol{b}'_t \boldsymbol{\mu}, \boldsymbol{b}'_t \Sigma \boldsymbol{b}_t). \tag{2.14}$$

The variance–covariance method offers a simple solution to the risk-measurement problem, but this convenience is achieved at the cost of two crude simplifying assumptions. First, linearization may not always offer a good approximation of the relationship between the true loss distribution and the risk-factor changes. Second, the assumption of normality is unlikely to be realistic for the distribution of the risk-factor changes, certainly for daily data and probably also for weekly and even monthly data. A stylized fact of empirical finance suggests that the distribution of financial risk-factor returns is leptokurtic and heavier tailed than the Gaussian distribution. In Section 3.1.2 we will present evidence for this observation in an analysis of daily, weekly, monthly and quarterly stock returns. The implication is that an assumption of Gaussian risk factors will tend to underestimate the tail of the loss distribution and thus underestimate the risk of the portfolio.

Remark 2.6. Note that we postpone a detailed discussion of how the model parameters $\boldsymbol{\mu}$ and Σ are estimated from historical risk-factor changes $(\boldsymbol{X}_s)_{s\leqslant t}$ until later chapters. We should, however, point out that when a dynamic model for $\boldsymbol{X}_{t+1}$ is considered, different estimation methods are possible depending on whether we focus on the conditional distribution of $\boldsymbol{X}_{t+1}$ given past values of the process or whether we consider the equilibrium distribution in a stationary model. These different approaches are said to constitute conditional and unconditional methods of computing the loss distribution—an issue we deal with in much more detail in Chapter 9.

Historical simulation. Instead of estimating the distribution of L_{t+1} in some explicit parametric model for $\boldsymbol{X}_{t+1}$, the historical-simulation method can be thought of as estimating the distribution of the loss using the *empirical distribution* of past risk-factor changes. Suppose we collect historical risk-factor change data over n time periods and denote these data by $\boldsymbol{X}_{t-n+1}, \dots, \boldsymbol{X}_t$. In historical simulation we construct the following univariate data set of imaginary losses:

$$\{\tilde{L}_s = -(f(t+1, \boldsymbol{z}_t + \boldsymbol{X}_s) - f(t, \boldsymbol{z}_t))\colon s = t-n+1, \dots, t\}.$$

The values $\tilde{L}_s$ show what would happen to the current portfolio if the risk-factor changes in period s were to recur. If we assume that the process of risk-factor changes is stationary with df $F_{\boldsymbol{X}}$, then (subject to further technical conditions) the empirical df of the historically simulated losses is a consistent estimator of the loss distribution. Estimators for any statistic of the loss distribution—such as the expected loss, the variance of the loss, or the value-at-risk (see Section 2.3.2 for a definition)—can be computed from the empirical df of the historically simulated losses. For instance, the expected loss can be estimated by $E(L_{t+1}) \approx n^{-1} \sum_{s=t-n+1}^{t} \tilde{L}_s$, and techniques like empirical quantile estimation can be used to derive estimates of value-at-risk. Further details can be found in Chapter 9.

The historical-simulation method has obvious attractions: it is easy to implement and it reduces the loss distribution calculation problem to a one-dimensional problem. However, the success of the approach is dependent on our ability to collect sufficient quantities of relevant synchronized historical data for all risk factors. As such, the method is mainly used in market-risk management for banks, where the issue of data availability is less of a problem (due to the relatively short risk-management time horizon).

Monte Carlo method. Any approach to risk measurement that involves the simulation of an explicit parametric model for risk-factor changes is known as a Monte Carlo method. The method does not solve the problem of finding a multivariate model for $\boldsymbol{X}_{t+1}$, and any results that are obtained will only be as good as the model that is used. For large portfolios the computational cost of the Monte Carlo approach can be considerable, as every simulation requires the revaluation of the portfolio. This is particularly problematic if the portfolio contains many derivatives that cannot be priced in closed form. Such derivative positions might have to be valued using Monte Carlo approximation techniques, which are also based on simulations. This leads to situations where Monte Carlo procedures are *nested* and simulations are being generated within simulations, which can be very slow.

Simulation techniques are frequently used in the management of credit portfolios (see, for example, Section 11.4). So-called *economic scenario generation* models, which are used in insurance, also fall under the heading of Monte Carlo methods. These are economically motivated and (typically) dynamic models for the evolution and interaction of different risk factors, and they can be used to generate realizations of $\boldsymbol{X}_{t+1}$.

Notes and Comments

The concept of mapping portfolio values to fundamental risk factors was pioneered by the RiskMetrics Group: see the RiskMetrics Technical Document (JPMorgan 1996) and Mina and Xiao (2001). We explore the topic in more detail, with further examples, in Chapter 9. Other textbooks that treat the mapping of positions include Dowd (1998), Jorion (2007) and Volume III of *Market Risk Analysis* by Alexander (2009). The use of first-order approximations to the portfolio value (the so-called delta approximation) may be found in Duffie and Pan (1997); for second-order approximations, see Section 9.1.2.

More details of the Black–Scholes valuation formula used in Example 2.2 may be found in many texts on options and derivatives, such as Haug (1998), Wilmot (2000) and Hull (2014). For annuity products similar to the one analysed in Example 2.4 and other standard life insurance products, good references are Hardy (2003), Møller and Steffensen (2007), Koller (2011), Dickson, Hardy and Waters (2013) and the classic book by Gerber (1997).

The best resource for more on International Financial Reporting Standard 7 and the fair-value accounting of financial instruments is the International Financial Reporting Standards website at www.ifrs.org. Shaffer (2011) considers the impact of fair-value accounting on financial institutions, while Laux and Leuz (2010) address the issue of whether fair-value accounting may have contributed to the financial crisis.

Market-consistent actuarial valuation in insurance is the subject of a textbook by Wüthrich, Bühlmann and Furrer (2010) (see also Wüthrich and Merz 2013). The fundamental theorem of asset pricing and the conceptual underpinnings of risk-neutral pricing are discussed in most textbooks on mathematical finance: see, for example, Björk (2004) or Shreve (2004b).

The analytical method for deriving loss distributions based on an assumption of normal risk-factor changes belongs to the original RiskMetrics methodology cited above. For the analysis of the distribution of losses in a bank's trading book, this has largely been supplanted by the use of historical simulation (Pérignon and Smith 2010). Monte Carlo approaches (economic scenario generators) are widely used in internal models for Solvency II in the insurance industry (see Varnell 2011).

2.3 Risk Measurement

In very general terms a risk measure associates a financial position with loss L with a real number that measures the "riskiness of L". In practice, risk measures are used for a variety of purposes. To begin with, they are used to determine the amount of capital a financial institution needs to hold as a buffer against unexpected future losses on its portfolio in order to satisfy a regulator who is concerned with the solvency of the institution. Similarly, they are used to determine appropriate margin requirements for investors trading at an organized exchange. Moreover, risk measures are often used by management as a tool for limiting the amount of risk a business unit within a firm may take. For instance, traders in a bank might be constrained by the rule that the daily 95% value-at-risk of their position should not exceed a given bound.

In Section 2.3.1 we give an overview of some different approaches to measuring risk before focusing on risk measures that are derived from loss distributions. We introduce the widely used value-at-risk measure in Section 2.3.2 and explain how VaR features in risk capital calculations in Section 2.3.3. In Section 2.3.4 alternative risk measures derived from loss distributions are presented, and in Section 2.3.5 an introduction to the subject of desirable risk measure properties is given, in which the notions of coherent and convex risk measures are defined and examples are discussed.

2.3.1 Approaches to Risk Measurement

Existing approaches to measuring the risk of a financial position can be grouped into three categories: the notional-amount approach, risk measures based on loss distributions, and risk measures based on scenarios.

Notional-amount approach. This is the oldest approach to quantifying the risk of a portfolio of risky assets. In the notional-amount approach the risk of a portfolio is defined as the sum of the notional values of the individual securities in the portfolio, where each notional value may be weighted by a factor representing an assessment of the riskiness of the broad asset class to which the security or instrument belongs. An example of this approach is the so-called *standardized approach* in the Basel regulatory framework (see Section 1.3.1 for a general description and Section 13.1.2 for the standardized approach as it applies to operational risk).

The advantage of the notional-amount approach is its apparent simplicity. However, as noted in Section 1.3.1, the approach is flawed from an economic viewpoint for a number of reasons. To begin with, the approach does not differentiate between long and short positions and there is no netting. For instance, the risk of a long position in corporate bonds hedged by an offsetting position in credit default swaps would be counted as twice the risk of the unhedged bond position. Moreover, the approach does not reflect the benefits of diversification on the overall risk of the portfolio. For example, if we use the notional-amount approach, a well-diversified credit portfolio consisting of loans to many companies appears to have the same risk as a portfolio in which the whole amount is lent to a single company. Finally, the notional-amount approach has problems in dealing with portfolios of derivatives, where the notional amount of the underlying and the economic value of the derivative position can differ widely.

Risk measures based on loss distributions. Most modern measures of the risk in a portfolio are statistical quantities describing the conditional or unconditional loss distribution of the portfolio over some predetermined horizon Δt. Examples include the variance, the VaR and the ES risk measures, which we discuss in more detail later in this chapter. Risk measures based on loss distributions have a number of advantages. The concept of a loss distribution makes sense on all levels of aggregation, from a portfolio consisting of a single instrument to the overall position of a financial institution. Moreover, if estimated properly, the loss distribution reflects netting and diversification effects.

Two issues should be borne in mind when working with loss distributions. First, any estimate of the loss distribution is based on past data. If the laws governing financial markets change, these past data are of limited use in predicting future risk. Second, even in a stationary environment it is difficult to estimate the loss distribution accurately, particularly for large portfolios. Many seemingly sophisticated risk-management systems are based on relatively crude statistical models for the loss distribution (incorporating, for example, untenable assumptions of normality). These issues call for continual improvements in the way that loss distributions are estimated and, of course, for prudence in the practical application of risk-management models based on estimated loss distributions. In particular, risk measures based on the loss distribution should be complemented by information from hypothetical scenarios. Moreover, forward-looking information reflecting the expectations of market participants, such as implied volatilities, should be used in conjunction with statistical estimates (which are necessarily based on past information) in calibrating models of the loss distribution.

Scenario-based risk measures. In the scenario-based approach to measuring the risk of a portfolio, one considers a number of possible future risk-factor changes (scenarios), such as a 10% rise in key exchange rates, a simultaneous 20% drop in major stock market indices or a simultaneous rise in key interest rates around the globe. The risk of the portfolio is then measured as the maximum loss of the portfolio under all scenarios. The scenarios can also be weighted for plausibility. This approach to risk measurement is the one that is typically adopted in stress testing.

We now give a formal description. Fix a set $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ of risk-factor changes (the scenarios) and a vector $\boldsymbol{w} = (w_1, \dots, w_n)' \in [0,1]^n$ of weights. Denote by $L(\boldsymbol{x})$ the loss the portfolio would suffer if the hypothetical scenario $\boldsymbol{x}$ were to occur. Using the notation of Section 2.2.1 we get

$$L(\boldsymbol{x}) := -(f(t+1, z_t + \boldsymbol{x}) - f(t, z_t)), \quad \boldsymbol{x} \in \mathbb{R}^d.$$

The risk of the portfolio is then measured by

$$\psi_{[\mathcal{X},\boldsymbol{w}]} := \max\{w_1 L(\boldsymbol{x}_1), \dots, w_n L(\boldsymbol{x}_n)\}. \tag{2.15}$$

Many risk measures that are used in practice are of the form (2.15). The following is a simplified description of a system for determining margin requirements developed by the Chicago Mercantile Exchange (see Chicago Mercantile Exchange 2010). To compute the initial margin for a simple portfolio consisting of a position in a futures contract and call and put options on this contract, sixteen different scenarios are considered. The first fourteen consist of an up move or a down move of volatility combined with no move, an up move or a down move of the futures price by $\frac{1}{3}$, $\frac{2}{3}$ or $\frac{3}{3}$ of a unit of a specified range. The weights w_i, $i = 1, \dots, 14$, of these scenarios are equal to 1. In addition, there are two extreme scenarios with weights $w_{15} = w_{16} = 0.35$. The amount of capital required by the exchange as margin for the portfolio is then computed according to (2.15).

Remark 2.7. We can give a slightly different mathematical interpretation to formula (2.15), which will be useful in Section 2.3.5. Assume for the moment that $L(\mathbf{0}) = 0$, i.e. that the value of the position is unchanged if all risk factors stay the same. This is reasonable, at least for a short risk-management horizon Δt. In that case, the expression $w_i L(\boldsymbol{x}_i)$ can be viewed as the expected value of L under a probability measure on the space of risk-factor changes; this measure associates a mass of $w_i \in [0,1]$ to the point $\boldsymbol{x}_i$ and a mass of $1 - w_i$ to the point $\mathbf{0}$. Denote by $\delta_{\boldsymbol{x}}$ the probability measure associating a mass of one to the point $\boldsymbol{x} \in \mathbb{R}^d$ and by $\mathcal{P}_{[\mathcal{X},\boldsymbol{w}]}$ the following set of probability measures on $\mathbb{R}^d$:

$$\mathcal{P}_{[\mathcal{X},\boldsymbol{w}]} = \{w_1\delta_{\boldsymbol{x}_1} + (1 - w_1)\delta_{\mathbf{0}}, \dots, w_n\delta_{\boldsymbol{x}_n} + (1 - w_n)\delta_{\mathbf{0}}\}.$$

Then $\psi_{[\mathcal{X},\boldsymbol{w}]}$ can be written as

$$\psi_{[\mathcal{X},\boldsymbol{w}]} = \max\{E^P(L(\boldsymbol{X})) \colon P \in \mathcal{P}_{[\mathcal{X},\boldsymbol{w}]}\}. \tag{2.16}$$

A risk measure of the form (2.16), where $\mathcal{P}_{[\mathcal{X},\boldsymbol{w}]}$ is replaced by some arbitrary subset $\mathcal{P}$ of the set of all probability measures on the space of risk-factor changes, is termed a *generalized scenario*. Generalized scenarios play an important role in the theory of coherent risk measures (see Section 8.1).

Scenario-based risk measures are a very useful risk-management tool for portfolios exposed to a relatively small set of risk factors, as in the Chicago Mercantile Exchange example. Moreover, they provide useful complementary information to measures based on statistics of the loss distribution. The main problem in setting up a scenario-based risk measure is, of course, determining an appropriate set of scenarios and weighting factors.

2.3.2 Value-at-Risk

VaR is probably the most widely used risk measure in financial institutions. It has a prominent role in the Basel regulatory framework and has also been influential in Solvency II.

Consider a portfolio of risky assets and a fixed time horizon Δt, and denote by $F_L(l) = P(L \leqslant l)$ the df of the corresponding loss distribution. We want to define a statistic based on F_L that measures the severity of the risk of holding our portfolio over the time period Δt. An obvious candidate is the maximum possible loss, given by $\inf\{l \in \mathbb{R}\colon F_L(l) = 1\}$. However, for most distributions of interest, the maximum loss is infinity. Moreover, by using the maximum loss, any probability information in F_L is neglected. The idea in the definition of VaR is to replace "maximum loss" by "maximum loss that is not exceeded with a given high probability".

Definition 2.8 (value-at-risk). Given some confidence level $\alpha \in (0, 1)$, the VaR of a portfolio with loss L at the confidence level α is given by the smallest number l such that the probability that the loss L exceeds l is no larger than $1 - \alpha$. Formally,

$$\mathrm{VaR}_\alpha = \mathrm{VaR}_\alpha(L) = \inf\{l \in \mathbb{R}\colon P(L > l) \leqslant 1 - \alpha\} = \inf\{l \in \mathbb{R}\colon F_L(l) \geqslant \alpha\}. \tag{2.17}$$

In probabilistic terms, VaR is therefore simply a *quantile* of the loss distribution. Typical values for α are $\alpha = 0.95$ or $\alpha = 0.99$; in market-risk management, the time horizon Δt is usually one or ten days, while in credit risk management and operational risk management, Δt is usually one year. Note that by its very definition the VaR at confidence level α does not give any information about the severity of losses that occur with a probability of less than $1 - \alpha$. This is clearly a drawback of VaR as a risk measure. For a small case study that illustrates this problem numerically we refer to Example 2.16 below.

Figure 2.2 illustrates the notion of VaR. The probability density function of a loss distribution is shown with a vertical line at the value of the 95% VaR. Note that the mean loss is negative ($E(L) = -2.6$), indicating that we expect to make a profit, but the right tail of the loss distribution is quite long in comparison with the left tail. The 95% VaR value is approximately 2.2, indicating that there is a 5% chance that we lose at least this amount.

Since quantiles play an important role in risk management, we recall the precise definition.

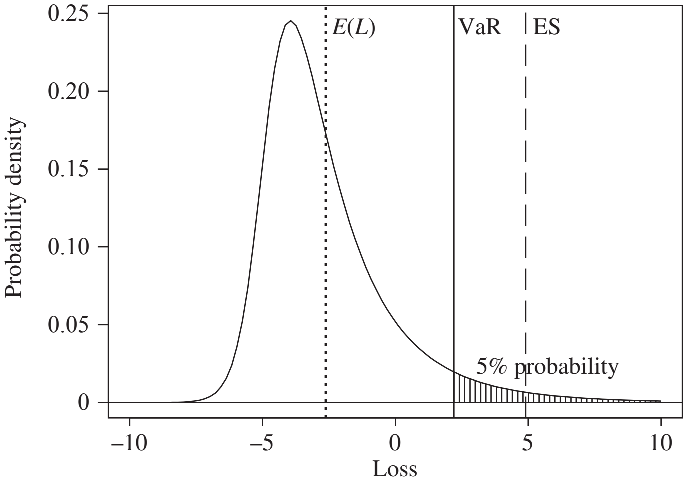

Figure 2.2. An example of a loss distribution with the 95% VaR marked as a vertical line; the mean loss is shown with a dotted line and an alternative risk measure known as the 95% ES (see Section 2.3.4 and Definition 2.12) is marked with a dashed line.

Definition 2.9 (the generalized inverse and the quantile function).

(i) Given some increasing function $T: \mathbb{R} \to \mathbb{R}$, the *generalized inverse* of T is defined by $T^{\leftarrow}(y) := \inf\{x \in \mathbb{R}: T(x) \geqslant y\}$, where we use the convention that the infimum of an empty set is ∞.

(ii) Given some df F, the generalized inverse $F^{\leftarrow}$ is called the *quantile function* of F. For $\alpha \in (0, 1)$ the α-quantile of F is given by

$$q_\alpha(F) := F^{\leftarrow}(\alpha) = \inf\{x \in \mathbb{R}: F(x) \geqslant \alpha\}.$$

For an rv X with df F we often use the alternative notation $q_\alpha(X) := q_\alpha(F)$. If F is continuous and strictly increasing, we simply have $q_\alpha(F) = F^{-1}(\alpha)$, where F^{-1} is the ordinary inverse of F. To compute quantiles in more general cases we may use the following simple criterion.

Lemma 2.10. *A point $x_0 \in \mathbb{R}$ is the α-quantile of some df F if and only if the following two conditions are satisfied: $F(x_0) \geqslant \alpha$; and $F(x) < \alpha$ for all $x < x_0$.*

The lemma follows immediately from the definition of the generalized inverse and the right-continuity of F. Examples of the computation of quantiles in certain tricky cases and further properties of generalized inverses are given in Section A.1.2.

Example 2.11 (VaR for normal and *t* loss distributions). Suppose that the loss distribution F_L is normal with mean μ and variance σ^2. Fix $\alpha \in (0, 1)$. Then

$$\mathrm{VaR}_\alpha = \mu + \sigma\Phi^{-1}(\alpha), \tag{2.18}$$

where Φ denotes the standard normal df and $\Phi^{-1}(\alpha)$ is the α-quantile of Φ. The proof is easy: since F_L is strictly increasing, by Lemma 2.10 we only have to show that $F_L(\mathrm{VaR}_\alpha) = \alpha$. Now,

$$P(L \leqslant \mathrm{VaR}_\alpha) = P\left(\frac{L-\mu}{\sigma} \leqslant \Phi^{-1}(\alpha)\right) = \Phi(\Phi^{-1}(\alpha)) = \alpha.$$

This result is routinely used in the *variance–covariance* approach (also known as the delta-normal approach) to computing risk measures.

Of course, a similar result is obtained for any location-scale family, and another useful example is the Student t loss distribution. Suppose that our loss L is such that $(L - \mu)/\sigma$ has a standard t distribution with ν degrees of freedom; we denote this loss distribution by $L \sim t(\nu, \mu, \sigma^2)$ and note that the moments are given by $E(L) = \mu$ and $\mathrm{var}(L) = \nu\sigma^2/(\nu-2)$ when $\nu > 2$, so σ is not the standard deviation of the distribution. We get

$$\mathrm{VaR}_\alpha = \mu + \sigma t_\nu^{-1}(\alpha), \tag{2.19}$$

where t_ν denotes the df of a standard t distribution, which is available in most statistical computer packages along with its inverse.

In the remainder of this section we discuss a number of further issues relating to the use of VaR as a risk measure in practice.

Choice of VaR parameters. In working with VaR the parameters Δt and α need to be chosen. There is of course no single optimal value for these parameters, but there are some considerations that might influence the choice of regulators or internal-model builders.

The risk-management horizon Δt should reflect the time period over which a financial institution is committed to hold its portfolio, which will be affected by contractual and legal constraints as well as liquidity considerations. In choosing a horizon for enterprise-wide risk management, a financial institution has little choice but to use the horizon appropriate for the market in which its core business activities lie. For example, insurance companies are typically bound to hold their portfolio of liabilities for one year, during which time they are not able to alter the portfolio or renegotiate the premiums they receive; one year is therefore an appropriate time horizon for measuring the risk in the liability and asset portfolios of an insurer. Moreover, a financial institution can be forced to hold a loss-making position in a risky asset if the market for that asset is not very liquid, so a relatively long horizon may be appropriate for illiquid assets.

There are other, more practical, considerations that suggest that Δt should be relatively small. The assumption that the composition of the portfolio remains unchanged is tenable only for a short holding period. Moreover, the calibration and testing of statistical models for historical risk-factor changes $(\boldsymbol{X}_t)$ are easier if Δt is small, since this typically means that we have more data at our disposal.

For the confidence level α, different values are also appropriate for different purposes. In order to set limits for traders, a bank would typically take α to be 95% and Δt to be one day. For capital adequacy purposes higher confidence levels are generally used. For instance, the Basel capital charges for market risk in the trading book of a bank are based on the use of VaR at the 99% level and a ten-day horizon. The Solvency II framework uses a value of α equal to 0.995 and a one-year horizon. On the other hand, the backtesting of models that produce VaR figures often needs to be carried out at lower confidence levels using shorter horizons in order to have sufficient statistical power to detect poor model performance.

Model risk and market liquidity. In practice, VaR numbers are sometimes given a very literal interpretation; the statement that the daily VaR at confidence level $\alpha = 99\%$ for a particular portfolio is equal to l is understood to mean that "there is a probability of exactly 1% that the loss on this position will be larger than l". This interpretation is misleading because it neglects estimation error, model risk and market liquidity risk.

We recall that model risk is the risk that our model for the loss distribution is misspecified. For instance, we might work with a normal distribution to model losses, whereas the true distribution is heavy tailed, or we might fail to recognize the presence of volatility clustering or tail dependence (see Chapter 3) in modelling the distribution of the risk-factor changes underlying the losses. Of course, these problems are most pronounced if we are trying to estimate VaR at very high confidence levels. Liquidity risk refers to the fact that any attempt to liquidate a large loss-making position is likely to move the price against us, thus exacerbating the loss.

2.3.3 VaR in Risk Capital Calculations

Quantile-based risk measures are used in many risk capital calculations in practice. In this section we give two examples.

VaR in regulatory capital calculations for the trading book. The VaR risk measure is applied to calculate a number of regulatory capital charges for the trading book of a bank. Under the internal-model approach a bank calculates a daily VaR measure for the distribution of possible ten-day trading book losses based on recent data on risk-factor changes under the assumption that the trading book portfolio is held fixed over this time period. We describe the statistical methodology that is typically used for this calculation in Section 9.2.

While exact details may vary from one national regulator to another, the basic capital charge on day t is usually calculated according to a formula of the form

$$\mathrm{RC}^t = \max\left\{ \mathrm{VaR}^{t,10}_{0.99}, \frac{k}{60}\sum_{i=1}^{60} \mathrm{VaR}^{t-i+1,10}_{0.99} \right\}, \tag{2.20}$$

where $\mathrm{VaR}^{j,10}_{0.99}$ stands for the ten-day VaR at the 99% confidence level, calculated on day j, and where k is a multiplier in the range 3–4 that is determined by the regulator as a function of the overall quality of the bank's internal model. The averaging of the last sixty daily VaR numbers obviously tends to lead to smooth changes in the capital charge over time unless the most recent number $\mathrm{VaR}^{t,10}_{0.99}$ is particularly large.

A number of additional capital charges are added to RC^t. These include a stressed VaR charge and an incremental risk charge, as well as a number of charges that are designed to take into account so-called specific risks due to idiosyncratic price movements in certain instruments that are not explained by general market-risk factors. The stressed VaR charge is calculated using similar VaR methodology to the standard charge but with data taken from a historical window in which markets were particularly volatile. The incremental risk charge is an estimate of the 99.9% quantile of the distribution of annual losses due to defaults and downgrades for credit-risky instruments in the trading book (excluding securitizations).

The solvency capital requirement in Solvency II. An informal definition of the solvency capital requirement is "the level of capital that enables the insurer to meet its obligations over a one-year time horizon with a high confidence level (99.5%)" (this is taken from a 2007 factsheet produced by De Nederlandsche Bank). We will give an argument that leads to the use of a VaR-based risk measure.

Consider the balance sheet of the insurer in Table 2.2 and assume that the current equity capital is given by $V_t = A_t - B_t$, i.e. the difference between the value of assets and the value of liabilities, or the net asset value; this is also referred to under Solvency II as *own funds*. The liabilities B_t are considered to include all technical provisions computed in a market-consistent way, including risk margins for non-hedgeable risks where necessary.

The insurer wants to ensure that it is solvent in one year's time with high probability α. It considers the possibility that it may need to raise extra capital and makes the following thought calculation. Given its current balance sheet and business model it attempts to determine the minimum amount of extra capital x_0 that it would have to raise now at time t and place in a risk-free investment in order to be solvent in one year's time with probability α. In mathematical notation it needs to determine

$$x_0 = \inf\{x \colon P(V_{t+1} + x(1 + r_{t,1}) \geqslant 0) = \alpha\},$$

where $r_{t,1}$ is the simple risk-free rate for a one-year investment and V_{t+1} is the net asset value in one year's time. If x_0 is negative, then the company is well capitalized at time t and money could be taken out of the company.

An easy calculation gives

$$\begin{aligned} x_0 &= \inf\{x \colon P(-V_{t+1} \leqslant x(1 + r_{t,1})) = \alpha\} \\ &= \inf\{x \colon P(V_t - V_{t+1}/(1 + r_{t,1}) \leqslant x + V_t) = \alpha\}, \end{aligned}$$

which shows that

$$V_t + x_0 = q_\alpha(V_t - V_{t+1}/(1 + r_{t,1})).$$

The sum $V_t + x_0$ gives the solvency capital requirement: namely, the available capital corrected by the amount x_0. Hence, we see that the solvency capital requirement is a quantile of the distribution of $V_t - V_{t+1}/(1 + r_{t,1})$, a loss distribution that takes into account the time value of money through discounting, as discussed in Section 2.2.1. For a well-capitalized company with $x_0 < 0$, the amount $-x_0 = V_t - q_\alpha(V_t - V_{t+1}/(1 + r_{t,1}))$ (own funds minus the solvency capital requirement) is called the excess capital.

2.3.4 Other Risk Measures Based on Loss Distributions

In this section we provide short notes on a number of other statistical summaries of the loss distribution that are frequently used as risk measures in financial and insurance risk management.

Variance. Historically, the variance of the P&L distribution (or, equivalently, the standard deviation) has been the dominating risk measure in finance. To a large extent this is due to the huge impact that the portfolio theory of Markowitz, which uses variance as a measure of risk, has had on theory and practice in finance (see, for example, Markowitz 1952). Variance is a well-understood concept that is easy to use analytically. However, as a risk measure it has two drawbacks. On the technical side, if we want to work with variance, we have to assume that the second moment of the loss distribution exists. While unproblematic for most return distributions in finance, this can cause problems in certain areas of non-life insurance or for the analysis of operational losses (see Section 13.1.4). On the conceptual side, since it makes no distinction between positive and negative deviations from the mean, variance is a good measure of risk only for distributions that are (approximately) symmetric around the mean, such as the normal distribution or a (finite-variance) Student t distribution. However, in many areas of risk management, such as in credit and operational risk management, we deal with loss distributions that are highly skewed.

Lower and upper partial moments. Partial moments are measures of risk based on the lower or upper part of a distribution. In most of the literature on risk management the main concern is with the risk inherent in the lower tail of a P&L distribution, and lower partial moments are used to measure this risk. Under our sign convention we are concerned with the risk inherent in the upper tail of a loss distribution, so we focus on upper partial moments. Given an exponent $k \geqslant 0$ and a reference point q, the upper partial moment $\mathrm{UPM}(k, q)$ is defined as

$$\mathrm{UPM}(k, q) = \int_q^\infty (l - q)^k \, \mathrm{d}F_L(l) \in [0, \infty]. \tag{2.21}$$

Some combinations of k and q have a special interpretation: for $k = 0$ we obtain $P(L \geqslant q)$; for $k = 1$ we obtain $E((L - q)I_{\{L \geqslant q\}})$; for $k = 2$ and $q = E(L)$ we obtain the *upper semivariance* of L. Of course, the higher the value we choose for k, the more conservative our risk measure becomes, since we give more and more weight to large deviations from the reference point q.

Expected shortfall. ES is closely related to VaR and there is an ongoing debate in the risk-management community on the strengths and weaknesses of both risk measures.

Definition 2.12 (expected shortfall). For a loss L with $E(|L|) < \infty$ and df F_L, the ES at confidence level $\alpha \in (0, 1)$ is defined as

$$\mathrm{ES}_\alpha = \frac{1}{1 - \alpha} \int_\alpha^1 q_u(F_L) \, \mathrm{d}u, \tag{2.22}$$

where $q_u(F_L) = F_L^{\leftarrow}(u)$ is the quantile function of F_L.

The condition $E(|L|) < \infty$ ensures that the integral in (2.22) is well defined. By definition, ES is related to VaR by

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_u(L)\,\mathrm{d}u.$$

Instead of fixing a particular confidence level α, we average VaR over all levels $u \geqslant \alpha$ and thus "look further into the tail" of the loss distribution. Obviously, ES_α depends only on the distribution of L, and $\mathrm{ES}_\alpha \geqslant \mathrm{VaR}_\alpha$. See Figure 2.2 for a simple illustration of an ES value and its relationship to VaR. The 95% ES value of 4.9 is at least double the 95% VaR value of 2.2 in this case.

For continuous loss distributions an even more intuitive expression can be derived that shows that ES can be interpreted as the expected loss that is incurred in the event that VaR is exceeded.

Lemma 2.13. *For an integrable loss L with continuous df F_L and for any $\alpha \in (0,1)$ we have*

$$\mathrm{ES}_\alpha = \frac{E(L;\, L \geqslant q_\alpha(L))}{1-\alpha} = E(L \mid L \geqslant \mathrm{VaR}_\alpha), \tag{2.23}$$

where we have used the notation $E(X;A) := E(XI_A)$ for a generic integrable rv X and a generic set $A \in \mathcal{F}$.

Proof. Denote by U an rv with uniform distribution on the interval $[0,1]$. It is a well-known fact from elementary probability theory that the rv $F_L^{\leftarrow}(U)$ has df F_L (see Proposition 7.2 for a proof). We have to show that $E(L;\, L \geqslant q_\alpha(L)) = \int_\alpha^1 F_L^{\leftarrow}(u)\,\mathrm{d}u$. Now,

$$E(L;\, L \geqslant q_\alpha(L)) = E(F_L^{\leftarrow}(U);\, F_L^{\leftarrow}(U) \geqslant F_L^{\leftarrow}(\alpha)) = E(F_L^{\leftarrow}(U);\, U \geqslant \alpha);$$

in the last equality we used the fact that $F_L^{\leftarrow}$ is strictly increasing since F_L is continuous (see Proposition A.3 (iii)). Thus we get $E(F_L^{\leftarrow}(U);\, U \geqslant \alpha) = \int_\alpha^1 F_L^{\leftarrow}(u)\,\mathrm{d}u$. The second representation follows since, for a continuous loss distribution F_L, we have $P(L \geqslant q_\alpha(L)) = 1-\alpha$. □

For an extension of this result to loss distributions with atoms, we refer to Proposition 8.13. Next we use Lemma 2.13 to calculate the ES for two common continuous distributions.

Example 2.14 (expected shortfall for Gaussian loss distribution). Suppose that the loss distribution F_L is normal with mean μ and variance σ^2. Fix $\alpha \in (0,1)$. Then

$$\mathrm{ES}_\alpha = \mu + \sigma\frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}, \tag{2.24}$$

where ϕ is the density of the standard normal distribution. The proof is elementary. First note that

$$\mathrm{ES}_\alpha = \mu + \sigma E\biggl(\frac{L-\mu}{\sigma} \biggm| \frac{L-\mu}{\sigma} \geqslant q_\alpha\biggl(\frac{L-\mu}{\sigma}\biggr)\biggr);$$

Table 2.3. VaR_α and ES_α in the normal and t models for different values of α.

α	0.90	0.95	0.975	0.99	0.995
VaR_α (normal model)	162.1	208.1	247.9	294.3	325.8
VaR_α (t model)	137.1	190.7	248.3	335.1	411.8
ES_α (normal model)	222.0	260.9	295.7	337.1	365.8
ES_α (t model)	223.5	286.5	357.2	466.9	565.7

hence, it suffices to compute the ES for the standard normal rv $\tilde{L} := (L - \mu)/\sigma$. Here we get

$$\mathrm{ES}_\alpha(\tilde{L}) = \frac{1}{1-\alpha}\int_{\Phi^{-1}(\alpha)}^{\infty} l\phi(l)\,\mathrm{d}l = \frac{1}{1-\alpha}[-\phi(l)]_{\Phi^{-1}(\alpha)}^{\infty} = \frac{\phi(\Phi^{-1}(\alpha))}{1-\alpha}.$$

Example 2.15 (expected shortfall for the Student t loss distribution). Suppose the loss L is such that $\tilde{L} = (L-\mu)/\sigma$ has a standard t distribution with ν degrees of freedom, as in Example 2.11. Suppose further that $\nu > 1$. By the reasoning of Example 2.14, which applies to any location-scale family, we have $\mathrm{ES}_\alpha = \mu + \sigma\,\mathrm{ES}_\alpha(\tilde{L})$. The ES of the standard t distribution is easily calculated by direct integration to be

$$\mathrm{ES}_\alpha(\tilde{L}) = \frac{g_\nu(t_\nu^{-1}(\alpha))}{1-\alpha}\left(\frac{\nu + (t_\nu^{-1}(\alpha))^2}{\nu - 1}\right), \tag{2.25}$$

where t_ν denotes the df and g_ν the density of standard t.

Since ES_α can be thought of as an average over all losses that are greater than or equal to VaR_α, it is sensitive to the severity of losses exceeding VaR_α. This advantage of ES is illustrated in the following example.

Example 2.16 (VaR and ES for stock returns). We consider daily losses on a position in a particular stock; the current value of the position equals $V_t = 10\,000$. Recall from Example 2.1 that the loss for this portfolio is given by $L_{t+1}^{\Delta} = -V_t X_{t+1}$, where X_{t+1} represents daily log-returns of the stock. We assume that X_{t+1} has mean 0 and standard deviation $\sigma = 0.2/\sqrt{250}$, i.e. we assume that the stock has an annualized volatility of 20%. We compare two different models for the distribution: namely, (i) a normal distribution, and (ii) a t distribution with $\nu = 4$ degrees of freedom scaled to have standard deviation σ. The t distribution is a symmetric distribution with heavy tails, so that large absolute values are much more probable than in the normal model; it is also a distribution that has been shown to fit well in many empirical studies (see Example 6.14). In Table 2.3 we present VaR_α and ES_α for both models and various values of α. In case (i) these values have been computed using (2.24); the ES for the t model has been computed using (2.25).

Most risk managers would argue that the t model is riskier than the normal model, since under the t distribution large losses are more likely. However, if we use VaR at the 95% or 97.5% confidence level to measure risk, the normal distribution appears to be at least as risky as the t model; only above a confidence level of 99% does the higher risk in the tails of the t model become apparent. On the other hand, if we use ES, the risk in the tails of the t model is reflected in our risk measurement for lower values of α. Of course, simply going to a 99% confidence level in quoting VaR numbers does not help to overcome this deficiency of VaR, as there are other examples where the higher risk becomes apparent only for confidence levels beyond 99%.

Remark 2.17. It is possible to derive results on the asymptotics of the *shortfall-to-quantile ratio* $\mathrm{ES}_\alpha/\mathrm{VaR}_\alpha$ for $\alpha \to 1$. For the normal distribution we have $\lim_{\alpha\to 1}\mathrm{ES}_\alpha/\mathrm{VaR}_\alpha = 1$; for the t distribution with $\nu > 1$ degrees of freedom we have $\lim_{\alpha\to 1}\mathrm{ES}_\alpha/\mathrm{VaR}_\alpha = \nu/(\nu-1) > 1$. This shows that for a heavy-tailed distribution, the difference between ES and VaR is more pronounced than for the normal distribution. We will take up this issue in more detail in Section 5.2.3 (see also Section 8.4.4).

2.3.5 Coherent and Convex Risk Measures

The premise of this section is the idea of approaching risk measurement by first writing down a list of properties (axioms) that a good risk measure should have. For applications in risk management, such axioms have been proposed by Artzner et al. (1999) (coherent risk measures) and Föllmer and Schied (2002) (convex risk measures). In this section we discuss these axioms in relation to specific examples of risk measures. A longer and more theoretical treatment of coherent and convex risk measures will be given in Chapter 8. It should be mentioned that the idea of having axiomatic systems for risk measures bears some relationship to similar systems for premium principles in the actuarial literature, which have a long and independent history (see, for example, Goovaerts et al. (2003), as well as further references in the Notes and Comments section below).

Axioms for risk measures For the purposes of this section risk measures are real-valued functions defined on a linear space of random variables $\mathcal{M}$, assumed to include constants. There are two possible interpretations of the elements of $\mathcal{M}$. First, elements of $\mathcal{M}$ could be considered as future net asset values of portfolios or positions; in that case, elements of $\mathcal{M}$ will be denoted by V and the current net asset value will be denoted by V_0. Second, elements of $\mathcal{M}$ could represent losses L, where, of course, these are related to future values by the formula $L = -(V - V_0)$ (ignoring any discounting for simplicity).

Correspondingly, there are two possible notions of risk measures on $\mathcal{M}$. On the one hand, we can view the risk measure as the amount of *additional capital* that needs to be added to a position with future net asset value V to make the position acceptable to a regulator or a prudent manager; in this case we write the risk measure as $\tilde{\varrho}(V)$. On the other hand, we might interpret the risk measure as the *total amount of equity capital* that is necessary to back a position with loss L; in this case we write the risk measure as $\varrho(L)$.

These two notions are related by

$$\varrho(L) = V_0 + \tilde{\varrho}(V),$$

since "total capital" is equal to "available capital" plus "additional capital". However, these two different notions do have a bearing on the way in which the axioms are presented and understood. Since in this book we mostly focus on loss distributions, we present the axioms for losses and consider a risk measure $\varrho\colon L \mapsto \varrho(L)$. Note that the alternative notion is frequently found in the literature.

Axiom 2.18 (monotonicity). For $L_1, L_2 \in \mathcal{M}$ such that $L_1 \leqslant L_2$ almost surely, we have $\varrho(L_1) \leqslant \varrho(L_2)$.

From an economic viewpoint this axiom is obvious: positions that lead to higher losses in every state of the world require more risk capital. Positions with $\varrho(L) \leqslant 0$ do not require any capital.

Axiom 2.19 (translation invariance). For all $L \in \mathcal{M}$ and every $l \in \mathbb{R}$ we have $\varrho(L + l) = \varrho(L) + l$.

Axiom 2.19 states that by adding or subtracting a deterministic quantity l to a position leading to the loss L, we alter our capital requirements by exactly that amount. In terms of the alternative notion of a risk measure defined on future net asset values, this axiom implies that $\tilde{\varrho}(V + k) = \tilde{\varrho}(V) - k$ for $k \in \mathbb{R}$. It follows that $\tilde{\varrho}(V + \tilde{\varrho}(V)) = 0$, so a position with future net asset value $V + \tilde{\varrho}(V)$ is immediately acceptable without further injection of capital. This makes sense and implies that risk is measured in monetary terms.

Axiom 2.20 (subadditivity). For all $L_1, L_2 \in \mathcal{M}$ we have $\varrho(L_1 + L_2) \leqslant \varrho(L_1) + \varrho(L_2)$.

The rationale behind Axiom 2.20 is summarized by Artzner et al. (1999) in the statement that "a merger does not create extra risk" (ignoring, of course, any problematic practical aspects of a merger!). Axiom 2.20 is the most debated of the four axioms characterizing coherent risk measures, probably because it rules out VaR as a risk measure in certain situations. We provide some arguments explaining why subadditivity is indeed a reasonable requirement. First, subadditivity reflects the idea that risk can be reduced by diversification, a time-honoured principle in finance and economics. Second, if a regulator uses a non-subadditive risk measure in determining the regulatory capital for a financial institution, that institution has an incentive to legally break up into various subsidiaries in order to reduce its regulatory capital requirements. Similarly, if the risk measure used by an organized exchange in determining the margin requirements of investors is non-subadditive, an investor could reduce the margin he has to pay by opening a different account for every position in his portfolio. Finally, subadditivity makes decentralization of risk-management systems possible. Consider as an example two trading desks with positions leading to losses L_1 and L_2. Imagine that a risk manager wants to ensure that $\varrho(L)$, the risk of the overall loss $L = L_1 + L_2$, is smaller than some number M. If he uses a subadditive risk measure ϱ, he may simply choose bounds M_1 and M_2 such that $M_1 + M_2 \leqslant M$ and impose on each of the desks the constraint that $\varrho(L_i) \leqslant M_i$; subadditivity of ϱ then automatically ensures that $\varrho(L) \leqslant M_1 + M_2 \leqslant M$.

Axiom 2.21 (positive homogeneity). For all $L \in \mathcal{M}$ and every $\lambda > 0$ we have $\varrho(\lambda L) = \lambda\varrho(L)$.

Axiom 2.21 is easily justified if we assume that Axiom 2.20 holds. Subadditivity implies that, for $n \in \mathbb{N}$,

$$\varrho(nL) = \varrho(L + \cdots + L) \leqslant n\varrho(L). \tag{2.26}$$

Since there is no netting or diversification between the losses in this portfolio, it is natural to require that equality should hold in (2.26), which leads to positive homogeneity. Note that subadditivity and positive homogeneity imply that the risk measure ϱ is *convex* on $\mathcal{M}$.

Definition 2.22 (coherent risk measure). A risk measure ϱ whose domain includes the convex cone $\mathcal{M}$ is called *coherent* (on $\mathcal{M}$) if it satisfies Axioms 2.18–2.21.

Axiom 2.21 (positive homogeneity) has been criticized and, in particular, it has been suggested that for large values of the multiplier λ we should have $\varrho(\lambda L) > \lambda\varrho(L)$ in order to penalize a concentration of risk and to account for liquidity risk in a large position. As shown in (2.26), this is impossible for a subadditive risk measure. This problem has led to the study of the larger class of *convex risk measures*. In this class the conditions of subadditivity and positive homogeneity have been relaxed; instead one requires only the weaker property of convexity.

Axiom 2.23 (convexity). For all $L_1, L_2 \in \mathcal{M}$ and all $\lambda \in [0, 1]$ we have $\varrho(\lambda L_1 + (1 - \lambda)L_2) \leqslant \lambda\varrho(L_1) + (1 - \lambda)\varrho(L_2)$.

The economic justification for convexity is again the idea that diversification reduces risk.

Definition 2.24 (convex risk measure). A risk measure ϱ on $\mathcal{M}$ is called *convex* (on $\mathcal{M}$) if it satisfies Axioms 2.18, 2.19 and 2.23.

While every coherent risk measure is convex, the converse is not true. In particular, within the class of convex risk measures it is possible to find risk measures that penalize concentration of risk in the sense that $\varrho(\lambda L) \geqslant \varrho(L)$ for $\lambda > 1$ (see, for example, Example 8.8). On the other hand, for risk measures that are positive, homogeneous convexity and subadditivity are equivalent.

Examples. In view of its practical relevance we begin with a discussion of VaR. It is immediate from the definition of VaR as a quantile of the loss distribution that VaR is translation invariant, monotone and positive homogeneous. However, the following example shows that VaR is in general not subadditive, and hence, in general, neither is it a convex nor a coherent measure of risk.

Example 2.25 (non-subadditivity of VaR for defaultable bonds). Consider a portfolio of two zero-coupon bonds with a maturity of one year that default independently. The default probability of both bonds is assumed to be identical and equal to $p = 0.9\%$. The current price of the bonds and the face value of the bonds is equal to 100, and the bonds pay an interest rate of 5%. If there is no default, the holder of the bond therefore receives a payment of size 105 in one year; in the case of a default, he receives nothing, i.e. we assume a recovery rate of zero. Denote by L_i the loss incurred by holding one unit of bond i. We have

$$\begin{aligned} P(L_i = -5) &= 1 - p = 0.991 \quad \text{(no default)}, \\ P(L_i = 100) &= p \qquad\;\; = 0.009 \quad \text{(default)}. \end{aligned}$$

Set $\alpha = 0.99$. We have $P(L_i < -5) = 0$ and $P(L_i \leqslant -5) = 0.991 > \alpha$, so $\mathrm{VaR}_\alpha(L_i) = -5$.

Now consider a portfolio of one bond from each firm, with corresponding loss $L = L_1 + L_2$. Since the default events of the two firms are independent, we get

$$\begin{aligned} P(L = -10) &= (1-p)^2 &&= 0.982 &&\text{(no default)}, \\ P(L = 95) &= 2p(1-p) &&= 0.07838 &&\text{(one default)}, \\ P(L = 200) &= p^2 &&= 0.000081 &&\text{(two defaults)}. \end{aligned}$$

In particular, $P(L \leqslant -10) = 0.982 < 0.99$ and $P(L \leqslant 95) > 0.99$, so $\mathrm{VaR}_\alpha(L) = 95 > -10 = \mathrm{VaR}_\alpha(L_1) + \mathrm{VaR}_\alpha(L_2)$. Hence, VaR is non-subadditive. In fact, in the example, VaR_α punishes diversification, as

$$\mathrm{VaR}_\alpha(0.5L_1 + 0.5L_2) = 0.5\,\mathrm{VaR}_\alpha(L) = 47.5 > \mathrm{VaR}_\alpha(L_1).$$

In Example 2.25 the non-subadditivity of VaR is caused by the fact that the assets making up the portfolio have very skewed loss distributions; such a situation can clearly occur if we have defaultable bonds or options in our portfolio. Note, however, that the assets in this example have an innocuous dependence structure because they are independent.

In fact, the non-subadditivity of VaR can be seen in many different examples. The following is a list of the situations that we will encounter in this book.

- Independent losses with highly skewed discrete distributions, as in Example 2.25.
- Independent losses with continuous light-tailed distributions but low values of α. This will be demonstrated for exponentially distributed losses in Example 7.30 and discussed further in Section 8.3.3.
- Dependent losses with continuous symmetric distributions when the *dependence structure* takes a special form. This will be demonstrated for normally distributed losses in Example 8.39.
- Independent losses with continuous but very heavy-tailed distributions. This can be seen in Example 8.40 for the extreme case of infinite-mean Pareto risks. While less relevant for modelling market and credit risks, infinite-mean distributions are sometimes used to model certain kinds of insurance losses as well as losses due to operational risk (see Chapter 13 for more discussion).

Note that the domain $\mathcal{M}$ is an integral part of the definition of a convex or coherent risk measure. We will often encounter risk measures that are coherent or convex if restricted to a sufficiently small domain. For example, VaR is subadditive in the idealized situation where all portfolios can be represented as linear combinations of the same set of underlying multivariate normal or, more generally, elliptically distributed risk factors (see Proposition 8.28).

There is an ongoing debate about the practical relevance of the non-subadditivity of VaR. The non-subadditivity can be particularly problematic if VaR is used to set risk limits for traders, as this can lead to portfolios with a high degree of concentration risk. Consider, for instance, in the set-up of Example 2.25 a trader who wants to maximize the expected return of a portfolio in the two defaultable bonds under the constraint that the VaR of his position is smaller than some given positive number; no short selling is permitted. Clearly, an optimal strategy for this trader is to invest all funds in one of the two bonds, a very concentrated position. For an elaboration of this toy example we refer to Frey and McNeil (2002).

Example 2.26 (coherence of expected shortfall). ES, on the other hand, is a coherent risk measure. Translation invariance, monotonicity and positive homogeneity are immediate from the corresponding properties of the quantile. For instance, it holds that

$$\mathrm{ES}_\alpha(\lambda L) = \frac{1}{1-\alpha}\int_\alpha^1 q_u(\lambda L)\,\mathrm{d}u = \frac{1}{1-\alpha}\int_\alpha^1 \lambda q_u(L)\,\mathrm{d}u = \lambda\,\mathrm{ES}_\alpha(L),$$

and similar arguments apply for translation invariance and monotonicity. A general proof of subadditivity is given in Theorem 8.14. Here, we give a simple argument for the case where L_1, L_2 and $L_1 + L_2$ have a continuous distribution. We recall from Lemma 2.13 that for a random variable L with a continuous distribution, it holds that

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha}E(LI_{\{L\geqslant q_\alpha(L)\}}).$$

To simplify the notation let $I_1 := I_{\{L_1\geqslant q_\alpha(L_1)\}}$, $I_2 := I_{\{L_2\geqslant q_\alpha(L_2)\}}$ and $I_{12} := I_{\{L_1+L_2\geqslant q_\alpha(L_1+L_2)\}}$. We calculate that

$$\begin{aligned}(1-\alpha)(\mathrm{ES}_\alpha(L_1) + \mathrm{ES}_\alpha(L_1) - \mathrm{ES}_\alpha(L_1+L_2))&\\ = E(L_1I_1) + E(L_2I_2) - E((L_1+L_2)I_{12})&\\ = E(L_1(I_1 - I_{12})) + E(L_2(I_2 - I_{12})).&\end{aligned}$$

Consider the first term and suppose that $\{L_1 \geqslant q_\alpha(L_1)\}$. It follows that $I_1 - I_{12} \geqslant 0$ and hence that $L_1(I_1 - I_{12}) \geqslant q_\alpha(L_1)(I_1 - I_{12})$. Suppose, on the other hand, that $\{L_1 < q_\alpha(L_1)\}$. It follows that $I_1 - I_{12} \leqslant 0$ and hence that $L_1(I_1 - I_{12}) \geqslant q_\alpha(L_1)(I_1 - I_{12})$. The same reasoning applies to L_2, so in either case we conclude that

$$\begin{aligned}(1-\alpha)(\mathrm{ES}_\alpha(L_1) &+ \mathrm{ES}_\alpha(L_1) - \mathrm{ES}_\alpha(L_1+L_2))\\ &\geqslant E(q_\alpha(L_1)(I_1 - I_{12})) + E(q_\alpha(L_2)(I_2 - I_{12}))\\ &\geqslant q_\alpha(L_1)E(I_1 - I_{12}) + q_\alpha(L_2)E(I_2 - I_{12})\\ &\geqslant q_\alpha(L_1)((1-\alpha) - (1-\alpha)) + q_\alpha(L_2)((1-\alpha) - (1-\alpha))\\ &= 0,\end{aligned}$$

which proves subadditivity.

We have now seen two advantages of ES over VaR: ES reflects tail risk better (see Example 2.16) and it is always subadditive. On the other hand, VaR has practical advantages: it is easier to estimate, in particular for heavy-tailed distributions, and VaR estimates are easier to backtest than estimates of ES. We will come back to this point in our discussion of backtesting in Section 9.3.

Example 2.27 (generalized scenario risk measure). The generalized scenario risk measure in (2.16) is another example of a coherent risk measure. Translation invariance, positive homogeneity and monotonicity are clear, so it only remains to check subadditivity. For $i = 1, 2$ denote by $L_i(\boldsymbol{x})$ the hypothetical loss of position i under the scenario $\boldsymbol{x}$ for the risk-factor changes. We observe that

$$\begin{aligned}&\max\{E^P(L_1(\boldsymbol{X})) + L_2(\boldsymbol{X}))\colon P \in \mathcal{P}_{[\mathcal{X},\boldsymbol{w}]}\}\\ &\qquad \leqslant \max\{E^P(L_1(\boldsymbol{X}))\colon P \in \mathcal{P}_{[\mathcal{X},\boldsymbol{w}]}\} + \max\{E^P(L_2(\boldsymbol{X}))\colon P \in \mathcal{P}_{[\mathcal{X},\boldsymbol{w}]}\}.\end{aligned}$$

Example 2.28 (a coherent premium principle). In Fischer (2003), a class of coherent risk measures closely resembling certain actuarial premium principles is proposed. These risk measures are potentially useful for an insurance company that wants to compute premiums on a coherent basis without deviating too far from standard actuarial practice.

Given constants $p > 1$ and $\alpha \in [0, 1)$, this coherent premium principle $\varrho_{[\alpha,p]}$ is defined as follows. Let $\mathcal{M} := L^p(\Omega, \mathcal{F}, P)$, the space of all L with $\|L\|_p := E(|L|^p)^{1/p} < \infty$, and define, for $L \in \mathcal{M}$,

$$\varrho_{[\alpha,p]}(L) = E(L) + \alpha\|(L - E(L))^+\|_p. \tag{2.27}$$

Under (2.27) the risk associated with a loss L is measured by the sum of $E(L)$, the pure actuarial premium for the loss, and a *risk loading* given by a fraction α of the L^p-norm of the positive part of the centred loss $L - E(L)$. This loading can be written more explicitly as $(\int_{E(L)}^{\infty}(l - E(L))^p \, \mathrm{d}F_L(l))^{1/p}$. The higher the values of α and p, the more conservative the risk measure $\varrho_{[\alpha,p]}$ becomes.

The coherence of $\varrho_{[\alpha,p]}$ is easy to check. Translation invariance and positive homogeneity are immediate. To prove subadditivity observe that for any two rvs X and Y we have $(X + Y)^+ \leqslant X^+ + Y^+$. Hence, from Minkowski's inequality (the triangle inequality for the L^p-norm) we obtain that for any two $L_1, L_2 \in \mathcal{M}$,

$$\begin{aligned}\|(L_1 - E(L_1) + L_2 - E(L_2))^+\|_p &\leqslant \|(L_1 - E(L_1))^+ + (L_2 - E(L_2))^+\|_p\\ &\leqslant \|(L_1 - E(L_1))^+\|_p + \|(L_2 - E(L_2))^+\|_p,\end{aligned}$$

which shows that $\varrho_{[\alpha,p]}$ is subadditive. To verify monotonicity, assume that $L_1 \leqslant L_2$ almost surely and write $L = L_1 - L_2$. Since $L \leqslant 0$ almost surely, it follows that $(L - E(L))^+ \leqslant -E(L)$ almost surely, and hence that $\|(L - E(L))^+\|_p \leqslant -E(L)$ and $\varrho_{[\alpha,p]}(L) \leqslant 0$, since $\alpha < 1$. Using the fact that $L_1 = L_2 + L$ and the subadditivity property we obtain

$$\varrho_{[\alpha,p]}(L_1) \leqslant \varrho_{[\alpha,p]}(L_2) + \varrho_{[\alpha,p]}(L) \leqslant \varrho_{[\alpha,p]}(L_2).$$

Notes and Comments

An extensive discussion of different approaches to risk quantification is given in Crouhy, Galai and Mark (2001). Value-at-risk was introduced by JPMorgan in the first version of its RiskMetrics system and was quickly accepted by risk managers and regulators as an industry standard; see also Brown (2012) for a broader view of the history and use of VaR on Wall Street. A number of different notions of VaR are used in practice (see Alexander 2009, Volume 4) but all are related to the idea of a quantile of the P&L distribution.

Expected shortfall was made popular by Artzner et al. (1997, 1999). There are a number of variants of the ES risk measure with a variety of names, such as tail conditional expectation, worst conditional expectation and conditional VaR; all coincide for continuous loss distributions. Acerbi and Tasche (2002) discuss the relationships between the various notions. Risk measures based on loss distributions also appear in the literature under the (somewhat unfortunate) heading of *law-invariant* risk measures.

Example 2.25 is due to Albanese (1997) and Artzner et al. (1999). There are many different examples of the non-subadditivity of VaR in the literature, including the case of independent, infinite-mean Pareto risks (see Embrechts, McNeil and Straumann 2002, Example 7; Denuit and Charpentier 2004, Example 5.2.7). The implications of the non-subadditivity of VaR for portfolio optimization are discussed in Frey and McNeil (2002); see also papers by Basak and Shapiro (2001), Krokhmal, Palmquist and Uryasev (2001) and Emmer, Klüppelberg and Korn (2001).

A class of risk measures that are widely used throughout the hedge fund industry is based on the peak-to-bottom loss over a given period of time in the performance curve of an investment. These measures are typically referred to as (maximal) *drawdown* risk measures (see, for example, Chekhlov, Uryasev and Zabarankin 2005; Jaeger 2005).

The measurement of financial risk and the computation of actuarial premiums are at least conceptually closely related problems, so that the actuarial literature on premium principles is of relevance in financial risk management. We refer to Chapter 3 of Rolski et al. (1999) for an overview; Goovaerts, De Vylder and Haezendonck (1984) provides a specialist account.

Model risk has become a central issue in modern risk management. The problems faced by the hedge fund LTCM in 1998 provide a prime example of model risk in VaR-based risk-management systems. While LTCM had a seemingly sophisticated VaR system in place, errors in parameter estimation, unexpectedly large market moves (heavy tails) and, in particular, vanishing market liquidity drove the hedge fund into near-bankruptcy, causing major financial turbulence around the globe. Jorion (2000) contains an excellent discussion of the LTCM case, in particular comparing a Gaussian-based VaR model with a t-based approach. At a more general level, Jorion (2002a) discusses the various fallacies surrounding VaR-based market-risk-management systems.

3
Empirical Properties of Financial Data

In Chapter 2 we saw that the risk that a financial portfolio loses value over a given time period can be modelled in terms of changes in fundamental underlying risk factors, such as equity prices, interest rates and foreign exchange rates (see, in particular, the examples of Section 2.2.1). To build realistic models for risk-management purposes we need to consider the empirical properties of fundamental risk factors and develop models that share these properties.

In this chapter we first consider the univariate properties of single time series of risk-factor changes in Section 3.1, before reviewing some of the properties of multivariate series in Section 3.2. The features we describe motivate the statistical methodology of Part II of this book.

3.1 Stylized Facts of Financial Return Series

The *stylized facts* of financial time series are a collection of empirical observations, and inferences drawn from these observations, that apply to many time series of risk-factor changes, such as log-returns on equities, indices, exchange rates and commodity prices; these observations are now so entrenched in econometric experience that they have been accorded the status of facts.

The stylized facts that we describe typically apply to time series of daily log-returns and often continue to hold when we consider longer-interval series, such as weekly or monthly returns, or shorter-interval series, such as intra-daily returns. Most risk-management models are based on data collected at these frequencies.

Very-high-frequency financial time series, such as tick-by-tick data, have their own stylized facts, but this will not be a subject of this chapter. Moreover, the properties of very-low-frequency data (such as annual returns) are more difficult to pin down, due to the sparseness of such data and the difficulty of assuming that they are generated under long-term stationary regimes.

For a single time series of financial returns, a version of the stylized facts is as follows.

(1) Return series are not independent and identically distributed (iid), although they show little serial correlation.

(2) Series of absolute or squared returns show profound serial correlation.

(3) Conditional expected returns are close to zero.

(4) Volatility appears to vary over time.

(5) Extreme returns appear in clusters.

(6) Return series are leptokurtic or heavy tailed.

To discuss these observations further we denote a return series by $X_1, \dots, X_n$ and assume that the returns have been formed by logarithmic differencing of a price, index or exchange-rate series $(S_t)_{t=0,1,\dots,n}$, so $X_t = \ln(S_t/S_{t-1})$, $t = 1, \dots, n$.

3.1.1 Volatility Clustering

Evidence for the first two stylized facts is collected in Figures 3.1 and 3.2. Figure 3.1 (a) shows 2608 daily log-returns for the DAX index spanning a decade from 2 January 1985 to 30 December 1994, a period including both the stock market crash of 1987 and the reunification of Germany. Parts (b) and (c) show series of simulated iid data from a normal model and a Student t model, respectively; in both cases the model parameters have been set by fitting the model to the real return data using the method of maximum likelihood under the iid assumption. In the normal case, this means that we simply simulate iid data with distribution $N(\mu, \sigma^2)$, where $\mu = \bar{X} = n^{-1}\sum_{i=1}^{n} X_i$ and $\sigma^2 = n^{-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. In the t case, the likelihood has been maximized numerically and the estimated degrees of freedom parameter is $\nu = 3.8$.

The simulated normal data are clearly very different from the DAX return data and do not show the same range of extreme values. While the Student t model can generate comparable extreme values to the real data, more careful observation reveals that the real returns exhibit a phenomenon known as *volatility clustering*, which is not present in the simulated series. Volatility clustering is the tendency for extreme returns to be followed by other extreme returns, although not necessarily with the same sign. We can see periods such as the stock market crash of October 1987 or the political uncertainty in the period between late 1989 and German reunification in 1990 in the DAX data: they are marked by large positive and negative moves.

In Figure 3.2 the *correlograms* of the raw data and the absolute data for all three data sets are shown. The correlogram is a graphical display for estimates of serial correlation, and its construction and interpretation are discussed in Section 4.1.3. While there is very little evidence of serial correlation in the raw data for all data sets, the absolute values of the real financial data appear to show evidence of serial dependence. Clearly, more than 5% of the estimated correlations lie outside the dashed lines, which are the 95% confidence intervals for serial correlations when the underlying process consists of iid finite-variance rvs. This serial dependence in the absolute returns would be equally apparent in squared return values, and it seems to confirm the presence of volatility clustering. We conclude that, although there is no evidence against the iid hypothesis for the genuinely iid data, there is strong evidence against the iid hypothesis for the DAX return data.

Table 3.1 contains more evidence against the iid hypothesis for daily stock-return data. The Ljung–Box test of randomness (described in Section 4.1.3) has been performed for the stocks comprising the Dow Jones 30 index in the period 1993–2000.

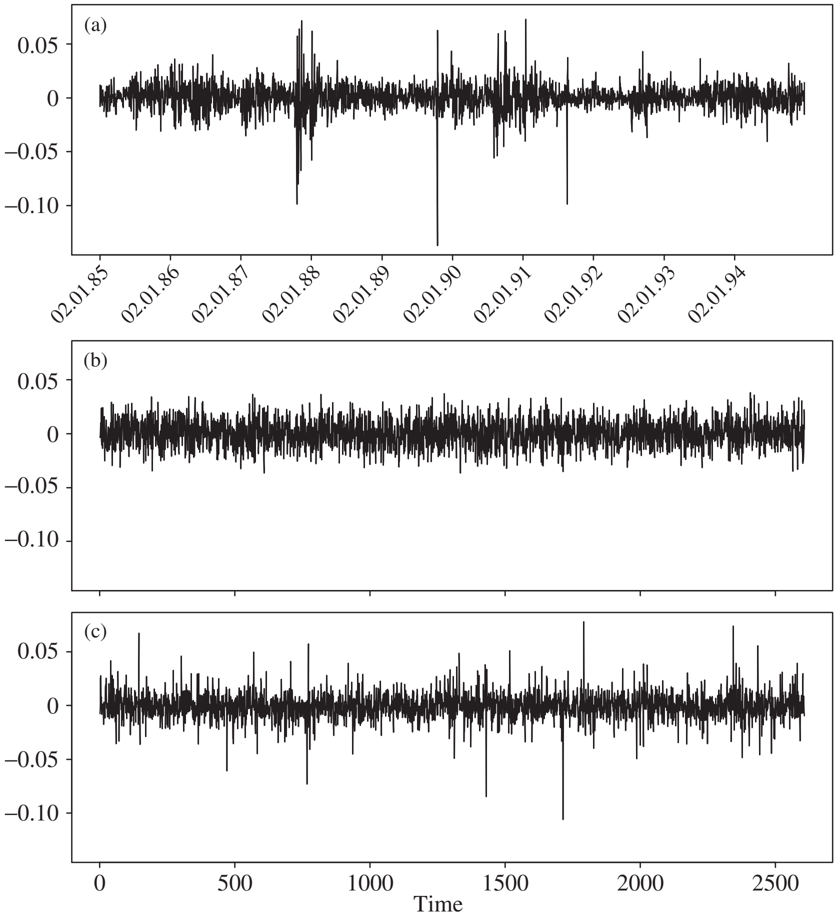

Figure 3.1. (a) Log-returns for the DAX index from 2 January 1985 to 30 December 1994 compared with simulated iid data from (b) a normal and (c) a t distribution, where the parameters have been determined by fitting the models to the DAX data.

In the two columns for daily returns the test is applied, respectively, to the raw return data (LBraw) and their absolute values (LBabs), and p-values are tabulated; these show strong evidence (particularly when applied to absolute values) against the iid hypothesis. If financial log-returns are not iid, then this contradicts the popular *random-walk* hypothesis for the discrete-time development of log-prices (or, in this case, index values). If log-returns are neither iid nor normal, then this contradicts the *geometric Brownian motion* hypothesis for the continuous-time development of prices on which the Black–Scholes–Merton pricing theory is based.

Moreover, if there is serial dependence in financial return data, then the question arises: to what extent can this dependence be used to make *predictions* about the future? This is the subject of the third and fourth stylized facts. It is very difficult to predict the return in the next time period based on historical data alone. This difficulty in predicting future returns is part of the evidence for the well-known

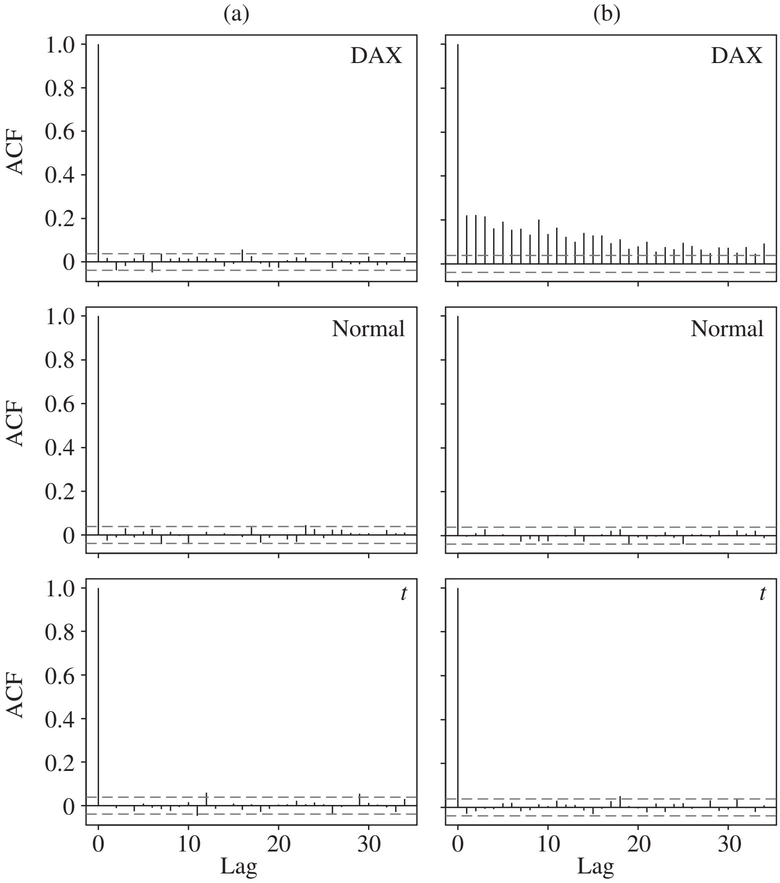

Figure 3.2. Correlograms for (a) the three data sets in Figure 3.1 and (b) the absolute values of these data. Dotted lines mark the standard 95% confidence intervals for the autocorrelations of a process of iid finite-variance rvs.

efficient markets hypothesis in finance, which says that prices react quickly to reflect all the available information about the asset in question.

In empirical terms, the lack of predictability of returns is shown by a lack of serial correlation in the raw return series data. For some series we do sometimes see evidence of correlations at the first lag (or first few lags). A small positive correlation at the first lag would suggest that there is some discernible tendency for a return with a particular sign (positive or negative) to be followed in the next period by a return with the same sign. However, this is not apparent in the DAX data, which suggests that our best estimate for tomorrow's return based on our observations up to today is zero. This idea is expressed in the assertion of the third stylized fact: that conditional expected returns are close to zero.

Volatility is often formally modelled as the *conditional standard deviation* of financial returns given historical information, and, although the conditional expected returns are consistently close to zero, the presence of volatility clustering suggests

Table 3.1. Tests of randomness for returns of Dow Jones 30 stocks in the eight-year period 1993–2000. The columns LBraw and LBabs show p-values for Ljung–Box tests applied to the raw and absolute values, respectively.

		Daily		Monthly	
Name	Symbol	LBraw	LBabs	LBraw	LBabs
Alcoa	AA	0.00	0.00	0.23	0.02
American Express	AXP	0.02	0.00	0.55	0.07
AT&T	T	0.11	0.00	0.70	0.02
Boeing	BA	0.03	0.00	0.90	0.17
Caterpillar	CAT	0.28	0.00	0.73	0.07
Citigroup	C	0.09	0.00	0.91	0.48
Coca-Cola	KO	0.00	0.00	0.50	0.03
DuPont	DD	0.03	0.00	0.75	0.00
Eastman Kodak	EK	0.15	0.00	0.61	0.54
Exxon Mobil	XOM	0.00	0.00	0.32	0.22
General Electric	GE	0.00	0.00	0.25	0.09
General Motors	GM	0.65	0.00	0.81	0.27
Hewlett-Packard	HWP	0.09	0.00	0.21	0.02
Home Depot	HD	0.00	0.00	0.00	0.41
Honeywell	HON	0.44	0.00	0.07	0.30
Intel	INTC	0.23	0.00	0.79	0.62
IBM	IBM	0.18	0.00	0.67	0.28
International Paper	IP	0.15	0.00	0.01	0.09
JPMorgan	JPM	0.52	0.00	0.43	0.12
Johnson & Johnson	JNJ	0.00	0.00	0.11	0.91
McDonald's	MCD	0.28	0.00	0.72	0.68
Merck	MRK	0.05	0.00	0.53	0.65
Microsoft	MSFT	0.28	0.00	0.19	0.13
3M	MMM	0.00	0.00	0.57	0.33
Philip Morris	MO	0.01	0.00	0.68	0.82
Procter & Gamble	PG	0.02	0.00	0.99	0.74
SBC	SBC	0.05	0.00	0.13	0.00
United Technologies	UTX	0.00	0.00	0.12	0.01
Wal-Mart	WMT	0.00	0.00	0.41	0.64
Disney	DIS	0.44	0.00	0.01	0.51

that conditional standard deviations are continually changing in a partly predictable manner. If we know that returns have been large in the last few days, due to market excitement, then there is reason to believe that the distribution from which tomorrow's return is "drawn" should have a large variance. It is this idea that lies behind the time-series models for changing volatility that we will examine in Chapter 4.

Further evidence for volatility clustering is given in Figure 3.3, where the time series of the 100 largest daily losses for the DAX returns and the 100 largest values for the simulated t data are plotted. In Section 5.3.1 we summarize the theory that suggests that the very largest values in iid data will occur like events in a

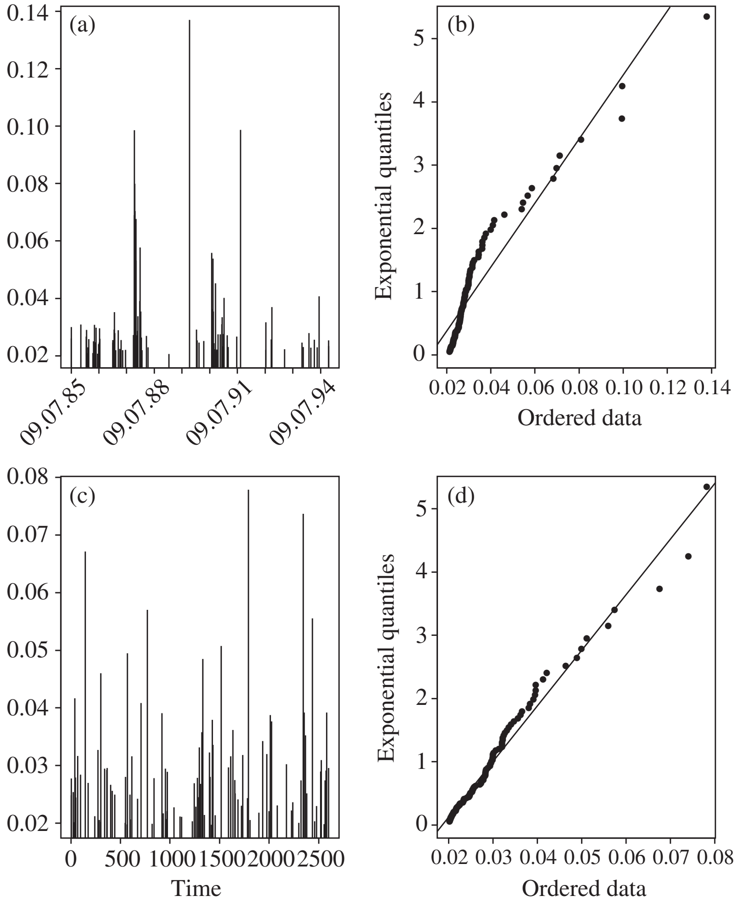

Figure 3.3. Time-series plots of the 100 largest negative values for (a) the DAX returns and (c) the simulated t data as well as (b), (d) Q–Q plots of the waiting times between these extreme values against an exponential reference distribution.

Poisson process, separated by waiting times that are iid with an exponential distribution. Parts (b) and (d) of the figure show Q–Q plots of these waiting times against an exponential reference distribution. While the hypothesis of the Poisson occurrence of extreme values for the iid data is supported, there are too many short waiting times and long waiting times caused by the clustering of extreme values in the DAX data to support the exponential hypothesis. The fifth stylized fact therefore constitutes further strong evidence against the iid hypothesis for return data.

In Chapter 4 we will introduce time-series models that have the volatility clustering behaviour that we observe in real return data. In particular, we will describe ARCH and GARCH models, which can replicate all of the stylized facts we have discussed so far, as well as the typical non-normality of returns addressed by the sixth stylized fact, to which we now turn.

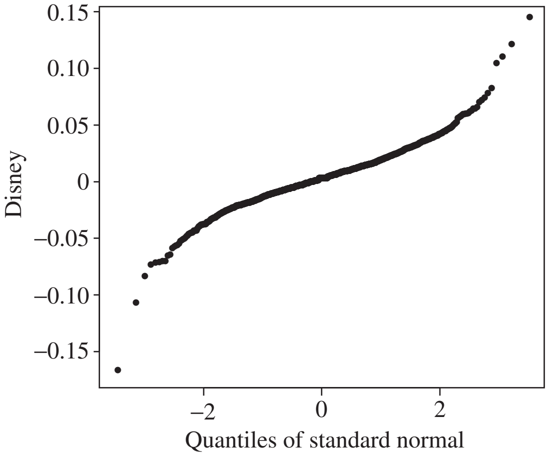

Figure 3.4. A Q–Q plot of daily returns of the Disney share price from 1993 to 2000 against a normal reference distribution.

3.1.2 Non-normality and Heavy Tails

The normal distribution is frequently observed to be a poor model for daily, weekly and even monthly returns. This can be confirmed using various well-known tests of normality, including the Q–Q plot against a standard normal reference distribution, as well as a number of formal numerical tests.

A Q–Q plot (quantile–quantile plot) is a standard visual tool for showing the relationship between empirical quantiles of the data and theoretical quantiles of a reference distribution. A lack of linearity in the Q–Q plot is interpreted as evidence against the hypothesized reference distribution. In Figure 3.4 we show a Q–Q plot of daily returns of the Disney share price from 1993 to 2000 against a normal reference distribution; the inverted S-shaped curve of the points suggests that the more extreme empirical quantiles of the data tend to be larger than the corresponding quantiles of a normal distribution, indicating that the normal distribution is a poor model for these returns.

Common numerical tests include those of Jarque and Bera, Anderson and Darling, Shapiro and Wilk, and D'Agostino. The Jarque–Bera test belongs to the class of omnibus moment tests, i.e. tests that assess simultaneously whether the *skewness* and *kurtosis* of the data are consistent with a Gaussian model. The sample skewness and kurtosis coefficients are defined by

$$b = \frac{(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^3}{((1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2)^{3/2}}, \qquad k = \frac{(1/n)\sum_{i=1}^{n}(X_i - \bar{X})^4}{((1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2)^2}. \tag{3.1}$$

These are designed to estimate the theoretical skewness and kurtosis, which are defined, respectively, by $\beta = E(X - \mu)^3/\sigma^3$ and $\kappa = E(X - \mu)^4/\sigma^4$, where $\mu = E(X)$ and $\sigma^2 = \mathrm{var}(X)$ denote mean and variance; β and κ take the values zero and three for a normal variate X. The Jarque–Bera test statistic is

$$T = \tfrac{1}{6}n(b^2 + \tfrac{1}{4}(k - 3)^2), \tag{3.2}$$

Table 3.2. Sample skewness (b) and kurtosis (k) coefficients as well as p-values for Jarque–Bera tests of normality for an arbitrary set of ten of the Dow Jones 30 stocks (see Table 3.1 for names of stocks).

	Daily returns, $n = 2020$			Weekly returns, $n = 416$		
Stock	b	k	p-value	b	k	p-value
AXP	0.05	5.09	0.00	−0.01	3.91	0.00
EK	−1.93	31.20	0.00	−1.13	14.40	0.00
BA	−0.34	10.89	0.00	−0.26	7.54	0.00
C	0.21	5.93	0.00	0.44	5.42	0.00
KO	−0.02	6.36	0.00	−0.21	4.37	0.00
MSFT	−0.22	8.04	0.00	−0.14	5.25	0.00
HWP	−0.23	6.69	0.00	−0.26	4.66	0.00
INTC	−0.56	8.29	0.00	−0.65	5.20	0.00
JPM	0.14	5.25	0.00	−0.20	4.93	0.00
DIS	−0.01	9.39	0.00	0.08	4.48	0.00

	Monthly returns, $n = 96$			Quarterly returns, $n = 32$		
Stock	b	k	p-value	b	k	p-value
AXP	−1.22	5.99	0.00	−1.04	4.88	0.01
EK	−1.52	10.37	0.00	−0.63	4.49	0.08
BA	−0.50	4.15	0.01	−0.15	6.23	0.00
C	−1.10	7.38	0.00	−1.61	7.13	0.00
KO	−0.49	3.68	0.06	−1.45	5.21	0.00
MSFT	−0.40	3.90	0.06	−0.56	2.90	0.43
HWP	−0.33	3.47	0.27	−0.38	3.64	0.52
INTC	−1.04	6.50	0.00	−0.42	3.10	0.62
JPM	−0.51	5.40	0.00	−0.78	7.26	0.00
DIS	0.04	3.26	0.87	−0.49	4.32	0.16

and it has an asymptotic chi-squared distribution with two degrees of freedom under the null hypothesis of normality; sample kurtosis values differing widely from three and skewness values differing widely from zero may lead to rejection of normality.

In Table 3.2 tests of normality are applied to an arbitrary subgroup of ten of the stocks comprising the Dow Jones index. We take eight years of data spanning the period 1993–2000 and form daily, weekly, monthly and quarterly logarithmic returns. For each stock we calculate sample skewness and kurtosis and apply the Jarque–Bera test to the univariate time series. The daily and weekly return data fail all tests; in particular, it is notable that there are some large values for the sample kurtosis. For the monthly data, the null hypothesis of normality is not formally rejected (that is, the p-value is greater than 0.05) for four of the stocks; for quarterly data, it is not rejected for five of the stocks, although here the sample size is small.

The Jarque–Bera test (3.2) clearly rejects the normal hypothesis. In particular, daily financial return data appear to have a much higher kurtosis than is consistent with the normal distribution; their distribution is said to be *leptokurtic*, meaning that it is more narrow than the normal distribution in the centre but has longer and heavier tails.

Further empirical analysis often suggests that the distribution of daily or other short-interval financial return data has tails that decay slowly according to a power law, rather than the faster, exponential-type decay of the tails of a normal distribution. This means that we tend to see rather more extreme values than might be expected in such return data; we discuss this phenomenon further in Chapter 5, which is devoted to extreme value theory.

3.1.3 Longer-Interval Return Series

As we progressively increase the interval of the returns by moving from daily to weekly, monthly, quarterly and yearly data, the phenomena we have identified tend to become less pronounced. Volatility clustering decreases and returns begin to look both more iid and less heavy tailed.

Beginning with a sample of n returns measured at some time interval (say daily or weekly), we can aggregate these to form longer-interval log-returns. The h-period log-return at time t is given by

$$X_t^{(h)} = \ln\left(\frac{S_t}{S_{t-h}}\right) = \ln\left(\frac{S_t}{S_{t-1}} \cdots \frac{S_{t-h+1}}{S_{t-h}}\right) = \sum_{j=0}^{h-1} X_{t-j}, \tag{3.3}$$

and from our original sample we can form a sample of *non-overlapping* h-period returns $\{X_t^{(h)} : t = h, 2h, \dots, \lfloor n/h \rfloor h\}$, where $\lfloor \cdot \rfloor$ denotes the integer part or *floor* function; $\lfloor x \rfloor = \max\{k \in \mathbb{Z} : k \leqslant x\}$ is the largest integer not greater than x.

Due to the sum structure of the h-period returns, it is to be expected that some central limit effect takes place, whereby their distribution becomes less leptokurtic and more normal as h is increased. Note that, although we have cast doubt on the iid model for daily data, a central limit theorem also applies to many stationary time-series processes, including the GARCH models that are a focus of Chapter 4.

In Table 3.1 the Ljung–Box tests of randomness have also been applied to non-overlapping monthly return data. For twenty out of thirty stocks the null hypothesis of iid data is not rejected at the 5% level in Ljung–Box tests applied to both the raw and absolute returns. It is therefore harder to find evidence of serial dependence in such monthly returns.

Aggregating data to form non-overlapping h-period returns reduces the sample size from n to $\lfloor n/h \rfloor$, and for longer-period returns (such as quarterly or yearly returns) this may be a very serious reduction in the amount of data. An alternative in this case is to form overlapping returns. For $1 \leqslant k < h$ we can form overlapping returns by taking

$$\{X_t^{(h)} : t = h, h+k, h+2k, \dots, h + \lfloor (n-h)/k \rfloor k\}, \tag{3.4}$$

which yields $1 + \lfloor (n-h)/k \rfloor$ values that overlap by an amount $h - k$. In forming overlapping returns we can preserve a large number of data points, but we do build additional serial dependence into the data. Even if the original data were iid, overlapping data would be profoundly dependent, which can greatly complicate their analysis.

Notes and Comments

A number of texts contain extensive empirical analyses of financial return series and discussions of their properties. We mention in particular Taylor (2008), Alexander (2001), Tsay (2002) and Zivot and Wang (2003). For more discussion of the random-walk hypothesis for stock returns, and its shortcomings, see Lo and MacKinlay (1999).

There are countless possible tests of univariate normality and a good starting point is the entry on "departures from normality, tests for" in Volume 2 of the *Encyclopedia of Statistics* (Kotz, Johnson and Read 1985). For an introduction to Q–Q plots see Rice (1995, pp. 353–357); for the widely applied Jarque–Bera test based on the sample skewness and kurtosis, see Jarque and Bera (1987).

3.2 Multivariate Stylized Facts

In risk-management applications we are usually interested in multiple series of financial risk-factor changes. To the stylized facts identified in Section 3.1 we may add a number of stylized facts of a multivariate nature.

We now consider multivariate return data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. Each *component series* $X_{1,j}, \dots, X_{n,j}$ for $j = 1, \dots, d$ is a series formed by logarithmic differencing of a daily price, index or exchange-rate series as before. Commonly observed multivariate stylized facts include the following.

(M1) Multivariate return series show little evidence of cross-correlation, except for contemporaneous returns.

(M2) Multivariate series of absolute returns show profound evidence of cross-correlation.

(M3) Correlations between series (i.e. between contemporaneous returns) vary over time.

(M4) Extreme returns in one series often coincide with extreme returns in several other series.

3.2.1 Correlation between Series

The first two observations are fairly obvious extensions of univariate stylized facts (1) and (2) from Section 3.1. Just as the stock returns for, say, Microsoft on days t and $t+h$ (for $h > 0$) show very little serial correlation, so we generally detect very little correlation between the Microsoft return on day t and, say, the Coca-Cola return on day $t + h$. Of course, stock returns on the same day may show non-negligible correlation, due to factors that affect the whole market on that day. When we look at absolute returns we should bear in mind that periods of high or low volatility are generally common to more than one stock. Returns of large magnitude in one stock may therefore tend to be followed on subsequent days by further returns of large magnitude for both that stock and other stocks, which can explain (M2). The issue of cross-correlation and its estimation is a topic in multivariate time-series analysis and is addressed with an example in Section 14.1.

Stylized fact (M3) is a multivariate counterpart to univariate observation (4): that volatility appears to vary with time. It can be interpreted in a couple of ways with reference to different underlying models. On the one hand, we could refer to a model in which there are so-called stationary regimes; during these regimes correlations are fixed but the regimes change from time to time. On the other hand, we could refer to more dynamic models in which a *conditional correlation* is changing all the time. Just as volatility is often formally modelled as the conditional standard deviation of returns given historical information, we can also devise models that feature a changing conditional correlation given historical information. Examples of such models include certain multivariate GARCH models, as discussed in Section 14.2. In the context of such models it is possible to demonstrate (M3) for many pairs of risk-factor return series.

However, we should be careful about drawing conclusions about changing correlations based on more ad hoc analyses. To explain this further we consider two data sets. The first, shown in Figure 3.5, comprises the BMW and Siemens daily log-return series for the period from 23 January 1985 to 22 September 1994; there are precisely 2000 values.

The second data set, shown in Figure 3.6, consists of an equal quantity of randomly generated data from a bivariate t distribution. The parameters have been chosen by fitting the distribution to the BMW-Siemens data by the method of maximum likelihood. The fitted model is estimated to have 2.8 degrees of freedom and estimated correlation 0.72 (see Section 6.2.1 for more details of the multivariate t distribution).

The two data sets show some superficial resemblance. The distribution of values is similar in both cases. However, the simulated data are independent and there is no serial dependence or volatility clustering.

We estimate rolling correlation coefficients for both series using a moving window of twenty-five days, which is approximately the number of trading days in a typical calendar month. These kinds of rolling empirical estimates are quite commonly used in practice to gather evidence of how key model parameters may change. In Figure 3.7 the resulting estimates are shown for the BMW–Siemens log-return data and the iid Student t-distributed data.

Remarkably, there are no obvious differences between the results for the two data sets; if anything, the range of estimated correlation values for the iid data is greater, despite the fact that they are generated from a single stationary model with a fixed correlation of 0.72.

This illustrates that simple attempts to demonstrate (M3) using empirical correlation estimates should be interpreted with care. There is considerable error involved in estimating correlations from small samples, particularly when the underlying distribution is a heavier-tailed bivariate distribution, such as a t distribution, rather than a Gaussian distribution (see also Example 6.30 in this context). The most reliable way to substantiate (M3) and to decide in exactly what way correlation changes is to fit different models for changing correlation and then to make formal statistical comparisons of the models.

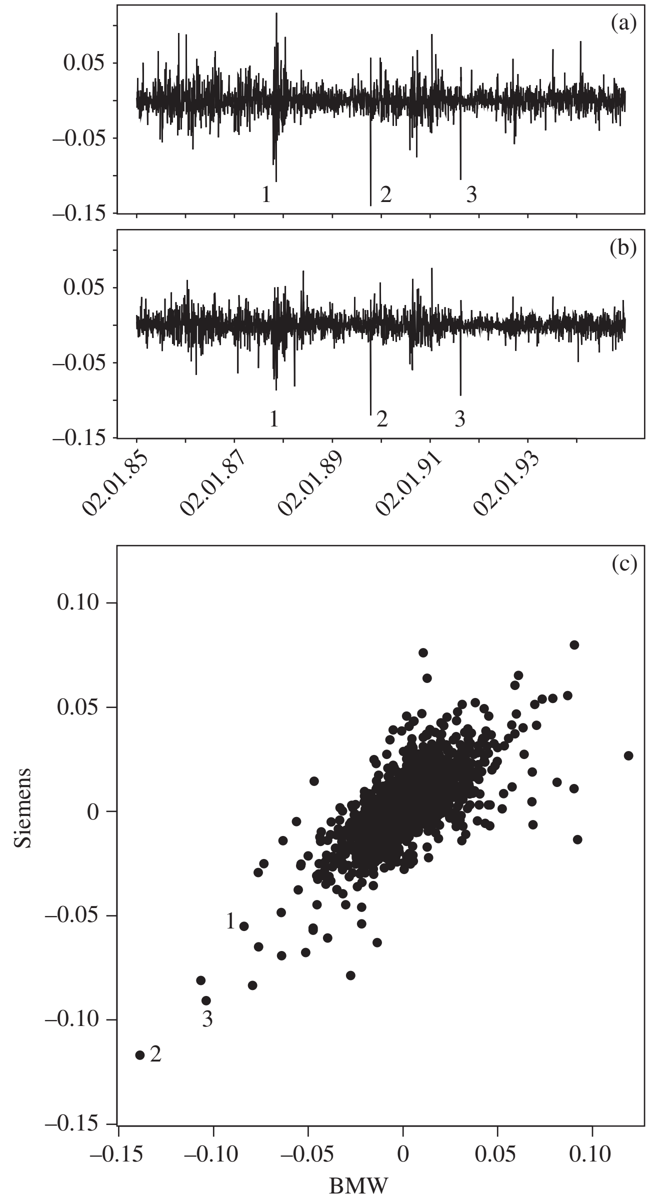

Figure 3.5. (a) BMW and (b) Siemens log-return data for the period from 23 January 1985 to 22 September 1994 together with (c) pairwise scatterplot. Three extreme days on which large negative returns occurred are marked. The dates are 19 October 1987, 16 October 1989 and 19 August 1991 (see Section 3.2.2 for historical commentary).

3.2.2 Tail Dependence

Stylized fact (M4) is often apparent when time series are compared. Consider again the BMW and Siemens log-returns in Figure 3.5. In both the time-series plots and the scatterplot, three days have been indicated with a number. These are days on

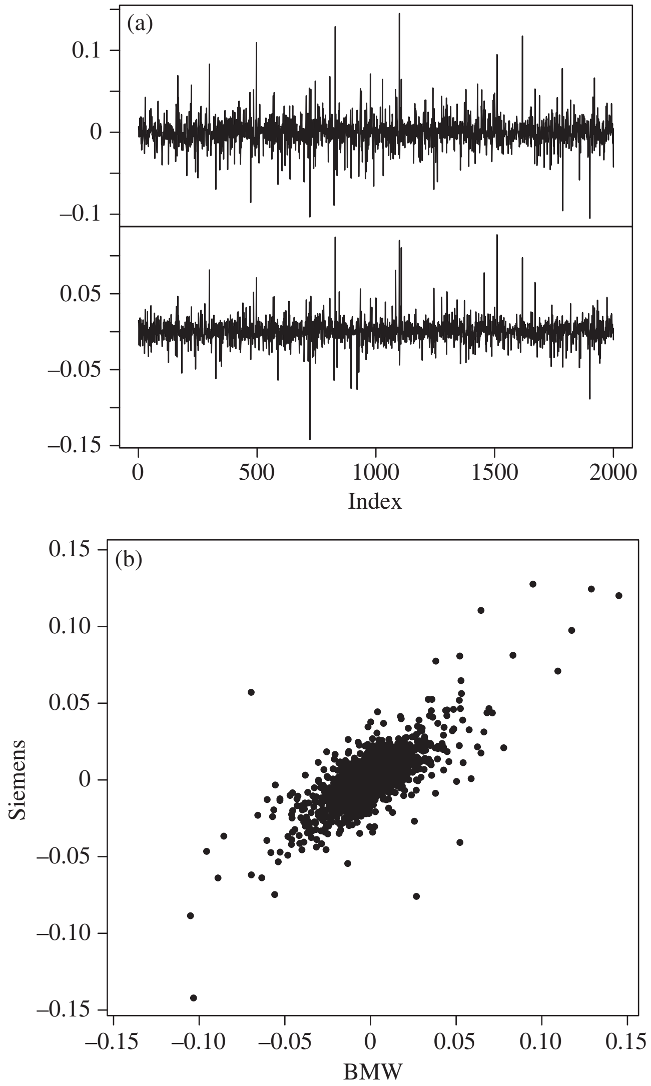

Figure 3.6. Two-thousand iid data points generated from a bivariate *t* distribution: (a) time series of components and (b) pairwise scatterplot. The parameters have been set by fitting the *t* distribution to the BMW–Siemens data in Figure 3.5.

which large negative returns were observed for both stocks, and all three occurred during periods of volatility on the German market. They are, respectively, 19 October 1987, Black Monday on Wall Street; 16 October 1989, when over 100 000 Germans protested against the East German regime in Leipzig during the chain of events that led to the fall of the Berlin Wall and German reunification; and 19 August 1991, the day of the coup by communist hardliners during the reforming era of Gorbachev in the USSR. Clearly, these are days on which momentous events led to joint extreme values.

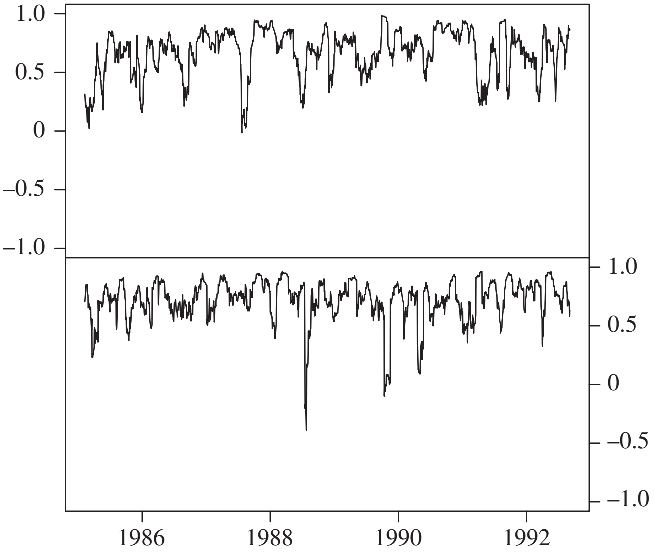

Figure 3.7. Twenty-five-day rolling correlation estimates for the empirical data of Figure 3.5 (top panel) and for the simulated iid data of Figure 3.6 (bottom panel).

Related to stylized fact (M4) is the idea that "correlations go to one in times of market stress". The three extreme days in Figure 3.5 correspond to points that are close to the diagonal of the scatterplot in the lower-left-hand corner, and it is easy to see why one might describe these as occasions on which correlations tend to one. It is quite difficult to formally test the hypothesis that model correlations are higher when volatilities are higher, and this should be done in the context of a multivariate time-series model incorporating either dynamic conditional correlations or regime changes. Once again, we should be cautious about interpreting simple analyses based on empirical correlation estimates, as we now show.

In Figure 3.8 we perform an analysis in which we split the 2000 bivariate return observations into eighty non-overlapping twenty-five-day blocks. In each block we estimate the empirical correlation between the return series and the volatility of the two series. We plot the *Fisher transform* of the estimated correlation against the estimated volatility of the BMW series and then regress the former on the latter. There is a strongly significant regression relationship (shown by the line) between the correlation and volatility estimates. It is tempting to say that in stress periods where volatility is high, correlation is also high. The Fisher transform is a well-known variance-stabilizing transform that is appropriate when correlation is the dependent variable in a regression analysis.

However, when exactly the same exercise is carried out for the data generated from a t distribution (Figure 3.6 (b)), the result is similar. In this case the observation that estimated correlation is higher in periods of higher estimated volatility is a pure artefact of estimation error for both quantities, since the true underlying correlation is fixed.

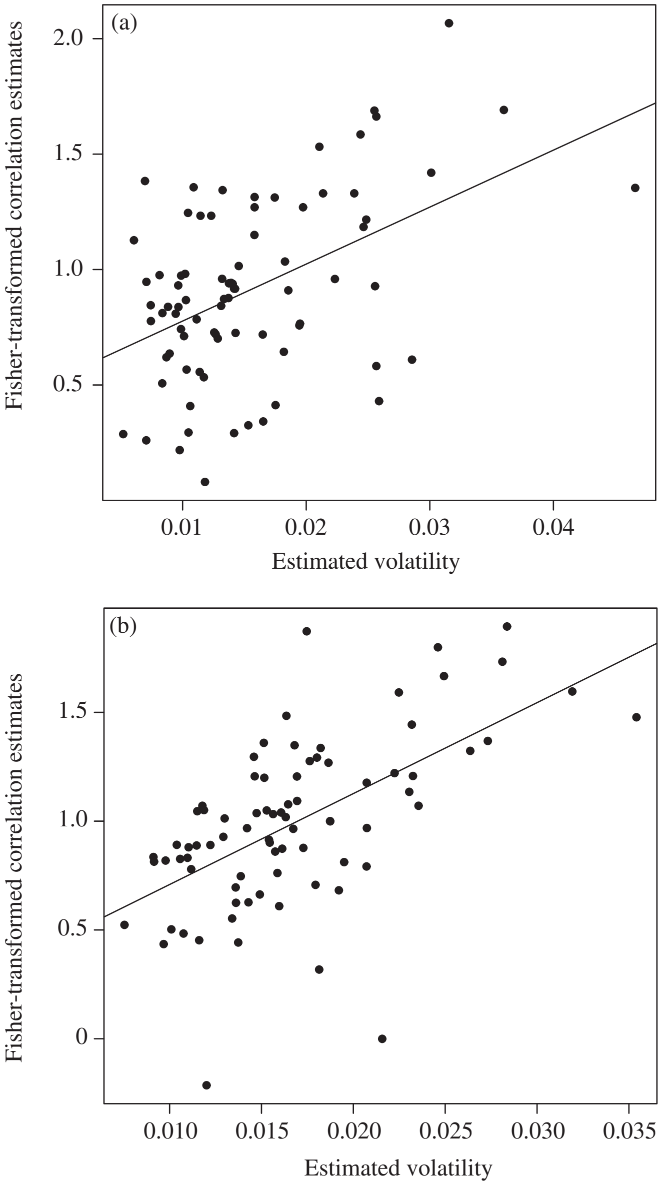

Figure 3.8. Fisher transforms of estimated correlations plotted against estimated volatilities for eighty non-overlapping blocks of twenty-five observations: (a) the BMW–Siemens log-return data in Figure 3.5; and (b) the simulated bivariate t data in Figure 3.6. In both cases there is a significant regression relationship between estimated correlations and estimated volatilities.

This example is not designed to argue against the view that correlations are higher when volatilities are higher; it is simply meant to show that it is difficult to demonstrate this using an ad hoc approach based on estimated correlations. Formal comparison of different multivariate volatility and correlation models with differing specifications for correlation is required; some models of this kind are described in Section 14.2. Moreover, while it may be partly true that useful multivariate time-series models for returns should have the property that conditional correlations tend to become large when volatilities are large, the phenomenon of simultaneous extreme values can also be addressed in other ways.

For example, we can choose distributions in multivariate models that have so-called *tail dependence* or *extremal dependence*. Loosely speaking, this means models in which the conditional probability of observing an extreme value for one risk factor given that we have observed an extreme value for another is non-negligible and, indeed, is in some cases quite large. A mathematical definition of this notion and a discussion of its importance may be found in Section 7.2.4.

Notes and Comments

Pitfalls in tests for changing correlations are addressed in an interesting paper by Boyer, Gibson and Loretan (1999), which argues against simplistic empirical analyses based on segmenting the data into normal and stressed regimes. See also Loretan and English (2000) for a discussion of correlation breakdowns during periods of market instability. An interesting book on the importance of correlation in risk management is Engle (2009).

Tail dependence has various definitions: see Joe (1997) and Coles, Heffernan and Tawn (1999). The importance of tail dependence in risk management was highlighted in Embrechts, McNeil and Straumann (1999), Embrechts, McNeil and Straumann (2002) and Mashal, Naldi and Zeevi (2003). It is now a recognized issue in the regulatory literature: see, for example, the discussion of tail correlation in the CEIOPS consultation paper on the use of correlation in the standard formula for the solvency capital requirement (CEIOPS 2009).

Part II

Methodology

4
Financial Time Series

Motivated by the discussion of the empirical properties of financial risk-factor change data in Chapter 3, in this chapter we present univariate time-series models that mimic the properties of real return data.

In Section 4.1 we review essential concepts in the analysis of time series, such as stationarity, autocorrelations and their estimation, white noise processes, and ARMA (autoregressive moving-average) processes. We then devote Section 4.2 to univariate ARCH and GARCH (generalized autoregressive conditionally heteroscedastic) processes for capturing the important phenomenon of volatility.

GARCH models are certainly not the only models for describing the volatility of financial returns. Other important classes of model include discrete-time stochastic volatility models, long-memory GARCH models, continuous-time models fitted to discrete data, and models based on realized volatility calculated from high-frequency data; these alternative approaches are not handled in this book.

Our emphasis on GARCH has two main motivations, the first being a practical one. We recall that in risk management we are typically dealing with very large numbers of risk factors, and our philosophy, expounded in Section 1.5, is that broad-brush techniques that capture the main risk features of many time series are more important than very detailed analyses of single series. The GARCH model lends itself to this approach and proves relatively easy to fit. There are also some multivariate extensions (see Chapter 14) that build in simple ways on the univariate models and that may be calibrated to a multivariate series in stages. This ease of use contrasts with other models where the fitting of a single series often presents a computational challenge (e.g. estimation of a stochastic volatility model via filtering or Gibbs sampling), and multivariate extensions have not been widely considered. Moreover, an average financial enterprise will typically collect daily data on its complete set of risk factors for the purposes of risk management, and this rules out some more sophisticated models that require higher-frequency data.

Our second reason for concentrating on ARCH and GARCH models is didactic. These models for volatile return series have a status akin to ARMA models in classical time series; they belong, in our opinion, to the body of standard methodology to which a student of the subject should be exposed. A quantitative risk manager who understands GARCH has a good basis for understanding more complex models and a good framework for talking about historical volatility in a rational way. He/she may also appreciate more clearly the role of more ad hoc volatility estimation methods such as the exponentially weighted moving-average (EWMA) procedure.

4.1 Fundamentals of Time Series Analysis

This section provides a short summary of the essentials of classical univariate time-series analysis with a focus on that which is relevant for modelling risk-factor return series. We have based the presentation on Brockwell and Davis (1991, 2002), so these texts may be used as supplementary reading.

4.1.1 Basic Definitions

A time-series model for a single risk factor is a discrete-time stochastic process $(X_t)_{t\in\mathbb{Z}}$, i.e. a family of rvs, indexed by the integers and defined on some probability space $(\Omega, \mathcal{F}, P)$.

Moments of a time series. Assuming they exist, we define the *mean function* $\mu(t)$ and the *autocovariance function* $\gamma(t, s)$ of $(X_t)_{t\in\mathbb{Z}}$ by

$$\begin{aligned} \mu(t) &= E(X_t), & t &\in \mathbb{Z}, \\ \gamma(t, s) &= E((X_t - \mu(t))(X_s - \mu(s))), & t, s &\in \mathbb{Z}. \end{aligned}$$

It follows that the autocovariance function satisfies $\gamma(t, s) = \gamma(s, t)$ for all t, s, and $\gamma(t, t) = \operatorname{var}(X_t)$.

Stationarity. Generally, the processes we consider will be stationary in one or both of the following two senses.

Definition 4.1 (strict stationarity). The time series $(X_t)_{t\in\mathbb{Z}}$ is *strictly* stationary if

$$(X_{t_1}, \dots, X_{t_n}) \overset{\mathrm{d}}{=} (X_{t_1+k}, \dots, X_{t_n+k})$$

for all $t_1, \dots, t_n, k \in \mathbb{Z}$ and for all $n \in \mathbb{N}$.

Definition 4.2 (covariance stationarity). The time series $(X_t)_{t\in\mathbb{Z}}$ is *covariance* stationary (or *weakly* or *second-order* stationary) if the first two moments exist and satisfy

$$\begin{aligned} \mu(t) &= \mu, & t &\in \mathbb{Z}, \\ \gamma(t, s) &= \gamma(t + k, s + k), & t, s, k &\in \mathbb{Z}. \end{aligned}$$

Both these definitions attempt to formalize the notion that the behaviour of a time series is similar in any epoch in which we might observe it. Systematic changes in mean, variance or the covariances between equally spaced observations are inconsistent with stationarity.

It may be easily verified that a strictly stationary time series with finite variance is covariance stationary, but it is important to note that we may define infinite-variance processes (including certain ARCH and GARCH processes) that are strictly stationary but not covariance stationary.

Autocorrelation in stationary time series. The definition of covariance stationarity implies that for all s, t we have $\gamma(t-s, 0) = \gamma(t, s) = \gamma(s, t) = \gamma(s-t, 0)$, so that the covariance between X_t and X_s only depends on their temporal separation $|s-t|$, which is known as the *lag*. Thus, for a covariance-stationary process we write the autocovariance function as a function of one variable:

$$\gamma(h) := \gamma(h, 0), \quad \forall h \in \mathbb{Z}.$$

Noting that $\gamma(0) = \operatorname{var}(X_t)$, $\forall t$, we can now define the autocorrelation function of a covariance-stationary process.

Definition 4.3 (autocorrelation function). The *autocorrelation function* (ACF) $\rho(h)$ of a covariance-stationary process $(X_t)_{t\in\mathbb{Z}}$ is

$$\rho(h) = \rho(X_h, X_0) = \gamma(h)/\gamma(0), \quad \forall h \in \mathbb{Z}.$$

We speak of the autocorrelation or *serial correlation* $\rho(h)$ at lag h. In classical time-series analysis the set of serial correlations and their empirical analogues estimated from data are the objects of principal interest. The study of autocorrelations is known as *analysis in the time domain*.

White noise processes. The basic building blocks for creating useful time-series models are stationary processes without serial correlation, known as *white noise* processes and defined as follows.

Definition 4.4 (white noise). $(X_t)_{t\in\mathbb{Z}}$ is a white noise process if it is covariance stationary with autocorrelation function

$$\rho(h) = \begin{cases} 1, & h = 0, \\ 0, & h \neq 0. \end{cases}$$

A white noise process centred to have mean 0 with variance $\sigma^2 = \operatorname{var}(X_t)$ will be denoted $\mathrm{WN}(0, \sigma^2)$. A simple example of a white noise process is a series of iid rvs with finite variance, and this is known as a *strict white noise* process.

Definition 4.5 (strict white noise). $(X_t)_{t\in\mathbb{Z}}$ is a strict white noise process if it is a series of iid, finite-variance rvs.

A strict white noise (SWN) process centred to have mean 0 and variance σ^2 will be denoted $\mathrm{SWN}(0, \sigma^2)$. Although SWN is the easiest kind of noise process to understand, it is not the only noise that we will use. We will later see that covariance-stationary ARCH and GARCH processes are in fact white noise processes.

Martingale difference. One further noise concept that we use, particularly when we come to discuss volatility and GARCH processes, is that of a martingale-difference sequence. We recall that a *martingale* is a sequence of integrable rvs (M_t) such that the expected value of M_t given the previous history of the sequence is M_{t-1}. This implies that if we define (X_t) by taking first differences of the sequence (M_t), then the expected value of X_t given information about previous values is 0. We have observed in Section 3.1 that this property may be appropriate for financial return data. A martingale difference is often said to model our winnings in consecutive rounds of a *fair game*.

To discuss this concept more precisely, we assume that the time series $(X_t)_{t\in\mathbb{Z}}$ is adapted to some *filtration* $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ that represents the *accrual of information* over time. The sigma algebra $\mathcal{F}_t$ represents the available information at time t, and typically this will be the information contained in past and present values of the time series itself $(X_s)_{s\leqslant t}$, which we refer to as the *history* up to time t and denote by $\mathcal{F}_t = \sigma(\{X_s : s \leqslant t\})$; the corresponding filtration is known as the *natural filtration*.

Definition 4.6 (martingale difference). The time series $(X_t)_{t\in\mathbb{Z}}$ is known as a martingale-difference sequence with respect to the filtration $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ if $E|X_t| < \infty$, X_t is $\mathcal{F}_t$-measurable (*adapted*) and

$$E(X_t \mid \mathcal{F}_{t-1}) = 0, \quad \forall t \in \mathbb{Z}.$$

Obviously the unconditional mean of such a process is also zero:

$$E(X_t) = E(E(X_t \mid \mathcal{F}_{t-1})) = 0, \quad \forall t \in \mathbb{Z}.$$

Moreover, if $E(X_t^2) < \infty$ for all t, then autocovariances satisfy

$$\begin{aligned}\gamma(t,s) &= E(X_t X_s)\\ &= \begin{cases} E(E(X_t X_s \mid \mathcal{F}_{s-1})) = E(X_t E(X_s \mid \mathcal{F}_{s-1})) = 0, & t < s,\\ E(E(X_t X_s \mid \mathcal{F}_{t-1})) = E(X_s E(X_t \mid \mathcal{F}_{t-1})) = 0, & t > s.\end{cases}\end{aligned}$$

Thus a finite-variance martingale-difference sequence has zero mean and zero covariance. If the variance is constant for all t, it is a white noise process.

4.1.2 ARMA Processes

The family of classical ARMA processes are widely used in many traditional applications of time-series analysis. They are covariance-stationary processes that are constructed using white noise as a basic building block. As a general notational convention in this section and the remainder of the chapter we will denote white noise by $(\varepsilon_t)_{t\in\mathbb{Z}}$ and *strict* white noise by $(Z_t)_{t\in\mathbb{Z}}$.

Definition 4.7 (ARMA process). Let $(\varepsilon_t)_{t\in\mathbb{Z}}$ be $\text{WN}(0, \sigma_\varepsilon^2)$. The process $(X_t)_{t\in\mathbb{Z}}$ is a zero-mean ARMA(p,q) process if it is a covariance-stationary process satisfying difference equations of the form

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}, \quad \forall t \in \mathbb{Z}. \tag{4.1}$$

(X_t) is an ARMA process with mean μ if the centred series $(X_t - \mu)_{t\in\mathbb{Z}}$ is a zero-mean ARMA(p,q) process.

Note that, according to our definition, there is no such thing as a non-covariance-stationary ARMA process. Whether the process is strictly stationary or not will depend on the exact nature of the driving white noise, also known as the process of *innovations*. If the innovations are iid, or themselves form a strictly stationary process, then the ARMA process will also be strictly stationary.

For all practical purposes we can restrict our study of ARMA processes to *causal* ARMA processes. By this we mean processes satisfying the equations (4.1), which have a representation of the form

$$X_t = \sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}, \tag{4.2}$$

where the ψ_i are coefficients that must satisfy

$$\sum_{i=0}^{\infty} |\psi_i| < \infty. \tag{4.3}$$

Remark 4.8. The so-called absolute summability condition (4.3) is a technical condition that ensures that $E|X_t| < \infty$. This guarantees that the infinite sum in (4.2) converges absolutely, almost surely, meaning that both $\sum_{i=0}^{\infty} |\psi_i||\varepsilon_{t-i}|$ and $\sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}$ are finite with probability 1 (see Brockwell and Davis 1991, Proposition 3.1.1).

We now verify by direct calculation that causal ARMA processes are indeed covariance stationary and calculate the form of their autocorrelation function, before going on to look at some simple standard examples.

Proposition 4.9. *Any process satisfying (4.2) and (4.3) is covariance stationary, with an autocorrelation function given by*

$$\rho(h) = \frac{\sum_{i=0}^{\infty} \psi_i \psi_{i+|h|}}{\sum_{i=0}^{\infty} \psi_i^2}, \quad h \in \mathbb{Z}. \tag{4.4}$$

Proof. Obviously, for all t we have $E(X_t) = 0$ and $\operatorname{var}(X_t) = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} \psi_i^2 < \infty$, due to (4.3). Moreover, the autocovariances are given by

$$\operatorname{cov}(X_t, X_{t+h}) = E(X_t X_{t+h}) = E\biggl(\sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i} \sum_{j=0}^{\infty} \psi_j \varepsilon_{t+h-j}\biggr).$$

Since (ε_t) is white noise, it follows that $E(\varepsilon_{t-i}\varepsilon_{t+h-j}) \neq 0 \iff j = i + h$, and hence that

$$\gamma(h) = \operatorname{cov}(X_t, X_{t+h}) = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+|h|}, \quad h \in \mathbb{Z}, \tag{4.5}$$

which depends only on the lag h and not on t. The autocorrelation function follows easily. □

Example 4.10 (MA(q) process). It is clear that a pure moving-average process

$$X_t = \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{4.6}$$

forms a simple example of a causal process of the form (4.2). It is easily inferred from (4.4) that the autocorrelation function is given by

$$\rho(h) = \frac{\sum_{i=0}^{q-|h|} \theta_i \theta_{i+|h|}}{\sum_{i=0}^{q} \theta_i^2}, \quad |h| \in \{0, 1, \dots, q\},$$

where $\theta_0 = 1$. For $|h| > q$ we have $\rho(h) = 0$, and the autocorrelation function is said to *cut off* at lag q. If this feature is observed in the estimated autocorrelations of empirical data, it is often taken as an indicator of moving-average behaviour. A realization of an MA(4) process together with the theoretical form of its ACF is shown in Figure 4.1.

Example 4.11 (AR(1) process). The first-order AR process satisfies the set of difference equations

$$X_t = \phi_1 X_{t-1} + \varepsilon_t, \quad \forall t. \tag{4.7}$$

This process is causal if and only if $|\phi_1| < 1$, and this may be understood intuitively by iterating the equation (4.7) to get

$$\begin{aligned} X_t &= \phi_1(\phi_1 X_{t-2} + \varepsilon_{t-1}) + \varepsilon_{t-2} \\ &= \phi_1^{k+1} X_{t-k-1} + \sum_{i=0}^{k} \phi_1^i \varepsilon_{t-i}. \end{aligned}$$

Using more careful probabilistic arguments it may be shown that the condition $|\phi_1| < 1$ ensures that the first term disappears as $k \to \infty$ and the second term converges. The process

$$X_t = \sum_{i=0}^{\infty} \phi_1^i \varepsilon_{t-i} \tag{4.8}$$

turns out to be the unique solution of the defining equations (4.7). It may be easily verified that this is a process of the form (4.2) and that $\sum_{i=0}^{\infty} |\phi_1|^i = (1 - |\phi_1|)^{-1}$ so that (4.3) is satisfied. Looking at the form of the solution (4.8), we see that the AR(1) process can be represented as an MA(∞) process: an infinite-order moving-average process.

The autocovariance and autocorrelation functions of the process may be calculated from (4.5) and (4.4) to be

$$\gamma(h) = \frac{\phi_1^{|h|} \sigma_\varepsilon^2}{1 - \phi_1^2}, \quad \rho(h) = \phi_1^{|h|}, \quad h \in \mathbb{Z}.$$

Thus the ACF is exponentially decaying with possibly alternating sign. A realization of an AR(1) process together with the theoretical form of its ACF is shown in Figure 4.1.

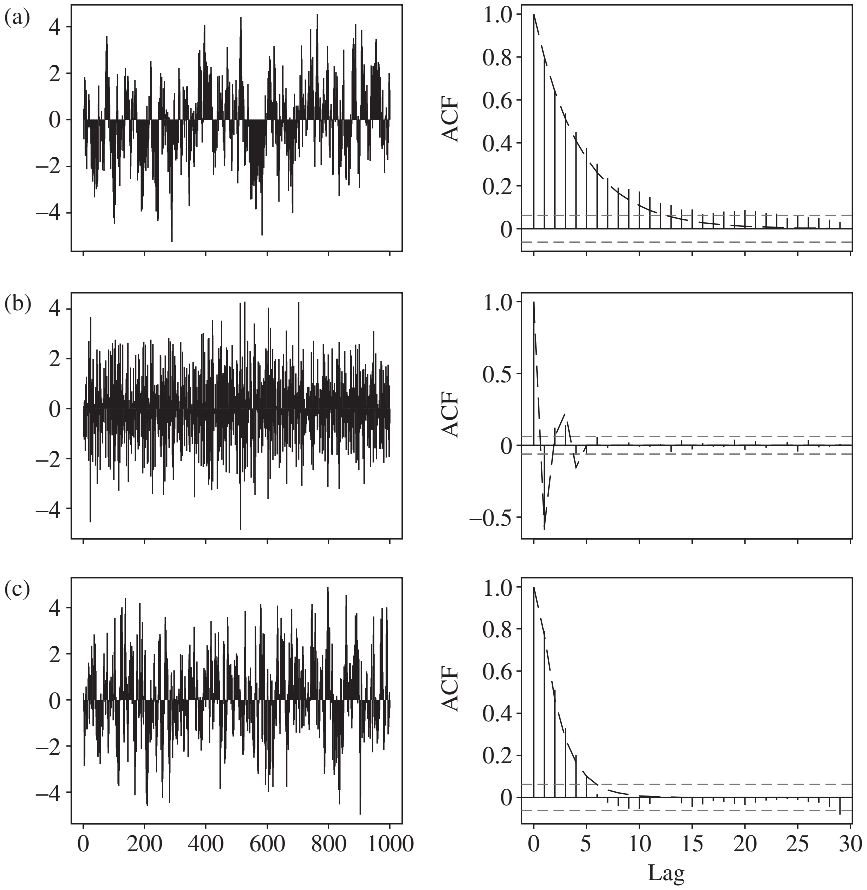

Figure 4.1. A number of simulated ARMA processes with their autocorrelation functions (dashed) and correlograms. Innovations are Gaussian. (a) AR(1), $\phi_1 = 0.8$. (b) MA(4), $\theta_1 = -0.8, 0.4, 0.2, -0.3$. (c) ARMA(1, 1), $\phi_1 = 0.6$, $\theta_1 = 0.5$.

Remarks on general ARMA theory. In the case of the general ARMA process of Definition 4.7, the issue of whether this process has a causal representation of the form (4.2) is resolved by the study of two polynomials in the complex plane, which are given in terms of the ARMA model parameters by

$$\begin{aligned}\tilde{\phi}(z) &= 1 - \phi_1 z - \cdots - \phi_p z^p,\\ \tilde{\theta}(z) &= 1 + \theta_1 z + \cdots + \theta_q z^q.\end{aligned}$$

Provided that $\tilde{\phi}(z)$ and $\tilde{\theta}(z)$ have no common roots, then the ARMA process is a causal process satisfying (4.2) and (4.3) if and only if $\tilde{\phi}(z)$ has no roots in the unit circle $|z| \leqslant 1$. The coefficients ψ_i in the representation (4.2) are determined by the equation

$$\sum_{i=0}^{\infty} \psi_i z^i = \frac{\tilde{\theta}(z)}{\tilde{\phi}(z)}, \quad |z| \leqslant 1.$$

Example 4.12 (ARMA(1, 1) process). For the process given by

$$X_t - \phi_1 X_{t-1} = \varepsilon_t + \theta_1 \varepsilon_{t-1}, \quad \forall t \in \mathbb{Z},$$

the complex polynomials are $\tilde{\phi}(z) = 1 - \phi_1 z$ and $\tilde{\theta}(z) = 1 + \theta_1 z$, and these have no common roots provided $\phi_1 + \theta_1 \neq 0$. The solution of $\tilde{\phi}(z) = 0$ is $z = 1/\phi_1$ and this is outside the unit circle provided $|\phi_1| < 1$, so that this is the condition for causality (as in the AR(1) model of Example 4.11).

The representation (4.2) can be obtained by considering

$$\sum_{i=0}^{\infty} \psi_i z^i = \frac{1 + \theta_1 z}{1 - \phi_1 z} = (1 + \theta_1 z)(1 + \phi_1 z + \phi_1^2 z^2 + \cdots), \quad |z| \leqslant 1,$$

and is easily calculated to be

$$X_t = \varepsilon_t + (\phi_1 + \theta_1) \sum_{i=1}^{\infty} \phi_1^{i-1} \varepsilon_{t-i}. \tag{4.9}$$

Using (4.4) we may calculate that for $h \neq 0$ the ACF is

$$\rho(h) = \frac{\phi_1^{|h|-1}(\phi_1 + \theta_1)(1 + \phi_1 \theta_1)}{1 + \theta_1^2 + 2\phi_1 \theta_1}.$$

A realization of an ARMA(1, 1) process together with the theoretical form of its ACF is shown in Figure 4.1.

Invertibility. Equation (4.9) shows how the ARMA(1, 1) process may be thought of as an MA(∞) process. In fact, if we impose the condition $|\theta_1| < 1$, we can also express (X_t) as the AR(∞) process given by

$$X_t = \varepsilon_t + (\phi_1 + \theta_1) \sum_{i=1}^{\infty} (-\theta_1)^{i-1} X_{t-i}. \tag{4.10}$$

If we rearrange this to be an equation for ε_t, then we see that we can, in a sense, "reconstruct" the latest innovation ε_t from the entire history of the process $(X_s)_{s \leqslant t}$. The condition $|\theta_1| < 1$ is known as an *invertibility* condition, and for the general ARMA(p, q) process the invertibility condition is that $\tilde{\theta}(z)$ should have no roots in the unit circle $|z| \leqslant 1$. In practice, the models we fit to real data will be both invertible and causal solutions of the ARMA-defining equations.

Models for the conditional mean. Consider a general invertible ARMA model with non-zero mean. For what comes later it will be useful to observe that we can write such models as

$$X_t = \mu_t + \varepsilon_t, \quad \mu_t = \mu + \sum_{i=1}^{p} \phi_i (X_{t-i} - \mu) + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}. \tag{4.11}$$

Since we have assumed invertibility, the terms ε_{t-j}, and hence μ_t, can be written in terms of the infinite past of the process up to time $t - 1$; μ_t is said to be *measurable* with respect to $\mathcal{F}_{t-1} = \sigma(\{X_s : s \leqslant t - 1\})$.

If we make the assumption that the white noise $(\varepsilon_t)_{t\in\mathbb{Z}}$ is a martingale-difference sequence (see Definition 4.6) with respect to $(\mathcal{F}_t)_{t\in\mathbb{Z}}$, then $E(X_t \mid \mathcal{F}_{t-1}) = \mu_t$. In other words, such an ARMA process can be thought of as putting a particular structure on the conditional mean μ_t of the process. ARCH and GARCH processes will later be seen to put structure on the conditional variance $\operatorname{var}(X_t \mid \mathcal{F}_{t-1})$.

ARIMA models. In traditional time-series analysis we often consider an even larger class of model known as ARIMA models, or autoregressive *integrated* moving-average models. Let ∇ denote the *difference operator*, so that for a time-series process $(Y_t)_{t\in\mathbb{Z}}$ we have $\nabla Y_t = Y_t - Y_{t-1}$. Denote repeated differencing by ∇^d, where

$$\nabla^d Y_t = \begin{cases} \nabla Y_t, & d = 1, \\ \nabla^{d-1}(\nabla Y_t) = \nabla^{d-1}(Y_t - Y_{t-1}), & d > 1. \end{cases} \tag{4.12}$$

The time series (Y_t) is said to be an ARIMA(p, d, q) process if the differenced series (X_t) given by $X_t = \nabla^d Y_t$ is an ARMA(p, q) process. For $d > 1$, ARIMA processes are non-stationary processes. They are popular in practice because the operation of differencing (once or more than once) can turn a data set that is obviously "non-stationary" into a data set that might plausibly be modelled by a stationary ARMA process. For example, if we use an ARMA(p, q) process to model daily log-returns of some price series (S_t), then we are really saying that the original logarithmic price series $(\ln S_t)$ follows an ARIMA$(p, 1, q)$ model.

When the word *integrated* is used in the context of time series it generally implies that we are looking at a non-stationary process that might be made stationary by differencing; see also the discussion of IGARCH models in Section 4.2.2.

4.1.3 Analysis in the Time Domain

We now assume that we have a sample $X_1, \dots, X_n$ from a covariance-stationary time-series model $(X_t)_{t\in\mathbb{Z}}$. Analysis in the time domain involves calculating empirical estimates of autocovariances and autocorrelations from this random sample and using these estimates to make inferences about the serial dependence structure of the underlying process.

Correlogram. The sample autocovariances are calculated according to

$$\hat{\gamma}(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h} - \bar{X})(X_t - \bar{X}), \quad 0 \leqslant h < n,$$

where $\bar{X} = \sum_{t=1}^{n} X_t/n$ is the sample mean, which estimates μ, the mean of the time series. From these we calculate the sample ACF:

$$\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0), \quad 0 \leqslant h < n.$$

The *correlogram* is the plot $\{(h, \hat{\rho}(h)) \colon h = 0, 1, 2, \dots\}$, which is designed to facilitate the interpretation of the sample ACF. Correlograms for various simulated ARMA processes are shown in Figure 4.1; note that the estimated correlations correspond reasonably closely to the theoretical ACF for these particular realizations.

To interpret such estimators of serial correlation, we need to know something about their behaviour for particular time series. The following general result is for *causal linear processes*, which are processes of the form (4.2) driven by *strict white noise*.

Theorem 4.13. *Let $(X_t)_{t\in\mathbb{Z}}$ be the linear process*

$$X_t - \mu = \sum_{i=0}^{\infty} \psi_i Z_{t-i}, \quad \text{where } \sum_{i=0}^{\infty} |\psi_i| < \infty, \ (Z_t)_{t\in\mathbb{Z}} \sim \mathrm{SWN}(0, \sigma_Z^2).$$

Suppose that either $E(Z_t^4) < \infty$ or $\sum_{i=0}^{\infty} i\psi_i^2 < \infty$. Then, for $h \in \{1, 2, \dots\}$, we have

$$\sqrt{n}(\hat{\boldsymbol{\rho}}(h) - \boldsymbol{\rho}(h)) \xrightarrow{\mathrm{d}} N_h(\mathbf{0}, W),$$

where

$$\hat{\boldsymbol{\rho}}(h) = (\hat{\rho}(1), \dots, \hat{\rho}(h))',$$
$$\boldsymbol{\rho}(h) = (\rho(1), \dots, \rho(h))',$$

N_h denotes an h-dimensional normal distribution (see Section 6.1.3), $\mathbf{0}$ is the h-dimensional vector of zeros, and W is a covariance matrix with elements

$$W_{ij} = \sum_{k=1}^{\infty} (\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k))(\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k)).$$

Proof. This follows as a special case of a result in Brockwell and Davis (1991, pp. 221–223). □

The condition $\sum_{i=0}^{\infty} i\psi_i^2 < \infty$ holds for ARMA processes, so ARMA processes driven by SWN fall under the scope of this theorem (regardless of whether fourth moments exist for the innovations).

Trivially, the theorem also applies to SWN itself. For SWN we have

$$\sqrt{n}\hat{\boldsymbol{\rho}}(h) \xrightarrow{\mathrm{d}} N_h(\mathbf{0}, I_h),$$

where I_h denotes the $h \times h$ identity matrix, so for sufficiently large n the sample autocorrelations of data from an SWN process will behave like iid normal observations with mean 0 and variance $1/n$. Ninety-five per cent of the estimated correlations should lie in the interval $(-1.96/\sqrt{n}, 1.96/\sqrt{n})$, and it is for this reason that correlograms are drawn with confidence bands at these values. If more than 5% of estimated correlations lie outside these bounds, then this is considered as evidence against the null hypothesis that the data are strict white noise.

Remark 4.14. In light of the discussion of the asymptotic behaviour of sample autocorrelations for SWN, it might be asked how these estimators behave for white noise. However, this is an extremely general question because white noise encompasses a variety of possible underlying processes (including the standard ARCH and GARCH processes we later address) that only share second-order properties (finiteness of variance and lack of serial correlation). In some cases the standard

Gaussian confidence bands apply; in some cases they do not. For a GARCH process the critical issue turns out to be the heaviness of the tail of the stationary distribution (see Mikosch and Stărică (2000) for more details).

Portmanteau tests. It is often useful to combine the visual analysis of the correlogram with a formal numerical test of the strict white noise hypothesis, and a popular test is that of Ljung and Box, as applied in Section 3.1.1. Under the null hypothesis of SWN, the statistic

$$Q_{\mathrm{LB}} = n(n+2)\sum_{j=1}^{h}\frac{\hat{\rho}(j)^2}{n-j}$$

has an asymptotic chi-squared distribution with h degrees of freedom. This statistic is generally preferred to the simpler Box–Pierce statistic $Q_{\mathrm{BP}} = n\sum_{j=1}^{h}\hat{\rho}(j)^2$, which also has an asymptotic χ_h^2 distribution under the null hypothesis, although the chi-squared approximation may not be so good in smaller samples. These tests are the most commonly applied portmanteau tests.

If a series of rvs forms an SWN process, then the series of absolute or squared variables must also be iid. It is a good idea to also apply the correlogram and Ljung–Box tests to absolute values as a further test of the SWN hypothesis. We prefer to perform tests of the SWN hypothesis on the absolute values rather than the squared values because the squared series is only an SWN (according to the definition we use) when the underlying series has a finite fourth moment. Daily log-return data often point to models with an infinite fourth moment.

4.1.4 Statistical Analysis of Time Series

In practice, the statistical analysis of time-series data $X_1, \dots, X_n$ follows a programme consisting of the following stages.

Preliminary analysis. The data are plotted and the plausibility of a single stationary model is considered. There are also a number of formal numerical tests of stationarity that may be carried out at this point (see Notes and Comments for details).

Since we concentrate here on differenced logarithmic value series, we will assume that at most minor preliminary manipulation of our data is required. Classical time-series analysis has many techniques for removing *trends and seasonalities* from "non-stationary" data; these techniques are discussed in all standard texts, including Brockwell and Davis (2002) and Chatfield (2003). While certain kinds of financial time series, such as earnings time series, certainly do show seasonal patterns, we will assume that such effects are relatively minor in the kinds of daily or weekly return series that are the basis of risk-management methods. If we were to base our risk management on high-frequency data, preliminary cleaning would be more of an issue, since these show clear diurnal cycles and other deterministic features (see Dacorogna et al. 2001).

Obviously, the assumption of stationarity becomes more questionable if we take long data windows, or if we choose windows in which well-known economic policy shifts have taken place. Although the markets change constantly there will always be a tension between our desire to use the most up-to-date data and our need to include enough data to have precision in statistical estimation. Whether half a year of data, one year, five years or ten years are appropriate will depend on the situation. It is certainly a good idea to perform a number of analyses with different data windows and to investigate the sensitivity of statistical inference to the amount of data.

Analysis in the time domain. Having settled on the data, the techniques of Section 4.1.3 come into play. By applying correlograms and portmanteau tests such as Ljung–Box to both the raw data and their absolute values, the SWN hypothesis is evaluated. If it cannot be rejected for the data in question, then the formal time-series analysis is over and simple distributional fitting could be used instead of dynamic modelling.

For daily risk-factor return series we expect to quickly reject the SWN hypothesis. Despite the fact that correlograms of the raw data may show little evidence of serial correlation, correlograms of the absolute data are likely to show evidence of strong serial dependence. In other words, the data may support a white noise model but not a strict white noise model. In this case, ARMA modelling is not required, but the volatility models of Section 4.2 may be useful.

If the correlogram does provide evidence of the kind of serial correlation patterns produced by ARMA processes, then we can attempt to fit ARMA processes to data.

Model fitting. A traditional approach to model fitting first attempts to *identify the order* of a suitable ARMA process using the correlogram and a further tool known as the partial correlogram (not described in this book but found in all standard texts). For example, the presence of a *cut-off* at lag q in the correlogram (see Example 4.10) is taken as a diagnostic for pure moving-average behaviour of order q (and similar behaviour in a partial correlogram indicates pure AR behaviour). With modern computing power it is now quite easy to simply fit a variety of MA, AR and ARMA models and to use a model-selection criterion like that of Akaike (described in Section A.3.6) to choose the "best" model. There are also automated model choice procedures such as the method of Tsay and Tiao (1984).

Sometimes there are a priori reasons for expecting certain kinds of model to be most appropriate. For example, suppose we analyse longer-period returns that overlap, as in (3.4). Consider the case where the raw data are daily returns and we build weekly returns. In (3.4) we set $h = 5$ (to get weekly returns) and $k = 1$ (to get as much data as possible). Assuming that the underlying data are genuinely from a white noise process $(X_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$, the weekly aggregated returns at times t and $t + l$ satisfy

$$\operatorname{cov}(X_t^{(5)}, X_{t+l}^{(5)}) = \operatorname{cov}\left(\sum_{i=0}^{4} X_{t-i}, \sum_{j=0}^{4} X_{t+l-j}\right) = \begin{cases} (5-l)\sigma^2, & l = 0, \dots, 4, \\ 0, & l \geqslant 5, \end{cases}$$

so that the overlapping returns have the correlation structure of an MA(4) process, and this would be a natural choice of time-series model for them.

Having chosen the model to fit, there are a number of possible fitting methods, including specialized methods for AR processes, such as Yule–Walker, that make minimal assumptions concerning the distribution of the white noise innovations; we refer to the standard time-series literature for more details. In Section 4.2.4 we discuss the method of (conditional) maximum likelihood, which may be used to fit ARMA models with (or without) GARCH errors to data.

Residual analysis and model comparison. Recall the representation of a causal and invertible ARMA process in (4.11), and suppose we have fitted such a process and estimated the parameters ϕ_i and θ_j. The residuals are inferred values $\hat{\varepsilon}_t$ for the unobserved innovations ε_t and they are calculated recursively from the data and the fitted model using the equations

$$\hat{\varepsilon}_t = X_t - \hat{\mu}_t, \quad \hat{\mu}_t = \hat{\mu} + \sum_{i=1}^{p} \hat{\phi}_i (X_{t-i} - \hat{\mu}) + \sum_{j=1}^{q} \hat{\theta}_j \hat{\varepsilon}_{t-j}, \tag{4.13}$$

where the values $\hat{\mu}_t$ are sometimes known as the *fitted values*. Obviously, we have a problem calculating the first few values of $\hat{\varepsilon}_t$ due to the finiteness of our data sample and the infinite nature of the recursions (4.13). One of many possible solutions is to set $\hat{\varepsilon}_{-q+1} = \hat{\varepsilon}_{-q+2} = \cdots = \hat{\varepsilon}_0 = 0$ and $X_{-p+1} = X_{-p+2} = \cdots = X_0 = \bar{X}$ and then to use (4.13) for $t = 1, \dots, n$. Since the first few values will be influenced by these starting values, they might be ignored in later analyses.

The residuals $(\hat{\varepsilon}_t)$ should behave like a realization of a white noise process, since this is our model assumption for the innovations, and this can be assessed by constructing their correlogram. If there is still evidence of serial correlation in the correlogram, then this suggests that a good ARMA model has not yet been found. Moreover, we can use portmanteau tests to test formally that the residuals behave like a realization of a strict white noise process. If the residuals behave like SWN, then no further time-series modelling is required; if they behave like WN but not SWN, then the volatility models of Section 4.2 may be required.

It is usually possible to find more than one reasonable ARMA model for the data, and formal model-comparison techniques may be required to decide on an overall best model or models. The Akaike information criterion described in Section A.3.6 might be used, or one of a number of variants on this criterion that are often preferred for time series (see Brockwell and Davis 2002, Section 5.5.2).

4.1.5 Prediction

There are many approaches to the forecasting or prediction of time series, and we summarize two that extend easily to the case of GARCH models. The first strategy makes use of fitted ARMA (or ARIMA) models and is sometimes called the *Box–Jenkins approach* (Box and Jenkins 1970). The second strategy is a model-free approach to forecasting known as *exponential smoothing*, which is related to the exponentially weighted moving-average technique for predicting volatility.

Prediction using ARMA models. Consider the invertible ARMA model and its representation in (4.11). Let $\mathcal{F}_t$ denote the history of the process up to and including time t, as before, and assume that the innovations $(\varepsilon_t)_{t\in\mathbb{Z}}$ have the martingale-difference property with respect to $(\mathcal{F}_t)_{t\in\mathbb{Z}}$.

For the prediction problem it will be convenient to denote our sample of n data by $X_{t-n+1}, \dots, X_t$. We assume that these are realizations of rvs following a particular ARMA model. Our aim is to predict X_{t+1} or, more generally, X_{t+h}, and we denote our prediction by $P_t X_{t+h}$. The method we describe assumes that we have access to the infinite history of the process up to time t and derives a formula that is then approximated for our finite sample.

As a predictor of X_{t+h} we use the conditional expectation $E(X_{t+h} \mid \mathcal{F}_t)$. Among all predictions $P_t X_{t+h}$ based on the infinite history of the process up to time t, this predictor minimizes the mean squared prediction error $E((X_{t+h} - P_t X_{t+h})^2)$.

The basic idea is that, for $h \geqslant 1$, the prediction $E(X_{t+h} \mid \mathcal{F}_t)$ is recursively evaluated in terms of $E(X_{t+h-1} \mid \mathcal{F}_t)$. We use the fact that $E(\varepsilon_{t+h} \mid \mathcal{F}_t) = 0$ (the martingale-difference property of innovations) and that the rvs $(X_s)_{s\leqslant t}$ and $(\varepsilon_s)_{s\leqslant t}$ are "known" at time t. The assumption of invertibility (4.10) ensures that the innovation ε_t can be written as a function of the infinite history of the process $(X_s)_{s\leqslant t}$. To illustrate the approach it will suffice to consider an ARMA$(1, 1)$ model, the generalization to ARMA(p, q) models following easily.

Example 4.15 (prediction for the ARMA(1, 1) model). Suppose an ARMA$(1, 1)$ model of the form (4.11) has been fitted to the data, and its parameters μ, ϕ_1 and θ_1 have been determined. Our one-step prediction for X_{t+1} is

$$E(X_{t+1} \mid \mathcal{F}_t) = \mu_{t+1} = \mu + \phi_1(X_t - \mu) + \theta_1\varepsilon_t,$$

since $E(\varepsilon_{t+1} \mid \mathcal{F}_t) = 0$. For a two-step prediction we get

$$\begin{aligned} E(X_{t+2} \mid \mathcal{F}_t) = E(\mu_{t+2} \mid \mathcal{F}_t) &= \mu + \phi_1(E(X_{t+1} \mid \mathcal{F}_t) - \mu) \\ &= \mu + \phi_1^2(X_t - \mu) + \phi_1\theta_1\varepsilon_t, \end{aligned}$$

and in general we have

$$E(X_{t+h} \mid \mathcal{F}_t) = \mu + \phi_1^h(X_t - \mu) + \phi_1^{h-1}\theta_1\varepsilon_t.$$

Without knowing all historical values of $(X_s)_{s\leqslant t}$ this predictor cannot be evaluated exactly, because we do not know ε_t exactly, but it can be accurately approximated if n is reasonably large. The easiest way of doing this is to substitute the model residual $\hat{\varepsilon}_t$ calculated from (4.13) for ε_t. Note that $\lim_{h\to\infty} E(X_{t+h} \mid \mathcal{F}_t) = \mu$, almost surely, so that the prediction converges to the estimate of the unconditional mean of the process for longer time horizons.

Exponential smoothing. This is a popular technique that is used for both prediction of time-series and trend estimation. Here we do not necessarily assume that the data come from a stationary model, although we do assume that there is no deterministic seasonal component in the model. In general, the method is less well suited to return series with frequently changing signs and is better suited to undifferenced price or value series. It forms the basis of a very common method of volatility prediction (see Section 4.2.5).

Suppose our data represent realizations of rvs $Y_{t-n+1}, \dots, Y_t$, considered without reference to any concrete parametric model. As a forecast for Y_{t+1} we use a prediction of the form

$$P_t Y_{t+1} = \sum_{i=0}^{n-1} \lambda(1-\lambda)^i Y_{t-i}, \quad \text{where } 0 < \lambda < 1.$$

Thus we weight the data from most recent to most distant with a sequence of exponentially decreasing weights that sum to almost one. It is easily calculated that

$$\begin{aligned} P_t Y_{t+1} = \sum_{i=0}^{n-1} \lambda(1-\lambda)^i Y_{t-i} &= \lambda Y_t + (1-\lambda) \sum_{j=0}^{n-2} \lambda(1-\lambda)^j Y_{t-1-j} \\ &= \lambda Y_t + (1-\lambda) P_{t-1} Y_t, \end{aligned} \tag{4.14}$$

so that the prediction at time t is obtained from the prediction at time $t-1$ by a simple recursive scheme. The choice of λ is subjective; the larger the value, the more weight is put on the most recent observation. Empirical validation studies with different data sets can be used to determine a value of λ that gives good results; Chatfield (2003) reports that values between 0.1 and 0.3 are commonly used in practice.

Note that, although the method is commonly seen as a model-free forecasting technique, it can be shown to be the natural prediction method based on conditional expectation for a non-stationary ARIMA(0, 1, 1) model.

Notes and Comments

There are many texts covering the subject of classical time-series analysis, including Box and Jenkins (1970), Priestley (1981), Abraham and Ledolter (1983), Brockwell and Davis (1991, 2002), Hamilton (1994) and Chatfield (2003). Our account of basic concepts, ARMA models and analysis in the time domain closely follows Brockwell and Davis (1991), which should be consulted for the rigorous background to ideas we can only summarize. We have not discussed analysis of time series in the frequency domain, which is less common for financial time series; for this subject see, again, Brockwell and Davis (1991) or Priestley (1981).

For more on tests of the strict white noise hypothesis (that is, tests of randomness), see Brockwell and Davis (2002). Original references for the Box–Pierce and Ljung–Box tests are Box and Pierce (1970) and Ljung and Box (1978).

There is a large econometrics literature on tests of stationarity and unit-root tests, where the latter are effectively tests of the null hypothesis of non-stationary random-walk behaviour. Particular examples are the Dickey–Fuller and Phillips–Perron unit-root tests (Dickey and Fuller 1979; Phillips and Perron 1988) and the KPSS test of stationarity (Kwiatkowski et al. 1992).

There is a vast literature on forecasting and prediction in linear models. A good non-mathematical introduction is found in Chatfield (2003). The approach we describe based on the infinite history of the time series is discussed in greater detail in Hamilton (1994). Brockwell and Davis (2002) concentrate on exact linear prediction methods for finite samples. A general review of exponential smoothing is found in Gardner (1985).

4.2 GARCH Models for Changing Volatility

The most important models for daily risk-factor return series are addressed in this section. We give definitions of ARCH (autoregressive conditionally heteroscedastic) and GARCH (generalized ARCH) models and discuss some of their mathematical properties before going on to talk about their use in practice.

4.2.1 ARCH Processes

Definition 4.16. Let $(Z_t)_{t\in\mathbb{Z}}$ be SWN(0, 1). The process $(X_t)_{t\in\mathbb{Z}}$ is an ARCH(p) process if it is strictly stationary and if it satisfies, for all $t \in \mathbb{Z}$ and some strictly positive-valued process $(\sigma_t)_{t\in\mathbb{Z}}$, the equations

$$X_t = \sigma_t Z_t, \tag{4.15}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2, \tag{4.16}$$

where $\alpha_0 > 0$ and $\alpha_i \geqslant 0$, $i = 1, \dots, p$.

Let $\mathcal{F}_t = \sigma(\{X_s : s \leqslant t\})$ again denote the sigma algebra representing the history of the process up to time t, so that $(\mathcal{F}_t)_{t\in\mathbb{Z}}$ is the natural filtration. The construction (4.16) ensures that σ_t is *measurable* with respect to $\mathcal{F}_{t-1}$, and the process $(\sigma_t)_{t\in\mathbb{Z}}$ is said to be previsible. This allows us to calculate that, provided $E(|X_t|) < \infty$,

$$E(X_t \mid \mathcal{F}_{t-1}) = E(\sigma_t Z_t \mid \mathcal{F}_{t-1}) = \sigma_t E(Z_t \mid \mathcal{F}_{t-1}) = \sigma_t E(Z_t) = 0, \tag{4.17}$$

so that the ARCH process has the martingale-difference property with respect to $(\mathcal{F}_t)_{t\in\mathbb{Z}}$. If the process is covariance stationary, it is simply a white noise, as discussed in Section 4.1.1.

Remark 4.17. Note that the independence of Z_t and $\mathcal{F}_{t-1}$ that we have assumed above follows from the fact that an ARCH process must be causal, i.e. the equations (4.15) and (4.16) must have a solution of the form $X_t = f(Z_t, Z_{t-1}, \dots)$ for some f, so that Z_t is independent of previous values of the process. This contrasts with ARMA models, where the equations can have non-causal solutions (see Brockwell and Davis 1991, Example 3.1.2).

If we simply assume that the process is a covariance-stationary white noise (for which we will give a condition in Proposition 4.18), then $E(X_t^2) < \infty$ and

$$\operatorname{var}(X_t \mid \mathcal{F}_{t-1}) = E(\sigma_t^2 Z_t^2 \mid \mathcal{F}_{t-1}) = \sigma_t^2 \operatorname{var}(Z_t) = \sigma_t^2.$$

Thus the model has the interesting property that its conditional standard deviation σ_t, or *volatility*, is a continually changing function of the previous squared values of the process. If one or more of $|X_{t-1}|, \dots, |X_{t-p}|$ are particularly *large*, then X_t is effectively drawn from a distribution with large variance, and may itself be large; in

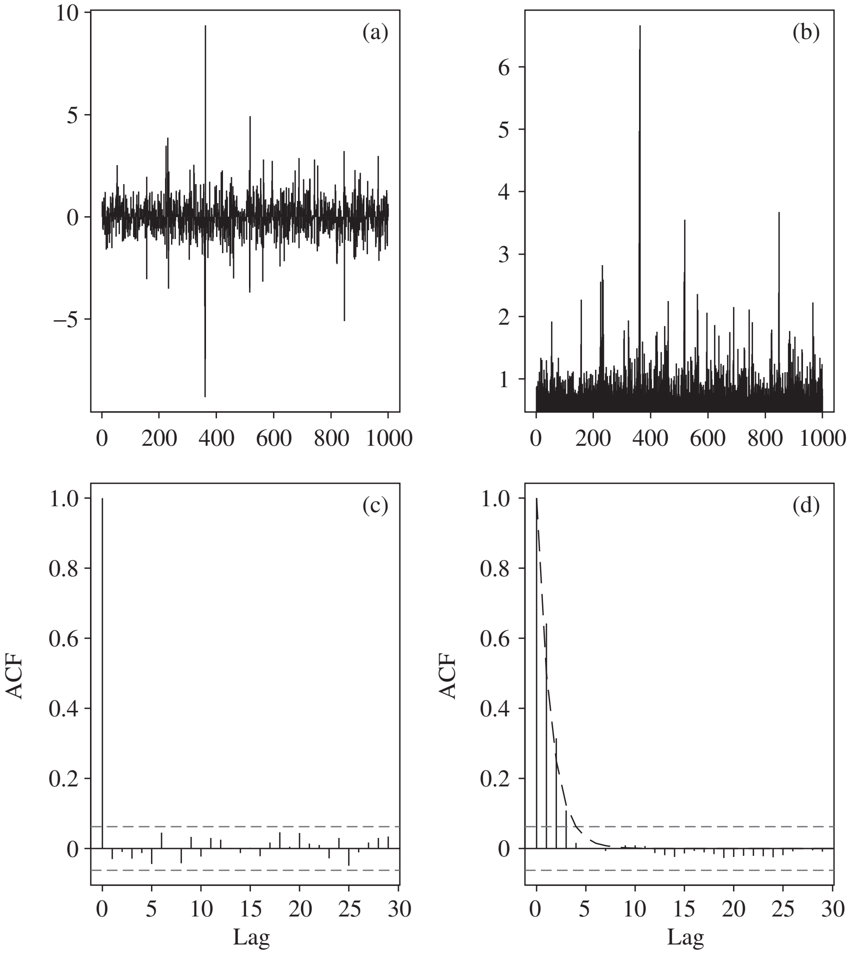

Figure 4.2. A simulated ARCH(1) process with Gaussian innovations and parameters $\alpha_0 = \alpha_1 = 0.5$: (a) the realization of the process; (b) the realization of the volatility; and correlograms of (c) the raw and (d) the squared values. The process is covariance stationary with unit variance and a finite fourth moment (since $\alpha_1 < 1/\sqrt{3}$) and the squared values follow an AR(1) process. The true form of the ACF of the squared values is represented by the dashed line in the correlogram.

this way the model generates volatility clusters. The name ARCH refers to this structure: the model is autoregressive, since X_t clearly depends on previous X_{t-i}, and conditionally heteroscedastic, since the conditional variance changes continually.

The distribution of the innovations $(Z_t)_{t\in\mathbb{Z}}$ can in principle be any zero-mean, unit-variance distribution. For statistical fitting purposes we may or may not choose to actually specify the distribution, depending on whether we implement a maximum likelihood (ML), quasi-maximum likelihood (QML) or non-parametric fitting method (see Section 4.2.4). For ML the most common choices are standard normal innovations or scaled t innovations. By the latter we mean that $Z_t \sim t_1(\nu, 0, (\nu - 2)/\nu)$, in the notation of Example 6.7, so that the variance of the distribution is one. We keep these choices in mind when discussing further theoretical properties of ARCH and GARCH models.

The ARCH(1) *model.* In the rest of this section we analyse some of the properties of the ARCH(1) model. These properties extend to the whole class of ARCH and GARCH models but are most easily introduced in the simplest case. A simulated realization of an ARCH(1) process with Gaussian innovations and the corresponding realization of the volatility process are shown in Figure 4.2.

Using $X_t^2 = \sigma_t^2 Z_t^2$ and (4.16) in the case $p = 1$, we deduce that the squared ARCH(1) process satisfies

$$X_t^2 = \alpha_0 Z_t^2 + \alpha_1 Z_t^2 X_{t-1}^2. \tag{4.18}$$

A detailed mathematical analysis of the ARCH(1) model involves the study of equation (4.18), which is a *stochastic recurrence equation* (SRE). Much as for the AR(1) model in Example 4.11, we would like to know when this equation has stationary solutions expressed in terms of the infinite history of the innovations, i.e. solutions of the form $X_t^2 = f(Z_t, Z_{t-1}, \dots)$.

For ARCH models we have to distinguish carefully between solutions that are covariance stationary and solutions that are only strictly stationary. It is possible to have ARCH(1) models with infinite variance, which obviously cannot be covariance stationary.

Stochastic recurrence relations. The detailed theory required to analyse stochastic recurrence relations of the form (4.18) is outside the scope of this book, and we give only brief notes to indicate the ideas involved. Our treatment is based on Brandt (1986), Mikosch (2003) and Mikosch (2013); see Notes and Comments at the end of this section for further references.

Equation (4.18) is a particular example of a class of recurrence equations of the form

$$Y_t = A_t Y_{t-1} + B_t, \tag{4.19}$$

where $(A_t)_{t\in\mathbb{Z}}$ and $(B_t)_{t\in\mathbb{Z}}$ are sequences of iid rvs. Sufficient conditions for a solution are that

$$E(\ln^+ |B_t|) < \infty \quad \text{and} \quad E(\ln |A_t|) < 0, \tag{4.20}$$

where $\ln^+ x = \max(0, \ln x)$. The unique solution is given by

$$Y_t = B_t + \sum_{i=1}^{\infty} B_{t-i} \prod_{j=0}^{i-1} A_{t-j}, \tag{4.21}$$

where the sum converges absolutely, almost surely.

We can develop some intuition for the conditions (4.20) and the form of the solution (4.21) by iterating equation (4.19) k times to obtain

$$\begin{aligned} Y_t &= A_t(A_{t-1}Y_{t-2} + B_{t-1}) + B_t \\ &= B_t + \sum_{i=1}^{k} B_{t-i} \prod_{j=0}^{i-1} A_{t-j} + Y_{t-k-1} \prod_{i=0}^{k} A_{t-i}. \end{aligned}$$

The conditions (4.20) ensure that the middle term on the right-hand side converges absolutely and the final term disappears. In particular, note that

$$\frac{1}{k+1}\sum_{i=0}^{k} \ln|A_{t-i}| \xrightarrow{\text{a.s.}} E(\ln|A_t|) < 0$$

by the strong law of large numbers. So

$$\prod_{i=0}^{k} |A_{t-i}| = \exp\left(\sum_{i=0}^{k} \ln|A_{t-i}|\right) \xrightarrow{\text{a.s.}} 0,$$

which shows the importance of the $E(\ln|A_t|) < 0$ condition. The solution (4.21) to the SRE is a strictly stationary process (being a function of iid variables $(A_s, B_s)_{s\leqslant t}$), and the $E(\ln|A_t|) < 0$ condition turns out to be the key to the strict stationarity of ARCH and GARCH models.

Stationarity of ARCH(1). The squared ARCH(1) model (4.18) is an SRE of the form (4.19) with $A_t = \alpha_1 Z_t^2$ and $B_t = \alpha_0 Z_t^2$. Thus the conditions in (4.20) translate into the requirements that $E(\ln^+|\alpha_0 Z_t^2|) < \infty$, which is automatically true for the ARCH(1) process as we have defined it, and $E(\ln(\alpha_1 Z_t^2)) < 0$. This is the condition for a strictly stationary solution of the ARCH(1) equations, and it can be shown that it is in fact a necessary and sufficient condition for strict stationarity (see Bougerol and Picard 1992). From (4.21), the solution of equation (4.18) takes the form

$$X_t^2 = \alpha_0 \sum_{i=0}^{\infty} \alpha_1^i \prod_{j=0}^{i} Z_{t-j}^2. \tag{4.22}$$

If the (Z_t) are standard normal innovations, then the condition for a strictly stationary solution is approximately $\alpha_1 < 3.562$; perhaps somewhat surprisingly, if the (Z_t) are scaled t innovations with four degrees of freedom and variance 1, the condition is $\alpha_1 < 5.437$. Strict stationarity depends on the distribution of the innovations but covariance stationarity does not; the necessary and sufficient condition for covariance stationarity is always $\alpha_1 < 1$, as we now prove.

Proposition 4.18. *The* ARCH(1) *process is a covariance-stationary white noise process if and only if* $\alpha_1 < 1$. *The variance of the covariance-stationary process is given by* $\alpha_0/(1-\alpha_1)$.

Proof. Assuming covariance stationarity, it follows from (4.18) and $E(Z_t^2) = 1$ that

$$\sigma_x^2 = E(X_t^2) = \alpha_0 + \alpha_1 E(X_{t-1}^2) = \alpha_0 + \alpha_1 \sigma_x^2.$$

Clearly, $\sigma_x^2 = \alpha_0/(1-\alpha_1)$ and we must have $\alpha_1 < 1$.

Conversely, if $\alpha_1 < 1$, then, by Jensen's inequality,

$$E(\ln(\alpha_1 Z_t^2)) \leqslant \ln(E(\alpha_1 Z_t^2)) = \ln(\alpha_1) < 0,$$

and we can use (4.22) to calculate that

$$E(X_t^2) = \alpha_0 \sum_{i=0}^{\infty} \alpha_1^i = \frac{\alpha_0}{1-\alpha_1}.$$

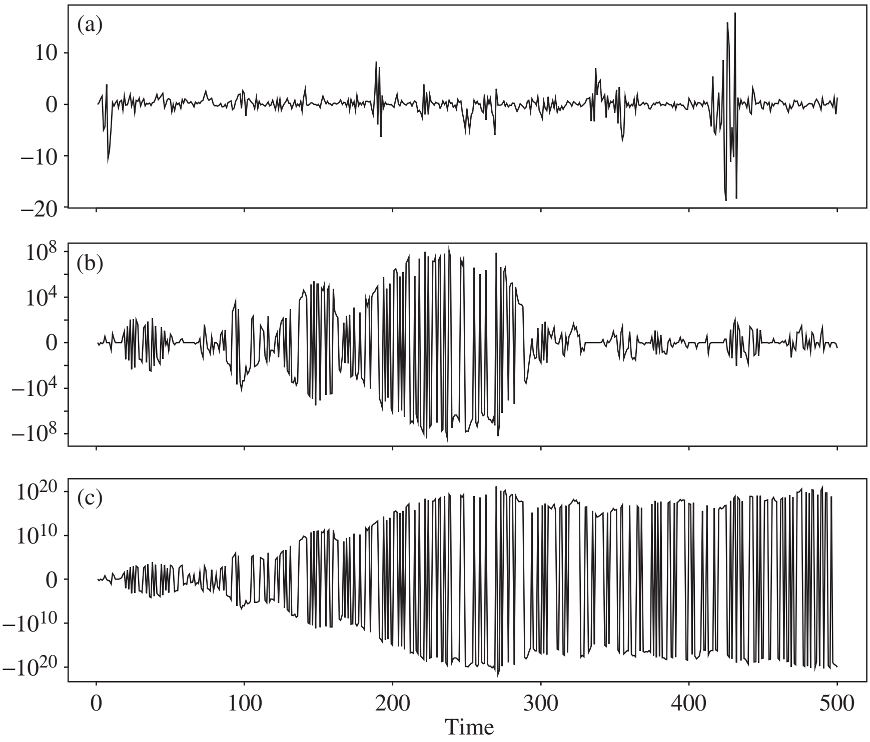

Figure 4.3. (a), (b) Strictly stationary ARCH(1) models with Gaussian innovations that are not covariance stationary ($\alpha_1 = 1.2$ and $\alpha_1 = 3$, respectively). (c) A non-stationary (explosive) process generated by the ARCH(1) equations with $\alpha_1 = 4$. Note that (b) and (c) use a special double-logarithmic y-axis where all values less than one in modulus are plotted at zero.

The process $(X_t)_{t \in \mathbb{Z}}$ is a martingale difference with a finite, non-time-dependent second moment. Hence it is a white noise process. □

See Figure 4.3 for examples of non-covariance-stationary ARCH(1) models as well as an example of a non-stationary (explosive) process generated by the ARCH(1) equations. The process in Figure 4.2 is covariance stationary.

On the stationary distribution of X_t. It is clear from (4.22) that the distribution of the (X_t) in an ARCH(1) model bears a complicated relationship to the distribution of the innovations (Z_t). Even if the innovations are Gaussian, the stationary distribution of the time series is not Gaussian, but rather a leptokurtic distribution with more slowly decaying tails.

Moreover, from (4.15) we see that the distribution of X_t is a normal mixture distribution of the kind discussed in Section 6.2. Its distribution depends on the distribution of σ_t, which has no simple form.

Proposition 4.19. *For $m \geqslant 1$, the strictly stationary* ARCH(1) *process has finite moments of order $2m$ if and only if $E(Z_t^{2m}) < \infty$ and $\alpha_1 < (E(Z_t^{2m}))^{-1/m}$.*

Proof. We rewrite (4.22) in the form $X_t^2 = Z_t^2 \sum_{i=0}^{\infty} Y_{t,i}$ for positive rvs $Y_{t,i} = \alpha_0 \alpha_1^i \prod_{j=1}^{i} Z_{t-j}^2$, $i \geqslant 1$, and $Y_{t,0} = \alpha_0$. For $m \geqslant 1$ the following inequalities hold (the latter being Minkowski's inequality):

$$E(Y_{t,1}^m) + E(Y_{t,2}^m) \leqslant E((Y_{t,1} + Y_{t,2})^m) \leqslant ((E(Y_{t,1}^m))^{1/m} + (E(Y_{t,2}^m))^{1/m})^m.$$

Since

$$E(X_t^{2m}) = E(Z_t^{2m}) E\biggl(\biggl(\sum_{i=0}^{\infty} Y_{t,i}\biggr)^m\biggr),$$

it follows that

$$E(Z_t^{2m}) \sum_{i=0}^{\infty} E(Y_{t,i}^m) \leqslant E(X_t^{2m}) \leqslant E(Z_t^{2m}) \biggl(\sum_{i=0}^{\infty} (E(Y_{t,i}^m))^{1/m}\biggr)^m.$$

Since $E(Y_{t,i}^m) = \alpha_0^m \alpha_1^{im} (E(Z_t^{2m}))^i$, it may be deduced that all three quantities are finite if and only if $E(Z_t^{2m}) < \infty$ and $\alpha_1^m E(Z_t^{2m}) < 1$. □

For example, for a finite fourth moment ($m = 2$) we require $\alpha_1 < 1/\sqrt{3}$ in the case of Gaussian innovations and $\alpha_1 < 1/\sqrt{6}$ in the case of t innovations with six degrees of freedom; for t innovations with four degrees of freedom, the fourth moment is undefined.

Assuming the existence of a finite fourth moment, it is easy to calculate its value, and also that of the kurtosis of the process. We square both sides of (4.18), take expectations of both sides and then solve for $E(X_t^4)$ to obtain

$$E(X_t^4) = \frac{\alpha_0^2 E(Z_t^4)(1 - \alpha_1^2)}{(1 - \alpha_1)^2 (1 - \alpha_1^2 E(Z_t^4))}.$$

The kurtosis of the stationary distribution κ_X can then be calculated to be

$$\kappa_X = \frac{E(X_t^4)}{E(X_t^2)^2} = \frac{\kappa_Z (1 - \alpha_1^2)}{(1 - \alpha_1^2 \kappa_Z)},$$

where $\kappa_Z = E(Z_t^4)$ denotes the kurtosis of the innovations. Clearly, when $\kappa_Z > 1$, the kurtosis of the stationary distribution is inflated in comparison with that of the innovation distribution; for Gaussian or t innovations, $\kappa_X > 3$, so the stationary distribution is leptokurtic. The kurtosis of the process in Figure 4.2 is 9.

Parallels with the AR(1) *process.* We now turn our attention to the serial dependence structure of the squared series in the case of covariance stationarity ($\alpha_1 < 1$). We write the squared process as

$$X_t^2 = \sigma_t^2 Z_t^2 = \sigma_t^2 + \sigma_t^2 (Z_t^2 - 1). \tag{4.23}$$

Setting $V_t = \sigma_t^2 (Z_t^2 - 1)$, we note that $(V_t)_{t \in \mathbb{Z}}$ forms a martingale-difference series, since $E|V_t| < \infty$ and $E(V_t \mid \mathcal{F}_{t-1}) = \sigma_t^2 E(Z_t^2 - 1) = 0$. Now we rewrite (4.23) as $X_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + V_t$, and observe that this closely resembles an AR(1) process for X_t^2, except that V_t is not necessarily a white noise process. If we restrict our attention to processes where $E(X_t^4)$ is finite, then V_t has a finite and constant second moment and is a white noise process. Under this assumption, X_t^2 is an AR(1) process, according to Definition 4.7, of the form

$$\left(X_t^2 - \frac{\alpha_0}{1-\alpha_1}\right) = \alpha_1\left(X_{t-1}^2 - \frac{\alpha_0}{1-\alpha_1}\right) + V_t.$$

It has mean $\alpha_0/(1-\alpha_1)$ and we can use Example 4.11 to conclude that the autocorrelation function is $\rho(h) = \alpha_1^{|h|}$, $h \in \mathbb{Z}$. Figure 4.2 shows an example of an ARCH(1) process with finite fourth moment whose squared values follow an AR(1) process.

4.2.2 GARCH Processes

Definition 4.20. Let $(Z_t)_{t\in\mathbb{Z}}$ be SWN(0, 1). The process $(X_t)_{t\in\mathbb{Z}}$ is a GARCH(p, q) process if it is strictly stationary and if it satisfies, for all $t \in \mathbb{Z}$ and some strictly positive-valued process $(\sigma_t)_{t\in\mathbb{Z}}$, the equations

$$X_t = \sigma_t Z_t, \qquad \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \tag{4.24}$$

where $\alpha_0 > 0$, $\alpha_i \geqslant 0$, $i = 1, \dots, p$, and $\beta_j \geqslant 0$, $j = 1, \dots, q$.

The GARCH processes are *generalized* ARCH processes in the sense that the squared volatility σ_t^2 is allowed to depend on previous squared volatilities, as well as previous squared values of the process.

The GARCH(1, 1) *model.* In practice, low-order GARCH models are most widely used and we will concentrate on the GARCH(1, 1) model. In this model periods of high volatility tend to be *persistent*, since $|X_t|$ has a chance of being large if *either* $|X_{t-1}|$ is large *or* σ_{t-1} is large; the same effect can be achieved in ARCH models of high order, but lower-order GARCH models achieve this effect more parsimoniously. A simulated realization of a GARCH(1, 1) process with Gaussian innovations and its volatility are shown in Figure 4.4; in comparison with the ARCH(1) model of Figure 4.2, it is clear that the volatility persists longer at higher levels before decaying to lower levels.

Stationarity. It follows from (4.24) that for a GARCH(1, 1) model we have

$$\sigma_t^2 = \alpha_0 + (\alpha_1 Z_{t-1}^2 + \beta_1)\sigma_{t-1}^2, \tag{4.25}$$

which is again an SRE of the form $Y_t = A_t Y_{t-1} + B_t$, as in (4.19). This time it is an SRE for $Y_t = \sigma_t^2$ rather than X_t^2, but its analysis follows easily from the ARCH(1) case.

The condition $E(\ln|A_t|) < 0$ for a strictly stationary solution of (4.19) translates to the condition $E(\ln(\alpha_1 Z_t^2 + \beta_1)) < 0$ for (4.25), and the general solution (4.21) becomes

$$\sigma_t^2 = \alpha_0 + \alpha_0 \sum_{i=1}^{\infty} \prod_{j=1}^{i} (\alpha_1 Z_{t-j}^2 + \beta_1). \tag{4.26}$$

If $(\sigma_t^2)_{t\in\mathbb{Z}}$ is a strictly stationary process, then so is $(X_t)_{t\in\mathbb{Z}}$, since $X_t = \sigma_t Z_t$ and $(Z_t)_{t\in\mathbb{Z}}$ is simply strict white noise. The solution of the GARCH(1, 1) defining equations is then

$$X_t = Z_t \sqrt{\alpha_0\biggl(1 + \sum_{i=1}^{\infty}\prod_{j=1}^{i}(\alpha_1 Z_{t-j}^2 + \beta_1)\biggr)}, \tag{4.27}$$

and we can use this to derive the condition for covariance stationarity.

Proposition 4.21. *The* GARCH(1, 1) *process is a covariance-stationary white noise process if and only if* $\alpha_1 + \beta_1 < 1$. *The variance of the covariance-stationary process is given by* $\alpha_0/(1 - \alpha_1 - \beta_1)$.

Proof. We use a similar argument to Proposition 4.18 and make use of (4.27). □

Fourth moments and kurtosis. Using a similar approach to Proposition 4.19 we can use (4.27) to derive conditions for the existence of higher moments of a covariance-stationary GARCH(1, 1) process. For the existence of a fourth moment, a necessary and sufficient condition is that $E((\alpha_1 Z_t^2 + \beta_1)^2) < 1$, or alternatively that

$$(\alpha_1 + \beta_1)^2 < 1 - (\kappa_Z - 1)\alpha_1^2.$$

Assuming this to be true, we calculate the fourth moment and kurtosis of X_t. We square both sides of (4.25) and take expectations to obtain

$$E(\sigma_t^4) = \alpha_0^2 + (\alpha_1^2\kappa_Z + \beta_1^2 + 2\alpha_1\beta_1)E(\sigma_t^4) + 2\alpha_0(\alpha_1 + \beta_1)E(\sigma_t^2).$$

Solving for $E(\sigma_t^4)$, recalling that $E(\sigma_t^2) = E(X_t^2) = \alpha_0/(1 - \alpha_1 - \beta_1)$, and setting $E(X_t^4) = \kappa_Z E(\sigma_t^4)$, we obtain

$$E(X_t^4) = \frac{\alpha_0^2\kappa_Z(1 - (\alpha_1 + \beta_1)^2)}{(1 - \alpha_1 - \beta_1)^2(1 - \alpha_1^2\kappa_Z - \beta_1^2 - 2\alpha_1\beta_1)},$$

from which it follows that

$$\kappa_X = \frac{\kappa_Z(1 - (\alpha_1 + \beta_1)^2)}{(1 - (\alpha_1 + \beta_1)^2 - (\kappa_Z - 1)\alpha_1^2)}.$$

Again it is clear that the kurtosis of X_t is greater than that of Z_t whenever $\kappa_Z > 1$, such as for Gaussian and scaled t innovations. The kurtosis of the GARCH(1, 1) model in Figure 4.4 is 3.77.

Parallels with the ARMA(1, 1) *process.* Using the same representation as in equation (4.23), the covariance-stationary GARCH(1, 1) process may be written as

$$X_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1\sigma_{t-1}^2 + V_t,$$

where V_t is a martingale difference, given by $V_t = \sigma_t^2(Z_t^2 - 1)$. Since $\sigma_{t-1}^2 = X_{t-1}^2 - V_{t-1}$, we may write

$$X_t^2 = \alpha_0 + (\alpha_1 + \beta_1)X_{t-1}^2 - \beta_1 V_{t-1} + V_t, \tag{4.28}$$

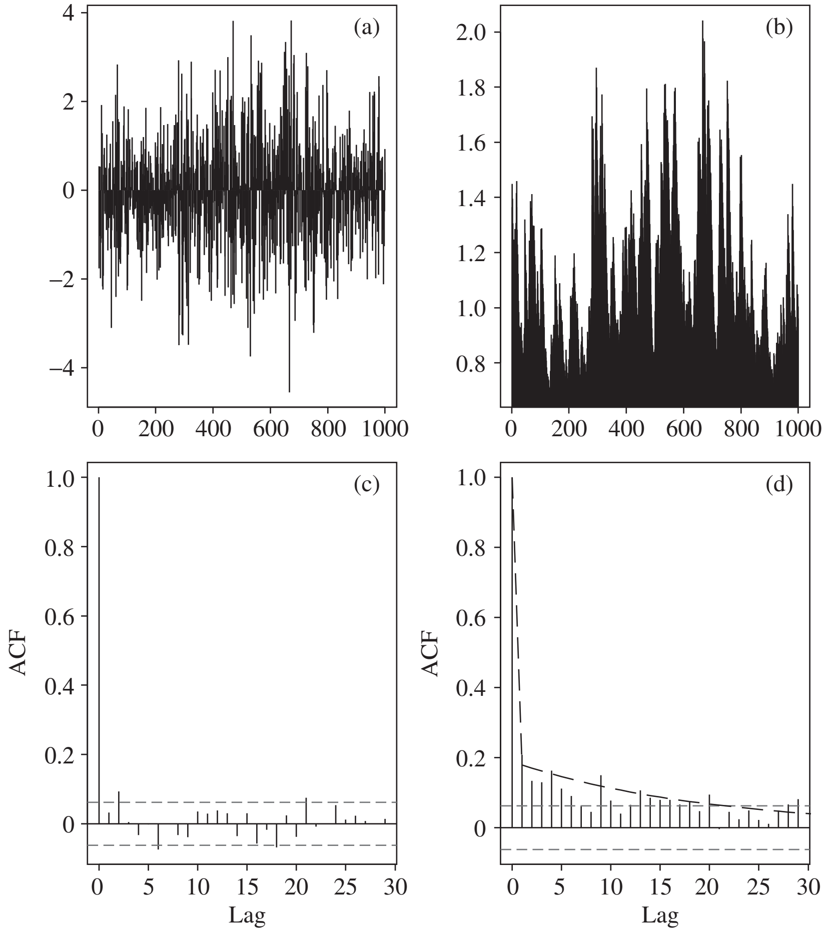

Figure 4.4. A GARCH(1, 1) process with Gaussian innovations and parameters $\alpha_0 = 0.5$, $\alpha_1 = 0.1, \beta_1 = 0.85$: (a) the realization of the process; (b) the realization of the volatility; and correlograms of (c) the raw and (d) the squared values. The process is covariance stationary with unit variance and a finite fourth moment and the squared values follow an ARMA(1, 1) process. The true form of the ACF of the squared values is shown by a dashed line in the correlogram.

which begins to resemble an ARMA(1, 1) process for X_t^2. If we further assume that $E(X_t^4) < \infty$, then, recalling that $\alpha_1 + \beta_1 < 1$, we have formally that

$$\left(X_t^2 - \frac{\alpha_0}{1 - \alpha_1 - \beta_1}\right) = (\alpha_1 + \beta_1)\left(X_{t-1}^2 - \frac{\alpha_0}{1 - \alpha_1 - \beta_1}\right) - \beta_1 V_{t-1} + V_t$$

is an ARMA(1, 1) process. Figure 4.4 shows an example of a GARCH(1, 1) process with finite fourth moment whose squared values follow an ARMA(1, 1) process.

The GARCH(p, q) *model.* Higher-order ARCH and GARCH models have the same general behaviour as ARCH(1) and GARCH(1, 1), but their mathematical analysis becomes more tedious. The condition for a strictly stationary solution of the defining SRE has been derived by Bougerol and Picard (1992), but it is complicated. The necessary and sufficient condition that this solution is covariance stationary is $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$.

A squared GARCH(p, q) process has the structure

$$X_t^2 = \alpha_0 + \sum_{i=1}^{\max(p,q)} (\alpha_i + \beta_i) X_{t-i}^2 - \sum_{j=1}^{q} \beta_j V_{t-j} + V_t,$$

where $\alpha_i = 0$ for $i = p + 1, \dots, q$ if $q > p$, or $\beta_j = 0$ for $j = q + 1, \dots, p$ if $p > q$. This resembles the ARMA$(\max(p, q), q)$ process and is formally such a process provided $E(X_t)^4 < \infty$.

Integrated GARCH. The study of integrated GARCH (or IGARCH) processes has been motivated by the fact that, in some applications of GARCH modelling to daily or higher-frequency risk-factor return series, the estimated ARCH and GARCH coefficients $(\alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q)$ are observed to sum to a number very close to 1, and sometimes even slightly larger than 1. In a model where $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j \geqslant 1$, the process has *infinite variance* and is thus non-covariance-stationary. The special case where $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j = 1$ is known as IGARCH and has received some attention.

For simplicity, consider the IGARCH$(1, 1)$ model. We use (4.28) to conclude that the squared process must satisfy

$$\nabla X_t^2 = X_t^2 - X_{t-1}^2 = \alpha_0 - (1 - \alpha_1) V_{t-1} + V_t,$$

where V_t is a noise sequence defined by $V_t = \sigma_t^2 (Z_t^2 - 1)$ and $\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + (1 - \alpha_1) \sigma_{t-1}^2$. This equation is reminiscent of an ARIMA$(0, 1, 1)$ model (see (4.12)) for X_t^2, although the noise V_t is not white noise, nor is it strictly speaking a martingale difference according to Definition 4.6. $E(V_t \mid \mathcal{F}_{t-1})$ is undefined since $E(\sigma_t^2) = E(X_t^2) = \infty$, and therefore $E|V_t|$ is undefined.

4.2.3 Simple Extensions of the GARCH Model

Many variants and extensions of the basic GARCH model have been proposed. We mention only a few (see Notes and Comments for further reading).

ARMA models with GARCH errors. We have seen that ARMA processes are driven by a white noise $(\varepsilon_t)_{t \in \mathbb{Z}}$ and that a covariance-stationary GARCH process is an example of a white noise. In this section we put the ARMA and GARCH models together by setting the ARMA error ε_t equal to $\sigma_t Z_t$, where σ_t follows a GARCH volatility specification in terms of historical values of ε_t. This gives us a flexible family of ARMA models with GARCH errors that combines the features of both model classes.

Definition 4.22. Let $(Z_t)_{t \in \mathbb{Z}}$ be SWN(0, 1). The process $(X_t)_{t \in \mathbb{Z}}$ is said to be an ARMA(p_1, q_1) process with GARCH(p_2, q_2) errors if it is covariance stationary and satisfies difference equations of the form

$$\begin{aligned}
X_t &= \mu_t + \sigma_t Z_t,\\
\mu_t &= \mu + \sum_{i=1}^{p_1} \phi_i (X_{t-i} - \mu) + \sum_{j=1}^{q_1} \theta_j (X_{t-j} - \mu_{t-j}),\\
\sigma_t^2 &= \alpha_0 + \sum_{i=1}^{p_2} \alpha_i (X_{t-i} - \mu_{t-i})^2 + \sum_{j=1}^{q_2} \beta_j \sigma_{t-j}^2,
\end{aligned}$$

where $\alpha_0 > 0$, $\alpha_i \geqslant 0$, $i = 1, \dots, p_2$, $\beta_j \geqslant 0$, $j = 1, \dots, q_2$, and $\sum_{i=1}^{p_2} \alpha_i + \sum_{j=1}^{q_2} \beta_j < 1$.

To be consistent with the previous definition of an ARMA process we build the covariance-stationarity condition for the GARCH errors into the definition. For the ARMA process to be a causal and invertible linear process, as before, the polynomials $\tilde{\phi}(z) = 1 - \phi_1 z - \cdots - \phi_{p_1} z^{p_1}$ and $\tilde{\theta}(z) = 1 + \theta_1 z + \cdots + \theta_{q_1} z^{q_1}$ should have no common roots and no roots inside the unit circle.

Let $(\mathcal{F}_t)_{t \in \mathbb{Z}}$ denote the natural filtration of $(X_t)_{t \in \mathbb{Z}}$, and assume that the ARMA model is invertible. The invertibility of the ARMA process ensures that μ_t is $\mathcal{F}_{t-1}$-measurable as in (4.11). Moreover, since σ_t depends on the infinite history $(X_s - \mu_s)_{s \leqslant t-1}$, the ARMA invertibility also ensures that σ_t is $\mathcal{F}_{t-1}$-measurable. Simple calculations show that $\mu_t = E(X_t \mid \mathcal{F}_{t-1})$ and $\sigma_t^2 = \operatorname{var}(X_t \mid \mathcal{F}_{t-1})$, so that μ_t and σ_t^2 are the conditional mean and variance of the new process.

GARCH with leverage. One of the main criticisms of the standard ARCH and GARCH models is the rigidly symmetric way in which the volatility reacts to recent returns, regardless of their sign. Economic theory suggests that market information should have an asymmetric effect on volatility, whereby bad news leading to a fall in the equity value of a company tends to increase the volatility. This phenomenon has been called a *leverage effect*, because a fall in equity value causes an increase in the debt-to-equity ratio, or so-called leverage, of a company and should consequently make the stock more volatile. At a less theoretical level it seems reasonable that falling stock values might lead to a higher level of investor nervousness than rises in value of the same magnitude.

One method of adding a leverage effect to a GARCH(1, 1) model is by introducing an additional parameter into the volatility equation (4.24) to get

$$\sigma_t^2 = \alpha_0 + \alpha_1 (X_{t-1} + \delta |X_{t-1}|)^2 + \beta_1 \sigma_{t-1}^2. \tag{4.29}$$

We assume that $\delta \in [-1, 1]$ and $\alpha_1 \geqslant 0$, as in the GARCH(1, 1) model. Observe that (4.29) may be written as

$$\sigma_t^2 = \begin{cases} \alpha_0 + \alpha_1 (1+\delta)^2 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2, & X_{t-1} \geqslant 0,\\ \alpha_0 + \alpha_1 (1-\delta)^2 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2, & X_{t-1} < 0, \end{cases}$$

and hence that

$$\frac{\partial \sigma_t^2}{\partial X_{t-1}^2} = \begin{cases} \alpha_1 (1+\delta)^2 \sigma_{t-1}^2, & X_{t-1} \geqslant 0,\\ \alpha_1 (1-\delta)^2 \sigma_{t-1}^2, & X_{t-1} < 0. \end{cases}$$

The response of volatility to the magnitude of the most recent return depends on the sign of that return, and we generally expect $\delta < 0$, so bad news has the greater effect.

Threshold GARCH. Observe that (4.29) may easily be rewritten in the form

$$\sigma_t^2 = \alpha_0 + \tilde{\alpha}_1 X_{t-1}^2 + \tilde{\delta}_1 I_{\{X_{t-1}<0\}} X_{t-1}^2 + \beta_1 \sigma_{t-1}^2, \tag{4.30}$$

where $\tilde{\alpha}_1 = \alpha_1(1+\delta)^2$ and $\tilde{\delta} = -4\delta\alpha_1$. Equation (4.30) gives the most common version of a threshold GARCH (or TGARCH) model. In effect, a threshold has been set at level zero, and at time t the dynamics depend on whether the previous value of the process X_{t-1} (or innovation Z_{t-1}) was below or above this threshold. However, it is also possible to set non-zero thresholds in TGARCH models, so this represents a more general class of model than GARCH with leverage.

In a less common version of threshold GARCH, the coefficients of the GARCH effects depend on the signs of previous values of the process; this gives a first-order process of the form

$$\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2 + \delta I_{\{X_{t-1}<0\}} \sigma_{t-1}^2. \tag{4.31}$$

Remark 4.23. Note, also, that a further way to introduce asymmetry into a GARCH model is to explicitly use an asymmetric innovation distribution (albeit normalized to have mean 0 and variance 1). Candidate distributions could come from the generalized hyperbolic family of Section 6.2.3.

4.2.4 Fitting GARCH Models to Data

Building the likelihood. In practice, the most widely used approach to fitting GARCH models to data is maximum likelihood. We consider in turn the fitting of the ARCH(1) and GARCH(1, 1) models, from which the fitting of general ARCH(p) and GARCH(p, q) models easily follows.

For the ARCH(1) and GARCH(1, 1) models, suppose we have a total of $n+1$ data values $X_0, X_1, \dots, X_n$. It is useful to recall that we can write the joint density of the corresponding rvs as

$$f_{X_0,\dots,X_n}(x_0,\dots,x_n) = f_{X_0}(x_0)\prod_{t=1}^{n} f_{X_t|X_{t-1},\dots,X_0}(x_t \mid x_{t-1},\dots,x_0). \tag{4.32}$$

For the pure ARCH(1) process, which is first-order Markovian, the conditional densities $f_{X_t|X_{t-1},\dots,X_0}$ in (4.32) depend on the past only through the value of σ_t or, equivalently, X_{t-1}. The conditional density is easily calculated to be

$$f_{X_t|X_{t-1},\dots,X_0}(x_t \mid x_{t-1},\dots,x_0) = f_{X_t|X_{t-1}}(x_t \mid x_{t-1}) = \frac{1}{\sigma_t} f_Z\left(\frac{x_t}{\sigma_t}\right), \tag{4.33}$$

where $\sigma_t = (\alpha_0 + \alpha_1 x_{t-1}^2)^{1/2}$ and $f_Z(z)$ denotes the density of the innovations $(Z_t)_{t\in\mathbb{Z}}$. We recall that this must have mean 0 and variance 1, and typical choices would be the standard normal density or the density of a t distribution scaled to have unit variance.

However, the marginal density f_{X_0} in (4.32) is not known in a tractable closed form for ARCH and GARCH models, and this poses a problem for basing a likelihood on (4.32). The solution employed in practice is to construct the *conditional likelihood* given X_0, which is calculated from

$$f_{X_1,\dots,X_n|X_0}(x_1,\dots,x_n \mid x_0) = \prod_{t=1}^{n} f_{X_t|X_{t-1},\dots,X_0}(x_t \mid x_{t-1},\dots,x_0). \qquad (4.34)$$

For the ARCH(1) model this follows from (4.33) and is

$$L(\alpha_0,\alpha_1;\boldsymbol{X}) = f_{X_1,\dots,X_n|X_0}(X_1,\dots,X_n \mid X_0) = \prod_{t=1}^{n} \frac{1}{\sigma_t} f_Z\biggl(\frac{X_t}{\sigma_t}\biggr),$$

with $\sigma_t = (\alpha_0 + \alpha_1 X_{t-1}^2)^{1/2}$. For an ARCH($p$) model we would use analogous arguments to write down a likelihood conditional on the first p values.

In the GARCH(1, 1) model, σ_t is recursively defined in terms of σ_{t-1}, and here, instead of using (4.34), we construct the joint density of $X_1,\dots,X_n$ conditional on realized values of both X_0 and σ_0, which is

$$f_{X_1,\dots,X_n|X_0,\sigma_0}(x_1,\dots,x_n \mid x_0,\sigma_0) = \prod_{t=1}^{n} f_{X_t|X_{t-1},\dots,X_0,\sigma_0}(x_t \mid x_{t-1},\dots,x_0,\sigma_0).$$

The conditional densities $f_{X_t|X_{t-1},\dots,X_0,\sigma_0}$ depend on the past only through the value of σ_t, which is given recursively from $\sigma_0, X_0,\dots,X_{t-1}$ using $\sigma_t^2 = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2$. This gives us the conditional likelihood

$$L(\alpha_0,\alpha_1,\beta_1;\boldsymbol{X}) = \prod_{t=1}^{n} \frac{1}{\sigma_t} f_Z\biggl(\frac{X_t}{\sigma_t}\biggr), \quad \sigma_t = \sqrt{\alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 \sigma_{t-1}^2}.$$

The problem remains that the value of σ_0^2 is not actually observed, and this is usually solved by choosing a starting value, such as the sample variance of $X_1,\dots,X_n$, or even simply zero.

For a GARCH(p,q) model, we would assume that we had $n+p$ data values labelled $X_{-p+1},\dots,X_0,X_1,\dots,X_n$. We would evaluate the likelihood conditional on the (observed) values of $X_{-p+1},\dots,X_0$ as well as the (unobserved) values of $\sigma_{-q+1},\dots,\sigma_0$, for which starting values would be used as above. For example, if $p=1$ and $q=3$, we require starting values for σ_0, σ_{-1} and σ_{-2}.

A similar approach can be used to develop a likelihood for an ARMA model with GARCH errors. In this case we would end up with a conditional likelihood of the form

$$L(\boldsymbol{\theta};\boldsymbol{X}) = \prod_{t=1}^{n} \frac{1}{\sigma_t} f_Z\biggl(\frac{X_t-\mu_t}{\sigma_t}\biggr),$$

where σ_t follows a GARCH specification and μ_t follows an ARMA specification as in Definition 4.22, and all unknown parameters (possibly including unknown parameters of the innovation distribution) have been collected in the vector $\boldsymbol{\theta}$. We could of course also consider models with leverage or threshold effects.

Deriving parameter estimates. Consider, then, a log-likelihood of the form

$$\ln L(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta}), \tag{4.35}$$

where l_t denotes the log-likelihood contribution arising from the tth observation. The maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ maximizes the (conditional) log-likelihood in (4.35) and, being in general a local maximum, solves the likelihood equations

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{t=1}^{n} \frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \tag{4.36}$$

where the left-hand side is also known as the *score vector* of the conditional likelihood. The equations (4.36) are usually solved numerically using so-called modified Newton–Raphson procedures. A particular method that is widely used for GARCH models is the BHHH method of Berndt et al. (1974).

In describing the behaviour of parameter estimates in the following paragraphs, we distinguish two situations. In the first situation we assume that the model that has been fitted has been *correctly specified*, so that the data are truly generated by a time-series model with both the assumed dynamic form and innovation distribution. We describe the asymptotic behaviour of the maximum likelihood estimates (MLEs) under this idealization.

In the second situation we assume that the correct dynamic form is fitted but that the innovations are erroneously assumed to be Gaussian. Under this misspecification, the model fitting procedure is known as *quasi-maximum likelihood* (QML) and the estimates obtained are QMLEs. Essentially, the Gaussian likelihood is treated as an objective function to be maximized rather than a proper likelihood; our intuition suggests that this may still give reasonable parameter estimates, and this turns out to be the case under appropriate assumptions about the true innovation distribution.

Properties of MLEs. It helps to recall at this point the asymptotic distribution theory for MLEs in the classical iid case, which is summarized in Section A.3. The asymptotic results we give for GARCH models have a similar form to the results in the iid case, but it is important to realize that this is not simply an application of these results. The asymptotics have been separately and laboriously derived in a series of papers for which starting references are given in Notes and Comments. We will give results for pure GARCH models without ARMA components or additional leverage structure, which have been studied rigorously, but the form of the results will apply more generally.

For a pure GARCH(p, q) model with Gaussian innovations it can be shown that (assuming the model has been correctly specified)

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\mathrm{d}} N_{p+q+1}(\mathbf{0}, I(\boldsymbol{\theta})^{-1}),$$

where

$$I(\boldsymbol{\theta}) = E\left(\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right) = -E\left(\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right) \tag{4.37}$$

is the Fisher information matrix arising from any single observation. Thus we have consistent and asymptotically normal estimates of the GARCH parameters. In practice, the *expected* information matrix $I(\boldsymbol{\theta})$ is approximated by an *observed* information matrix, and here we could take the observed information matrix coming from either of the equivalent forms for the expected information matrix in (4.37). That is, we could use

$$\bar{I}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right) \quad \text{or} \quad \bar{J}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}, \tag{4.38}$$

where the first matrix is said to have *outer-product* form and the second is said to have *Hessian* form. These matrices are estimated by evaluating them at the MLEs to get $\bar{I}(\hat{\boldsymbol{\theta}})$ or $\bar{J}(\hat{\boldsymbol{\theta}})$. In practice, the derivatives of the log-likelihood at the MLE are often approximated using first- and second-order differences.

If the model is correctly specified, the estimates $\bar{I}(\hat{\boldsymbol{\theta}})$ and $\bar{J}(\hat{\boldsymbol{\theta}})$ should be broadly similar, being estimators based on two different expressions for the same Fisher information matrix. In practice, we could also estimate $I(\boldsymbol{\theta})$ by $\bar{J}(\hat{\boldsymbol{\theta}})\bar{I}(\hat{\boldsymbol{\theta}})^{-1}\bar{J}(\hat{\boldsymbol{\theta}})$, and this anticipates the so-called *sandwich estimator* that is used in the QML procedure.

Properties of QMLEs. In this approach we assume that the true data-generating mechanism is a GARCH(p, q) model with non-Gaussian innovations, but we attempt to estimate the parameters of the process by maximizing the likelihood for a GARCH(p, q) model with Gaussian innovations. We still obtain consistent estimators of the model parameters and, if the true innovation distribution has a finite fourth moment, we again get asymptotic normality; however, the form of the asymptotic covariance matrix changes.

We now distinguish between matrices $I(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$, given by

$$I(\boldsymbol{\theta}) = E\left(\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right), \qquad J(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 l_t(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right),$$

where the expectation is now taken with respect to the true model (not the misspecified Gaussian model). The matrices $I(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ differ in general (unless the Gaussian model is correct). It may be shown that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{\text{d}} N_{p+q+1}(\mathbf{0}, J(\boldsymbol{\theta})^{-1}I(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}), \tag{4.39}$$

and the asymptotic covariance matrix is said to be of sandwich form; it can be estimated by $\bar{J}(\hat{\boldsymbol{\theta}})^{-1}\bar{I}(\hat{\boldsymbol{\theta}})\bar{J}(\hat{\boldsymbol{\theta}})^{-1}$, where $\bar{I}(\boldsymbol{\theta})$ and $\bar{J}(\boldsymbol{\theta})$ are defined in (4.38). If the model-checking procedures described below suggest that the dynamics have been adequately described by the GARCH model, but the Gaussian assumption seems doubtful, then standard errors for parameter estimates should be based on this covariance matrix estimate.

Model checking. As with ARMA models, it is usual to check fitted GARCH models using residuals. We consider a general ARMA–GARCH model of the form $X_t - \mu_t = \varepsilon_t = \sigma_t Z_t$, with μ_t and σ_t as in Definition 4.22. In this model we distinguish between *unstandardized* and *standardized* residuals. The former are the residuals

$\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n$ from the ARMA part of the model; they are calculated using the approach in (4.13), and under the hypothesized model they should behave like a realization of a pure GARCH process. The latter are reconstructed realizations of the SWN that is assumed to drive the GARCH part of the model, and they are calculated from the former by

$$\hat{Z}_t = \hat{\varepsilon}_t / \hat{\sigma}_t, \qquad \hat{\sigma}_t^2 = \hat{\alpha}_0 + \sum_{i=1}^{p_2} \hat{\alpha}_i \hat{\varepsilon}_{t-i}^2 + \sum_{j=1}^{q_2} \hat{\beta}_j \hat{\sigma}_{t-j}^2. \tag{4.40}$$

To use (4.40) we need some initial values, and one solution is to set required starting values of $\hat{\varepsilon}_t$ equal to zero and required starting values of the volatility $\hat{\sigma}_t$ equal to either the sample variance or zero. Because the first few values will be influenced by these starting values, as well as the starting values required to calculate the unstandardized residuals, they may be ignored in later analyses.

The standardized residuals should behave like an SWN and this can be investigated by constructing correlograms of raw and absolute values and applying portmanteau tests of strict white noise, as described in Section 4.1.3.

Assuming that the SWN hypothesis is not rejected, so that the dynamics have been satisfactorily captured, the validity of the distribution used in the ML fitting can also be investigated using Q–Q plots and goodness-of-fit tests for the normal or scaled t distributions. If the Gaussian likelihood does a reasonable job of estimating dynamics, but the residuals do not behave like iid standard normal observations, then the QML fitting philosophy can be adopted and standard errors can be estimated using the sandwich estimator implied by (4.39) above.

This opens up the possibility of *two-stage analyses*, where first the dynamics are estimated by QML methods and then the innovation distribution is modelled using the residuals from the dynamic model as data. The first stage is sometimes called *pre-whitening* of the data. In the second stage we might consider using heavier-tailed models than the Gaussian that also allow some asymmetry in the innovations.

A disadvantage of the two-stage approach is that the error from the time-series modelling propagates through to the distributional fitting in the second stage and the overall error is hard to quantify, but the procedure does lead to more transparency in model building and allows us to separate the tasks of volatility modelling and modelling the shocks that drive the process. In higher-dimensional risk-factor modelling, it may be a useful pragmatic approach.

Example 4.24 (GARCH model for Microsoft log-returns). We consider the Microsoft daily log-returns for the period 1997–2000 (1009 values), as shown in Figure 4.5. Although the raw returns show no evidence of serial correlation (see Figure 4.6), their absolute values do show serial correlation and they fail a Ljung–Box test (based on the first ten estimated correlations) at the 5% level.

For these data, models with Student t innovations are clearly preferred to models with Gaussian innovations, so we adopt an ML approach to fitting models with t innovations. We compare the standard GARCH(1, 1) model (with a constant mean term) with models that incorporate ARMA structure (AR(1), MA(1) and ARMA(1, 1))

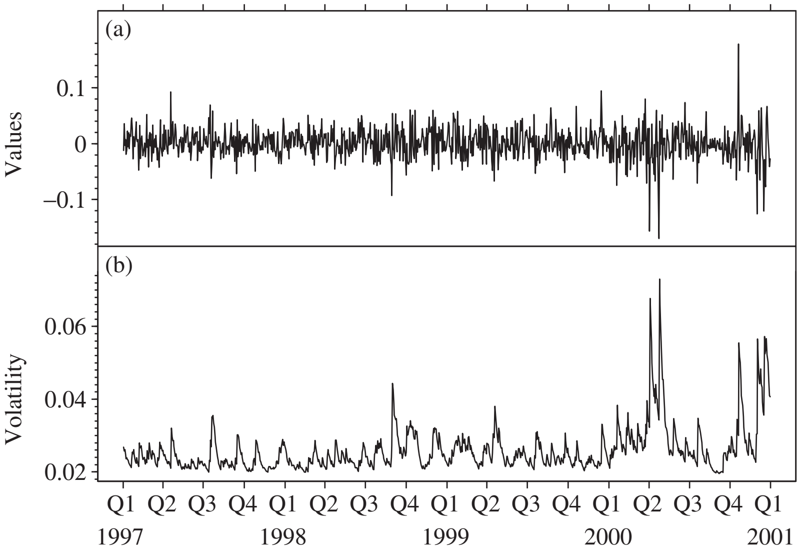

Figure 4.5. Microsoft log-returns 1997–2000; data and estimate of volatility from a GARCH(1, 1) model with a leverage term. (a) Original series. (b) Conditional standard deviation.

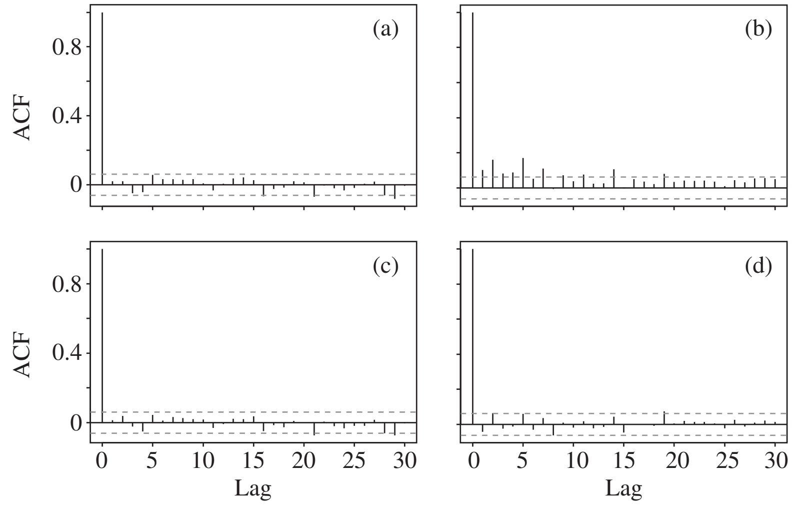

Figure 4.6. Microsoft log-returns 1997–2000; correlograms of data ((a) raw and (b) absolute values) and residuals ((c) raw and (d) absolute values) from a GARCH(1, 1) model.

for the conditional mean; the ARMA structure seems to offer little improvement in the model, and the basic GARCH(1, 1) model is favoured in an Akaike comparison. However, a model with a leverage term as in (4.29) does seem to offer an improvement. Both the raw and absolute standardized residuals obtained from this model show no visual evidence of serial correlation (see again Figure 4.6) and they do not fail Ljung–Box tests. The estimated degrees-of-freedom parameter of the (scaled)

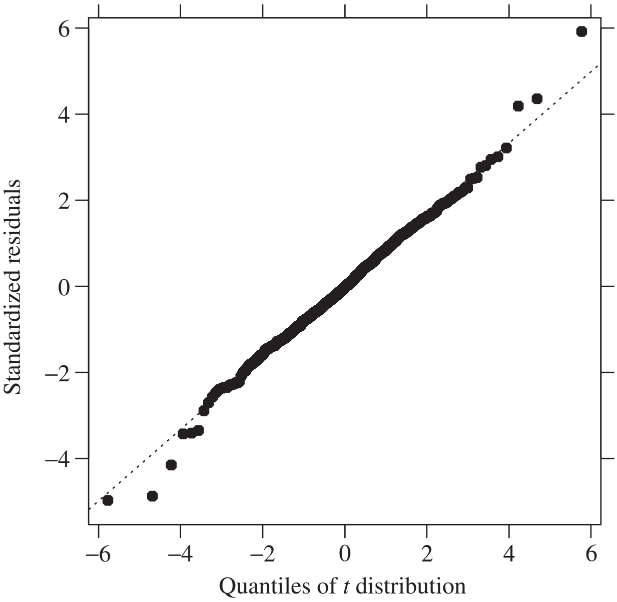

Figure 4.7. Microsoft log-returns 1997–2000; Q–Q plot of residuals from a GARCH(1, 1) model with leverage against a Student t distribution with 6.30 degrees of freedom.

Table 4.1. Analysis of Microsoft log-returns for the period 1997–2000; ML estimates of parameters and standard errors for a GARCH(1, 1) model with a leverage term under the assumption of t innovations.

Parameter	Estimate	Standard error	Ratio
μ	9.35×10^{-4}	7.21×10^{-4}	1.30
α_0	7.79×10^{-5}	3.07×10^{-5}	2.54
α_1	0.108	0.0369	2.91
β_1	0.778	0.0673	11.57
δ	−0.178	0.123	−1.45

t distribution is 6.30 (the standard error is 1.07) and a Q–Q plot of the residuals against this reference distribution reveals a satisfactory correspondence (see Figure 4.7). The estimates of the remaining parameters (with standard errors) in this model are given in Table 4.1.

4.2.5 Volatility Forecasting and Risk Measure Estimation

In this section we assume that our underlying model is a strictly and covariance-stationary time-series process (X_t) adapted to a filtration $(\mathcal{F}_t)$ satisfying equations of the form

$$X_t = \mu_t + \sigma_t Z_t, \tag{4.41}$$

where μ_t and σ_t are $\mathcal{F}_{t-1}$-measurable and Z_t is an innovation variable with mean 0 and variance 1 that is independent of $\mathcal{F}_{t-1}$. Examples fitting into the framework of (4.41) are any of the ARCH and GARCH models discussed in this chapter as well as causal and invertible ARMA models with GARCH errors.

Our task is to forecast σ_{t+h} for $h \geqslant 1$ based on a sample of n data $X_{t-n+1}, \dots, X_t$, which are assumed to be generated by the process (4.41). As in Section 4.1.5 we assume that we have observed the infinite history of the process up to time t and derive prediction formulas that we adapt to take account of the finiteness of the sample.

Since

$$E(\sigma_{t+h}^2 \mid \mathcal{F}_t) = E((X_{t+h} - \mu_{t+h})^2 \mid \mathcal{F}_t),$$

our forecasting problem is closely related to the problem of predicting $(X_{t+h} - \mu_{t+h})^2$, and we can use a similar approach to prediction to the one described in Section 4.1.5. We first derive prediction equations under explicit assumptions about the underlying model (i.e. when we specify the structure of σ_t and μ_t in (4.41)) before presenting the more ad hoc technique of exponentially weighted moving-average (EWMA) prediction. Finally, we describe how forecasts of volatility form the basis for estimates of value-at-risk and expected shortfall.

GARCH-based volatility prediction. Assume that a GARCH model has been fitted and its parameters estimated; we will suppress estimator notation for the parameters in the remainder of the section. We make calculations for two simple models, from which the general procedure for more complex models should be clear.

Example 4.25 (prediction in the GARCH(1, 1) model). Suppose that we use a pure GARCH(1, 1) model as in Definition 4.20, which conforms to (4.41) with $\mu_t = 0$. Since $E(X_{t+h} \mid \mathcal{F}_t) = 0$ (the martingale-difference property of the GARCH process), optimal predictions of X_{t+h} are zero. A natural prediction of X_{t+1}^2 based on $\mathcal{F}_t$ is its conditional mean σ_{t+1}^2 given by

$$E(X_{t+1}^2 \mid \mathcal{F}_t) = \sigma_{t+1}^2 = \alpha_0 + \alpha_1 X_t^2 + \beta_1 \sigma_t^2,$$

and, if $E(X_t^4) < \infty$, this is the optimal squared error prediction. Note that the prediction of the random variable X_{t+1}^2 based on the information $\mathcal{F}_t$ is the value of σ_{t+1}^2, which is *known* at time t, being a function of the history of the process.

In practice, we have to make an approximation based on this formula because the infinite series of past values that would allow us to calculate σ_t^2 is not available to us. A natural approach in applications is to approximate σ_t^2 by an estimate of squared volatility $\hat{\sigma}_t^2$ calculated from the residual equations (4.40). Our approximate forecast of X_{t+1}^2 also functions as an estimate of the squared volatility at time $t + 1$ and is given by

$$\hat{\sigma}_{t+1}^2 = \hat{E}(X_{t+1}^2 \mid \mathcal{F}_t) = \alpha_0 + \alpha_1 X_t^2 + \beta_1 \hat{\sigma}_t^2. \tag{4.42}$$

Thus equation (4.42) can be thought of as a recursive scheme for estimating volatility one step ahead.

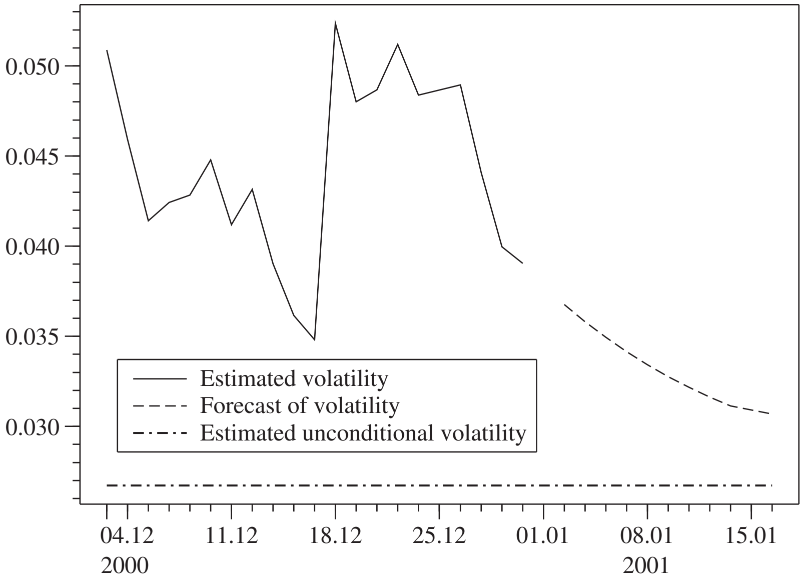

Figure 4.8. Estimate of volatility for the final days of the year 2000 and predictions of volatility for the first ten days of 2001 based on a GARCH(1, 1) model (without leverage) fitted to the Microsoft return data in Example 4.24.

When we look $h > 1$ steps ahead given the information at time t, both X_{t+h}^2 and σ_{t+h}^2 are rvs. Their predictions coincide and are

$$\begin{aligned} E(X_{t+h}^2 \mid \mathcal{F}_t) &= E(\sigma_{t+h}^2 \mid \mathcal{F}_t) \\ &= \alpha_0 + \alpha_1 E(X_{t+h-1}^2 \mid \mathcal{F}_t) + \beta_1 E(\sigma_{t+h-1}^2 \mid \mathcal{F}_t) \\ &= \alpha_0 + (\alpha_1 + \beta_1) E(X_{t+h-1}^2 \mid \mathcal{F}_t), \end{aligned}$$

so that a general formula is

$$E(X_{t+h}^2 \mid \mathcal{F}_t) = \alpha_0 \sum_{i=0}^{h-1} (\alpha_1 + \beta_1)^i + (\alpha_1 + \beta_1)^{h-1} (\alpha_1 X_t^2 + \beta_1 \sigma_t^2),$$

and we obtain a practical formula by substituting an estimate of squared volatility $\hat{\sigma}_t^2$ as before. As $h \to \infty$ we observe that $E(\sigma_{t+h}^2 \mid \mathcal{F}_t) \to \alpha_0/(1 - \alpha_1 - \beta_1)$, almost surely, so that the prediction of squared volatility converges to the unconditional variance of the process. A concrete example of volatility prediction in a GARCH(1, 1) model is given in Figure 4.8 for the Microsoft data analysed in Example 4.24.

We now consider a second example, which combines what we know about prediction in ARMA and GARCH models.

Example 4.26 (prediction in an ARMA(1, 1)–GARCH(1, 1) model). Suppose that we use an ARMA(1, 1) model with GARCH(1, 1) errors as in Definition 4.22. This also conforms to (4.41), and prediction formulas for this model follow easily from Examples 4.15 and 4.25. We calculate that

$$E(X_{t+h} \mid \mathcal{F}_t) = \mu + \phi_1^h(X_t - \mu) + \phi_1^{h-1}\theta_1\varepsilon_t, \tag{4.43}$$

$$\operatorname{var}(X_{t+h} \mid \mathcal{F}_t) = \alpha_0 \sum_{i=0}^{h-1}(\alpha_1 + \beta_1)^i + (\alpha_1 + \beta_1)^{h-1}(\alpha_1\varepsilon_t^2 + \beta_1\sigma_t^2), \tag{4.44}$$

and these are approximated by substituting inferred values for ε_t and σ_t obtained from the residual equations (4.40). Equation (4.43) yields predictions of μ_{t+h} or X_{t+h}, and equation (4.44) yields predictions of $(X_{t+h} - \mu_{t+h})^2$ or σ_{t+h}^2.

Exponential smoothing for volatility. Now suppose that we do not want to make detailed assumptions about the structure of σ_t and μ_t in (4.42). We consider a simpler scheme for forecasting volatility that builds on the exponential smoothing idea of Section 4.1.5. We recall from (4.14) that a forecast $P_t(X_{t+1})$ of X_{t+1} based on time-t information can be constructed using an updating scheme of the form

$$P_t X_{t+1} = \lambda X_t + (1 - \lambda)P_{t-1}X_t \tag{4.45}$$

for an appropriately chosen value of the parameter λ. If we apply this scheme to the prediction of $(X_{t+1} - \mu_{t+1})^2$, we obtain

$$P_t(X_{t+1} - \mu_{t+1})^2 = \alpha(X_t - \mu_t)^2 + (1 - \alpha)P_{t-1}(X_t - \mu_t)^2 \tag{4.46}$$

for an appropriately chosen value of the parameter α. Of course, in addition to choosing α, we also need to insert an estimate of the unobserved conditional mean μ_t to use (4.46).

Since $\sigma_{t+1}^2 = E((X_{t+1} - \mu_{t+1})^2 \mid \mathcal{F}_t)$, we can also use (4.46) as an exponential smoothing scheme for the unobserved squared volatility. This yields a recursive scheme for the one-step-ahead volatility forecast given by

$$\hat{\sigma}_{t+1}^2 = \alpha(X_t - \hat{\mu}_t)^2 + (1 - \alpha)\hat{\sigma}_t^2, \tag{4.47}$$

which defines the EWMA procedure. For many risk-factor return series, the conditional mean appears to be close to zero (recall the stylized facts of return series in Section 3.1) and we often set $\hat{\mu}_t = 0$. Alternatively, we can apply the exponential smoothing idea to the conditional mean and replace $\hat{\mu}_t$ by an estimate $P_{t-1}X_t$ derived using the recursive scheme (4.45). Typical values for α are generally small; for example, in the RiskMetrics methodology widely used by banks, a value of $\alpha = 0.06$ has been recommended (Mina and Xiao 2001).

If we compare (4.47) with the one-step-ahead volatility estimation scheme defined by a GARCH(1, 1) model in (4.42), it is tempting to say that EWMA corresponds to estimating volatility using a conditional-expectation-based technique in an IGARCH model, where the parameter α_0 equals zero. This analogy should be used with care; GARCH and IGARCH models with $\alpha_0 = 0$ are not well defined, and the solution of the stochastic recurrence relation in (4.27) vanishes. Moreover, IGARCH is not covariance stationary. It is better to regard EWMA as a sensible model-free approach to volatility forecasting based on the classical technique of exponential smoothing.

Estimates of VaR and expected shortfall. Finally, we suppose that the data $X_{t-n+1}, \dots, X_t$ can be interpreted as financial losses and we consider the application of risk measures based on loss distributions (see Section 2.3.1) to the conditional distribution $F_{X_{t+1}|\mathcal{F}_t}$. For example, the data may represent *negative* log-returns on an asset price rather than returns. In particular, we look at the estimation of value-at-risk and expected shortfall for the distribution $F_{X_{t+1}|\mathcal{F}_t}$.

Writing F_Z for the df of the innovations (Z_t), the $\mathcal{F}_t$-measurability of μ_{t+1} and σ_{t+1} implies that

$$F_{X_{t+1}|\mathcal{F}_t}(x) = P(\mu_{t+1} + \sigma_{t+1}Z_{t+1} \leqslant x \mid \mathcal{F}_t) = F_Z((x - \mu_{t+1})/\sigma_{t+1}).$$

Let VaR_α^t denote the α-quantile of $F_{X_{t+1}|\mathcal{F}_t}$ and let ES_α^t denote the corresponding expected shortfall. Using the approach of Examples 2.11 and 2.14 we obtain

$$\mathrm{VaR}_\alpha^t = \mu_{t+1} + \sigma_{t+1}q_\alpha(Z), \qquad \mathrm{ES}_\alpha^t = \mu_{t+1} + \sigma_{t+1}\,\mathrm{ES}_\alpha(Z), \tag{4.48}$$

where we write Z for a generic rv with df F_Z.

It is clear that if we can estimate μ_{t+1} and σ_{t+1}, then we only need to be able to estimate $q_\alpha(Z)$ and $\mathrm{ES}_\alpha(Z)$ for the innovation distribution to obtain estimates of the risk measures in (4.48). This task can be accomplished in both a parametric and a non-parametric (or semi-parametric) setting. If we estimate a fully specified GARCH-type model using the ML approach of Section 4.2.4, then it is mostly straightforward to calculate $q_\alpha(Z)$ and $\mathrm{ES}_\alpha(Z)$ for the estimated innovation distribution. If, on the other hand, we use a QML method to fit a GARCH-type model or, even more simply, we use exponential smoothing techniques to estimate the volatility and conditional mean, then we can form residuals $\hat{Z}_s = (X_s - \hat{\mu}_s)/\hat{\sigma}_s$ for $s = t - n + 1, \dots, n$ and apply quantile and expected shortfall estimation techniques to these residuals; statistical methods for estimating risk measures from data are discussed in Section 9.2.6.

Notes and Comments

The ARCH process was originally proposed by Engle (1982), and the GARCH process by Bollerslev (1986), who gave the condition for covariance stationarity. Overview texts on GARCH models include the books by Gouriéroux (1997) and Francq and Zakoïan (2010) and a number of useful review articles including Bollerslev, Chou and Kroner (1992), Bollerslev, Engle and Nelson (1994) and Shephard (1996). There are also substantial sections on GARCH models in the books by Alexander (2001), Tsay (2002) and Zivot and Wang (2003). The IGARCH model was first discussed by Engle and Bollerslev (1986).

The condition for strict stationarity of GARCH models was derived by Nelson (1990) in the case of the GARCH(1, 1) model and by Bougerol and Picard (1992) for GARCH(p, q). The necessary theory involves the study of stochastic recurrence relations and goes back to Kesten (1973); Brandt (1986) is also a useful reference. Readable accounts of this theory may be found in Embrechts, Klüppelberg and Mikosch (1997), Mikosch and Stărică (2000) and Mikosch (2003, 2013).

For more on the derivation of conditional likelihood functions for ARCH and GARCH models, see Hamilton (1994) and Tsay (2002). The BHHH algorithm (Berndt et al. 1974) is the most commonly used approach to numerically maximizing the likelihood. For an informative general discussion of numerical optimization procedures in the context of maximum likelihood, see Hamilton (1994, pp. 133–142). Standard general references on the QML approach are White (1981) and Gouriéroux, Montfort and Trognon (1984).

The essential asymptotic properties of MLEs and QMLEs in GARCH models are described in many publications, but a detailed mathematical proof has often lagged behind the assertions. Early papers appealed to regularity conditions for conditionally specified models such as those of Crowder (1976), which are essentially unverifiable. Lee and Hansen (1994) and Lumsdaine (1996) proved consistency and asymptotic normality of QMLEs in the GARCH$(1, 1)$ model. More recently, Berkes, Horváth and Kokoszka (2003) have extended this to the GARCH(p, q) model under minimal assumptions, and Straumann (2005) and Straumann and Mikosch (2006) have given similar results for a wide variety of first-order models.

From a more practical point-of-view, it is not easy to estimate GARCH model parameters to a high degree of accuracy because of the flatness of the typical likelihoods and the non-negligible influence of starting values in finite samples. Readers who write their own code may wish to compare their estimates with benchmark studies by McCullough and Renfro (1999) and Brooks, Burke and Persand (2001).

Alternative innovation distributions to the Gaussian and scaled t distributions that have been considered include the generalized error distribution (GED) in Nelson (1991) and the normal inverse Gaussian (NIG) in Venter and de Jongh (2002); the latter authors present extensive evidence that the NIG is a good choice of innovation distribution for practical work and that GARCH inference based on the NIG is relatively robust to misspecification of the distribution.

A great many extensions to the GARCH class have been proposed and thorough surveys may be found in Bollerslev, Engle and Nelson (1994) and Shephard (1996). Leverage effects in the GARCH model and the more general PGARCH (power GARCH) model are examined in Ding, Granger and Engle (1993). Various threshold GARCH models have been suggested; the model (4.30) is of the type suggested by Glosten, Jagannathan and Runkle (1993), while (4.31) is the switching-volatility GARCH (SV-GARCH) model of Fornari and Mele (1997). There have been proposals for non-parametric ARCH and GARCH modelling, including the multiplicative ARCH(p)-model of Yang, Härdle and Nielsen (1999) and the non-parametric GARCH procedure of Bühlmann and McNeil (2002). For long-memory processes modelling volatility, see the book by Beran et al. (2013).

The use of the EWMA (exponentially weighted moving-average) volatility estimation method based on exponential smoothing was popularized by the RiskMetrics Group at JPMorgan (JPMorgan 1996; Mina and Xiao 2001). See also Zivot and Wang (2003) for examples of the use of this method.

5
Extreme Value Theory

Extreme value theory (EVT) is a branch of probability concerned with limiting laws for extreme values in large samples. The theory contains many important results describing the behaviour of sample maxima and minima, upper-order statistics (such as the kth largest value in a sample) and sample values exceeding high thresholds. Our interest in this theory centres on the application of the results to developing models for the extremal behaviour of financial risk factors. In particular, we are interested in models for the tail of the distribution of financial risk-factor changes. We have observed at various points in Chapters 3 and 4 that risk-factor changes are frequently heavy tailed when compared with a normal distribution.

Much of this chapter is based on the presentation of EVT in Embrechts, Klüppelberg and Mikosch (1997) (henceforth, EKM), and whenever theoretical detail is missing the reader should consult that text. We concentrate on describing the statistical models suggested by EVT, while briefly summarizing the theoretical ideas on which the statistical methods are based.

We focus on two main kinds of model for extreme values. The most traditional models are the block maxima models described in Section 5.1: these are models for the largest observations collected from large samples of identically distributed observations. A more modern and powerful group of models are those for threshold exceedances, described in Section 5.2. These are models for all large observations that exceed some high level, and they are generally considered to be the most useful for practical applications, due to their more efficient use of the (often limited) data on extreme outcomes.

The models for threshold exceedances can be embedded in an elegant point process framework that simultaneously addresses their occurrence in time as well as the magnitude of excess losses over the threshold. This is the so-called peaks-over-threshold (POT) model, which is presented in Section 5.3. The POT model serves as a starting point for developing more dynamic descriptions of the occurrence and magnitude of extremes using self-exciting (Hawkes) processes. These advanced dynamic models are treated in Chapter 16 along with multivariate EVT.

5.1 Maxima

To begin with we consider a sequence of iid rvs $(X_i)_{i\in\mathbb{N}}$ representing financial losses. These may have a variety of interpretations, such as operational losses, insurance losses and losses on a credit portfolio over fixed time intervals. Later we relax the assumption of independence and consider that the rvs form a strictly stationary time series of dependent losses; they might be (negative) returns on an investment in a single stock, an index, or a portfolio of investments.

5.1.1 Generalized Extreme Value Distribution

Convergence of sums. The role of the generalized extreme value (GEV) distribution in the theory of extremes is analogous to that of the normal distribution (and, more generally, the stable laws) in the central limit theory for sums of rvs. Assuming that the underlying rvs $X_1, X_2, \dots$ are iid with a finite variance, and writing $S_n = X_1 + \cdots + X_n$ for the sum of the first n rvs, the standard version of the central limit theorem (CLT) says that appropriately normalized sums $(S_n - a_n)/b_n$ converge in distribution to the standard normal distribution as n goes to infinity. The appropriate normalization uses sequences of normalizing constants (a_n) and (b_n) defined by $a_n = nE(X_1)$ and $b_n = \sqrt{n \operatorname{var}(X_1)}$. In mathematical notation we have

$$\lim_{n\to\infty} P\left(\frac{S_n - a_n}{b_n} \leqslant x\right) = \Phi(x), \quad x \in \mathbb{R}.$$

Convergence of maxima. Classical EVT is concerned with limiting distributions for normalized maxima. We denote the maximum of n iid rvs $X_1, \dots, X_n$ by $M_n = \max(X_1, \dots, X_n)$ and refer to this also as an n-block maximum. The only possible non-degenerate limiting distributions for normalized maxima as n goes to infinity are in the GEV family.

Definition 5.1 (generalized extreme value distribution). The df of the (standard) GEV distribution is given by

$$H_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-1/\xi}), & \xi \neq 0, \\ \exp(-\mathrm{e}^{-x}), & \xi = 0, \end{cases}$$

where $1 + \xi x > 0$. A three-parameter family is obtained by defining $H_{\xi,\mu,\sigma}(x) := H_\xi((x-\mu)/\sigma)$ for a location parameter $\mu \in \mathbb{R}$ and a scale parameter $\sigma > 0$.

The parameter ξ is known as the *shape* parameter of the GEV distribution, and H_ξ defines a *type* of distribution, meaning a family of distributions specified up to location and scaling (see Section A.1.1 for a formal definition). The extreme value distribution in Definition 5.1 is generalized in the sense that the parametric form subsumes three types of distribution that are known by other names according to the value of ξ: when $\xi > 0$ the distribution is a Fréchet distribution; when $\xi = 0$ it is a Gumbel distribution; and when $\xi < 0$ it is a Weibull distribution. We also note that for fixed x we have $\lim_{\xi\to 0} H_\xi(x) = H_0(x)$ (from either side), so that the parametrization in Definition 5.1 is *continuous* in ξ, which facilitates the use of this distribution in statistical modelling.

The df and density of the GEV distribution are shown in Figure 5.1 for the three cases $\xi = 0.5$, $\xi = 0$ and $\xi = -0.5$, corresponding to Fréchet, Gumbel and Weibull types, respectively. Observe that the Weibull distribution is a short-tailed distribution with a so-called finite *right endpoint*. The right endpoint of a distribution will be

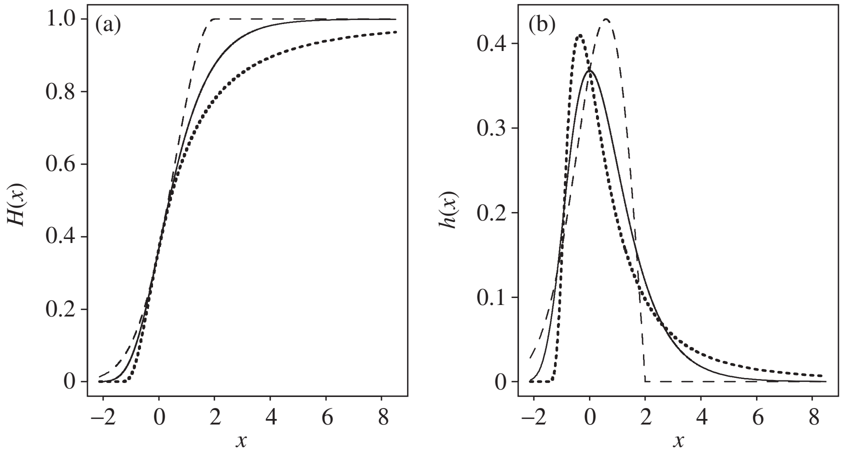

Figure 5.1. (a) The df of a standard GEV distribution in three cases: the solid line corresponds to $\xi = 0$ (Gumbel); the dotted line is $\xi = 0.5$ (Fréchet); and the dashed line is $\xi = -0.5$ (Weibull). (b) Corresponding densities. In all cases, $\mu = 0$ and $\sigma = 1$.

denoted by $x_F = \sup\{x \in \mathbb{R} \colon F(x) < 1\}$. The Gumbel and Fréchet distributions have infinite right endpoints, but the decay of the tail of the Fréchet distribution is much slower than that of the Gumbel distribution.

Suppose that maxima M_n of iid rvs converge in distribution as $n \to \infty$ under an appropriate normalization. Recalling that $P(M_n \leqslant x) = F^n(x)$, we observe that this convergence means that there exist sequences of real constants (d_n) and (c_n), where $c_n > 0$ for all n, such that

$$\lim_{n\to\infty} P((M_n - d_n)/c_n \leqslant x) = \lim_{n\to\infty} F^n(c_n x + d_n) = H(x) \tag{5.1}$$

for some non-degenerate df $H(x)$. The role of the GEV distribution in the study of maxima is formalized by the following definition and theorem.

Definition 5.2 (maximum domain of attraction). If (5.1) holds for some non-degenerate df H, then F is said to be in the maximum domain of attraction of H, written $F \in \mathrm{MDA}(H)$.

Theorem 5.3 (Fisher–Tippett, Gnedenko). *If $F \in \mathrm{MDA}(H)$ for some non-degenerate df H, then H must be a distribution of type H_ξ, i.e. a GEV distribution.*

Remarks 5.4.

(1) If convergence of normalized maxima takes place, the type of the limiting distribution (as specified by ξ) is uniquely determined, although the location and scaling of the limit law (μ and σ) depend on the exact normalizing sequences chosen; this is guaranteed by the so-called convergence to types theorem (EKM, p. 554). It is always possible to choose these sequences such that the limit appears in the standard form H_ξ.

(2) By non-degenerate df we mean a distribution that is not concentrated on a single point.

Examples. We calculate two examples to show how the GEV limit emerges for two well-known underlying distributions and appropriately chosen normalizing sequences. To discover how normalizing sequences may be constructed in general, we refer to Section 3.3 of EKM.

Example 5.5 (exponential distribution). If the underlying distribution is an exponential distribution with df $F(x) = 1 - \mathrm{e}^{-\beta x}$ for $\beta > 0$ and $x \geqslant 0$, then by choosing normalizing sequences $c_n = 1/\beta$ and $d_n = (\ln n)/\beta$ we can directly calculate the limiting distribution of maxima using (5.1). We get

$$\begin{aligned} F^n(c_n x + d_n) &= \left(1 - \frac{1}{n}\mathrm{e}^{-x}\right)^n, \quad && x \geqslant -\ln n, \\ \lim_{n\to\infty} F^n(c_n x + d_n) &= \exp(-\mathrm{e}^{-x}), && x \in \mathbb{R}, \end{aligned}$$

from which we conclude that $F \in \mathrm{MDA}(H_0)$.

Example 5.6 (Pareto distribution). If the underlying distribution is a Pareto distribution ($\mathrm{Pa}(\alpha, \kappa)$) with df $F(x) = 1 - (\kappa/(\kappa + x))^\alpha$ for $\alpha > 0$, $\kappa > 0$ and $x \geqslant 0$, we can take normalizing sequences $c_n = \kappa n^{1/\alpha}/\alpha$ and $d_n = \kappa n^{1/\alpha} - \kappa$. Using (5.1) we get

$$\begin{aligned} F^n(c_n x + d_n) &= \left(1 - \frac{1}{n}\left(1 + \frac{x}{\alpha}\right)^{-\alpha}\right)^n, \quad && 1 + \frac{x}{\alpha} \geqslant n^{-1/\alpha}, \\ \lim_{n\to\infty} F^n(c_n x + d_n) &= \exp\left(-\left(1 + \frac{x}{\alpha}\right)^{-\alpha}\right), && 1 + \frac{x}{\alpha} > 0, \end{aligned}$$

from which we conclude that $F \in \mathrm{MDA}(H_{1/\alpha})$.

Convergence of minima. The limiting theory for convergence of maxima encompasses the limiting behaviour of minima using the identity

$$\min(X_1, \dots, X_n) = -\max(-X_1, \dots, -X_n). \tag{5.2}$$

It is not difficult to see that normalized minima of iid samples with df F will converge in distribution if the df $\tilde{F}(x) = 1 - F(-x)$, which is the df of the rvs $-X_1, \dots, -X_n$, is in the maximum domain of attraction of an extreme value distribution. Writing $M_n^* = \max(-X_1, \dots, -X_n)$ and assuming that $\tilde{F} \in \mathrm{MDA}(H_\xi)$, we have

$$\lim_{n\to\infty} P\left(\frac{M_n^* - d_n}{c_n} \leqslant x\right) = H_\xi(x),$$

from which it follows easily, using (5.2), that

$$\lim_{n\to\infty} P\left(\frac{\min(X_1, \dots, X_n) + d_n}{c_n} \leqslant x\right) = 1 - H_\xi(-x).$$

Thus appropriate limits for minima are distributions of type $1 - H_\xi(-x)$. For a symmetric distribution F we have $\tilde{F}(x) = F(x)$, so that if H_ξ is the limiting type of distribution for maxima for a particular value of ξ, then $1 - H_\xi(-x)$ is the limiting type of distribution for minima.

5.1.2 Maximum Domains of Attraction

For most applications it is sufficient to note that essentially all the common continuous distributions of statistics or actuarial science are in $\mathrm{MDA}(H_\xi)$ for some value of ξ. In this section we consider the issue of which underlying distributions lead to which limits for maxima.

The Fréchet case. The distributions that lead to the Fréchet limit $H_\xi(x)$ for $\xi > 0$ have a particularly elegant characterization involving *slowly varying* or *regularly varying* functions.

Definition 5.7 (slowly varying and regularly varying functions).

(i) A positive, Lebesgue-measurable function L on $(0, \infty)$ is slowly varying at ∞ (written $L \in \mathcal{R}_0$) if

$$\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1, \quad t > 0.$$

(ii) A positive, Lebesgue-measurable function h on $(0, \infty)$ is regularly varying at ∞ with index $\rho \in \mathbb{R}$ if

$$\lim_{x\to\infty} \frac{h(tx)}{h(x)} = t^\rho, \quad t > 0.$$

Slowly varying functions are functions that, in comparison with power functions, change relatively slowly for large x, an example being the logarithm $L(x) = \ln(x)$. Regularly varying functions are functions that can be represented by power functions multiplied by slowly varying functions, i.e. $h(x) = x^\rho L(x)$ for some $L \in \mathcal{R}_0$.

Theorem 5.8 (Fréchet MDA, Gnedenko). *For* $\xi > 0$,

$$F \in \mathrm{MDA}(H_\xi) \iff \bar{F}(x) = x^{-1/\xi} L(x) \text{ for some function } L \in \mathcal{R}_0. \tag{5.3}$$

This means that distributions giving rise to the Fréchet case are distributions with tails that are regularly varying functions with a negative index of variation. Their tails decay essentially like a power function, and the rate of decay $\alpha = 1/\xi$ is often referred to as the *tail index* of the distribution. A consequence of Theorem 5.8 is that the right endpoint of any distribution in the Fréchet MDA satisfies $x_F = \infty$.

These distributions are the most studied distributions in EVT and they are of particular interest in financial applications because they are heavy-tailed distributions with infinite higher moments. If X is a non-negative rv whose df F is an element of $\mathrm{MDA}(H_\xi)$ for $\xi > 0$, then it may be shown that $E(X^k) = \infty$ for $k > 1/\xi$ (EKM, p. 568). If, for some small $\varepsilon > 0$, the distribution is in $\mathrm{MDA}(H_{(1/2)+\varepsilon})$, it is an infinite-variance distribution, and if the distribution is in $\mathrm{MDA}(H_{(1/4)+\varepsilon})$, it is a distribution with infinite fourth moment.

Example 5.9 (Pareto distribution). In Example 5.6 we verified by direct calculation that normalized maxima of iid Pareto variates converge to a Fréchet distribution. Observe that the tail of the Pareto df in (A.19) may be written $\bar{F}(x) = x^{-\alpha} L(x)$, where it may be easily checked that $L(x) = (\kappa^{-1} + x^{-1})^{-\alpha}$ is a slowly varying function; indeed, as $x \to \infty$, $L(x)$ converges to the constant κ^{α}. Thus we verify that the Pareto df has the form (5.3).

Further examples of distributions giving rise to the Fréchet limit for maxima include the Fréchet distribution itself and the inverse gamma, Student t, loggamma, F and Burr distributions. We will provide further demonstrations for some of these distributions in Section 16.1.1.

The Gumbel case. The characterization of distributions in this class is more complicated than in the Fréchet class. We have seen in Example 5.5 that the exponential distribution is in the Gumbel class and, more generally, it could be said that the distributions in this class have tails that have an essentially exponential decay. A positive-valued rv with a df in $\mathrm{MDA}(H_0)$ has finite moments of any positive order, i.e. $E(X^k) < \infty$ for every $k > 0$ (EKM, p. 148).

However, there is a great deal of variety in the tails of distributions in this class, so, for example, both the normal and lognormal distributions belong to the Gumbel class (EKM, pp. 145–147). The normal distribution, as discussed in Section 6.1.4, is thin tailed, but the lognormal distribution has much heavier tails, and we would need to collect a lot of data from the lognormal distribution before we could distinguish its tail behaviour from that of a distribution in the Fréchet class. Moreover, it should be noted that the right endpoints of distributions in this class satisfy $x_F \leqslant \infty$, so the case $x_F < \infty$ is possible.

In financial modelling it is often erroneously assumed that the only interesting models for financial returns are the power-tailed distributions of the Fréchet class. The Gumbel class is also interesting because it contains many distributions with much heavier tails than the normal, even if these are not regularly varying power tails. Examples are hyperbolic and generalized hyperbolic distributions (with the exception of the special boundary case that is Student t).

Other distributions in $\mathrm{MDA}(H_0)$ include the gamma, chi-squared, standard Weibull (to be distinguished from the Weibull special case of the GEV distribution) and Benktander type I and II distributions (which are popular actuarial loss distributions), and the Gumbel itself. We provide demonstrations for some of these examples in Section 16.1.2.

The Weibull case. This is perhaps the least important case for financial modelling, at least in the area of market risk, since the distributions in this class all have finite *right endpoints*. Although all potential financial and insurance losses are, in practice, bounded, we will still tend to favour models that have infinite support for loss modelling. An exception may be in the area of credit risk modelling, where we will see in Chapter 10 that probability distributions on the unit interval [0, 1] are very useful. A characterization of the Weibull class is as follows.

Theorem 5.10 (Weibull MDA, Gnedenko). *For* $\xi < 0$,

$$F \in \mathrm{MDA}(H_\xi) \iff x_F < \infty \text{ and } \bar{F}(x_F - x^{-1}) = x^{1/\xi} L(x) \quad \text{for some function } L \in \mathcal{R}_0.$$

It can be shown (EKM, p. 137) that a beta distribution with density $f_{\alpha,\beta}$ as given in (A.10) is in MDA($H_{-1/\beta}$). This includes the special case of the uniform distribution for $\beta = \alpha = 1$.

5.1.3 Maxima of Strictly Stationary Time Series

The standard theory of the previous sections concerns maxima of iid sequences. With financial time series in mind, we now look briefly at the theory for maxima of strictly stationary time series and find that the same types of limiting distribution apply.

In this section let $(X_i)_{i\in\mathbb{Z}}$ denote a strictly stationary time series with stationary distribution F, and let $(\tilde{X}_i)_{i\in\mathbb{Z}}$ denote the *associated iid process*, i.e. a strict white noise process with the same df F. Let $M_n = \max(X_1, \dots, X_n)$ and $\tilde{M}_n = \max(\tilde{X}_1, \dots, \tilde{X}_n)$ denote maxima of the original series and the iid series, respectively.

For many processes $(X_i)_{i\in\mathbb{N}}$, it may be shown that there exists a real number θ in (0, 1] such that

$$\lim_{n\to\infty} P\left(\frac{\tilde{M}_n - d_n}{c_n} \leqslant x\right) = H(x) \tag{5.4}$$

for a non-degenerate limit $H(x)$ if and only if

$$\lim_{n\to\infty} P\left(\frac{M_n - d_n}{c_n} \leqslant x\right) = H^{\theta}(x). \tag{5.5}$$

For such processes this value θ is known as the *extremal index* of the process (not to be confused with the tail index of distributions in the Fréchet class). A formal definition is more technical (see Notes and Comments) but the basic ideas behind (5.4) and (5.5) are easily explained.

For processes with an extremal index, normalized maxima converge in distribution provided that maxima of the associated iid process converge in distribution: that is, provided the underlying distribution F is in MDA(H_ξ) for some ξ. Moreover, since $H_\xi^\theta(x)$ can be easily verified to be a distribution of the same type as $H_\xi(x)$, the limiting distribution of the normalized maxima of the dependent series is a GEV distribution with exactly the same ξ parameter as the limit for the associated iid data; only the location and scaling of the distribution may change.

Writing $u = c_n x + d_n$, we observe that, for large enough n, (5.4) and (5.5) imply that

$$P(M_n \leqslant u) \approx P^{\theta}(\tilde{M}_n \leqslant u) = F^{n\theta}(u), \tag{5.6}$$

so that for u large, the probability distribution of the maximum of n observations from the time series with extremal index θ can be approximated by the distribution of the maximum of $n\theta < n$ observations from the associated iid series. In a sense, $n\theta$ can be thought of as counting the number of roughly independent *clusters* of observations in n observations, and θ is often interpreted as the reciprocal of the mean cluster size.

Table 5.1. Approximate values of the extremal index as a function of the parameter α_1 for the ARCH(1) process in (4.22).

α_1	0.1	0.3	0.5	0.7	0.9
θ	0.999	0.939	0.835	0.721	0.612

Not every strictly stationary process has an extremal index (see p. 418 of EKM for a counterexample) but, for the kinds of time-series processes that interest us in financial modelling, an extremal index generally exists. Essentially, we only have to distinguish between the cases when $\theta = 1$ and the cases when $\theta < 1$: for the former, there is no tendency to cluster at high levels, and large sample maxima from the time series behave exactly like maxima from similarly sized iid samples; for the latter, we must be aware of a tendency for extreme values to cluster.

- Strict white noise processes (iid rvs) have extremal index $\theta = 1$.
- ARMA processes with Gaussian strict white noise innovations have $\theta = 1$ (EKM, pp. 216–218). However, if the innovation distribution is in MDA(H_ξ) for $\xi > 0$, then $\theta < 1$ (EKM, pp. 415, 416).
- ARCH and GARCH processes have $\theta < 1$ (EKM, pp. 476–480).

The final fact is particularly relevant to our financial applications, since we saw in Chapter 4 that ARCH and GARCH processes provide good models for many financial return series.

Example 5.11 (the extremal index of the ARCH(1) process). In Table 5.1 we reproduce some results from de Haan et al. (1989), who calculate approximate values for the extremal index of the ARCH(1) process (see Definition 4.16) using a Monte Carlo simulation approach. Clearly, the stronger the ARCH effect (that is, the larger the magnitude of the parameter α_1), the greater the tendency of the process to cluster. For a process with parameter 0.9, the extremal index value $\theta = 0.612$ is interpreted as suggesting that the average cluster size is $1/\theta = 1.64$.

5.1.4 The Block Maxima Method

Fitting the GEV distribution. Suppose we have data from an unknown underlying distribution F, which we suppose lies in the domain of attraction of an extreme value distribution H_ξ for some ξ. If the data are realizations of iid variables, or variables from a process with an extremal index such as GARCH, the implication of the theory is that the true distribution of the n-block maximum M_n can be approximated for large enough n by a three-parameter GEV distribution $H_{\xi,\mu,\sigma}$.

We make use of this idea by fitting the GEV distribution $H_{\xi,\mu,\sigma}$ to data on the n-block maximum. Obviously we need repeated observations of an n-block maximum, and we assume that the data can be divided into m blocks of size n. This makes most sense when there are natural ways of blocking the data. The method has its origins in hydrology, where, for example, daily measurements of water levels might be divided into yearly blocks and the yearly maxima collected. Analogously, we will consider financial applications where daily return data (recorded on trading days) are divided into yearly (or semesterly or quarterly) blocks and the maximum daily falls within these blocks are analysed.

We denote the block maximum of the jth block by M_{nj}, so our data are $M_{n1}, \dots, M_{nm}$. The GEV distribution can be fitted using various methods, including maximum likelihood. An alternative is the method of probability-weighted moments (see Notes and Comments). In implementing maximum likelihood it will be assumed that the block size n is quite large so that, regardless of whether the underlying data are dependent or not, the block maxima observations can be taken to be independent. In this case, writing $h_{\xi,\mu,\sigma}$ for the density of the GEV distribution, the log-likelihood is easily calculated to be

$$\begin{aligned} l(\xi, \mu, \sigma; M_{n1}, \dots, M_{nm}) \\ &= \sum_{i=1}^{m} \ln h_{\xi,\mu,\sigma}(M_{ni}) \\ &= -m \ln \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{m} \ln \left(1 + \xi \frac{M_{ni} - \mu}{\sigma}\right) - \sum_{i=1}^{m} \left(1 + \xi \frac{M_{ni} - \mu}{\sigma}\right)^{-1/\xi}, \end{aligned}$$

which must be maximized subject to the parameter constraints that $\sigma > 0$ and $1 + \xi(M_{ni} - \mu)/\sigma > 0$ for all i. While this represents an irregular likelihood problem, due to the dependence of the parameter space on the values of the data, the consistency and asymptotic efficiency of the resulting MLEs can be established for the case when $\xi > -\frac{1}{2}$ using results in Smith (1985).

In determining the number and size of the blocks (m and n, respectively), a trade-off necessarily takes place: roughly speaking, a large value of n leads to a more accurate approximation of the block maxima distribution by a GEV distribution and a low bias in the parameter estimates; a large value of m gives more block maxima data for the ML estimation and leads to a low variance in the parameter estimates. Note also that, in the case of dependent data, somewhat larger block sizes than are used in the iid case may be advisable; dependence generally has the effect that convergence to the GEV distribution is slower, since the effective sample size is $n\theta$, which is smaller than n.

Example 5.12 (block maxima analysis of S&P return data). Suppose we turn the clock back and imagine it is the early evening of Friday 16 October 1987. An unusually turbulent week in the equity markets has seen the S&P 500 index fall by 9.12%. On that Friday alone the index is down 5.16% on the previous day, the largest one-day fall since 1962.

We fit the GEV distribution to annual maximum daily percentage falls in value for the S&P index. Using data going back to 1960, shown in Figure 5.2, gives us twenty-eight observations of the annual maximum fall (including the latest observation from the incomplete year 1987). The estimated parameter values are $\hat{\xi} = 0.29$, $\hat{\mu} = 2.03$ and $\hat{\sigma} = 0.72$ with standard errors 0.21, 0.16 and 0.14, respectively. Thus the fitted distribution is a heavy-tailed Fréchet distribution with an infinite fourth moment,

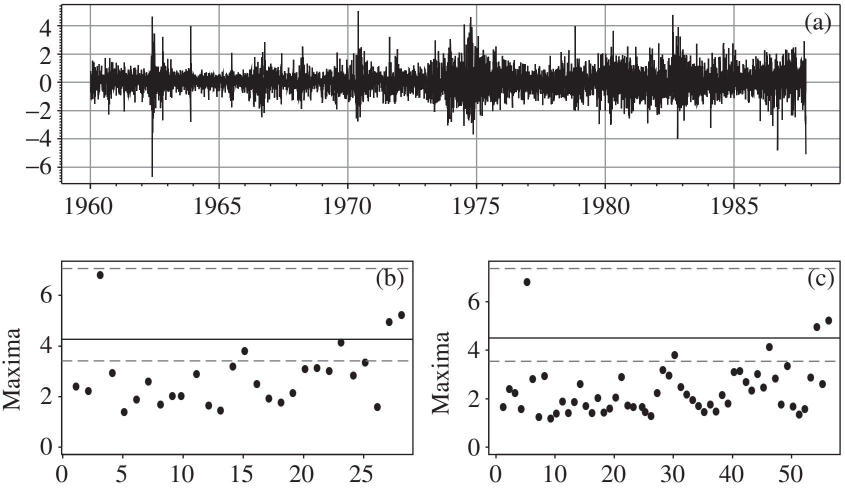

Figure 5.2. (a) S&P percentage returns for the period 1960 to 16 October 1987. (b) Annual maxima of daily falls in the index; superimposed is an estimate of the ten-year return level with associated 95% confidence interval (dotted lines). (c) Semesterly maxima of daily falls in the index; superimposed is an estimate of the 20-semester return level with associated 95% confidence interval. See Examples 5.12 and 5.15 for full details.

suggesting that the underlying distribution is heavy tailed. Note that the standard errors imply considerable uncertainty in our analysis, as might be expected with only twenty-four observations of maxima. In fact, in a likelihood ratio test of the null hypothesis that a Gumbel model fits the data ($H_0 \colon \xi = 0$), the null hypothesis cannot be rejected.

To increase the number of blocks we also fit a GEV model to 56 semesterly maxima and obtain the parameter estimates $\hat{\xi} = 0.33$, $\hat{\mu} = 1.68$ and $\hat{\sigma} = 0.55$ with standard errors 0.14, 0.09 and 0.07. This model has an even heavier tail, and the null hypothesis that a Gumbel model is adequate is now rejected.

Return levels and stress losses. The fitted GEV model can be used to estimate two related quantities that describe the occurrence of *stress events*. On the one hand, we can estimate the size of a stress event that occurs with prescribed frequency (the *return-level* problem). On the other hand, we can estimate the frequency of a stress event that has a prescribed size (the *return-period* problem).

Definition 5.13 (return level). Let H denote the df of the true distribution of the n-block maximum. The k n-block return level is $r_{n,k} = q_{1-1/k}(H)$, i.e. the $(1 - 1/k)$-quantile of H.

The k n-block return level can be roughly interpreted as that level which is exceeded in one out of every k n-blocks on average. For example, the ten-trading-year return level $r_{260,10}$ is that level which is exceeded in one out of every ten years on average. (In the notation we assume that every year has 260 trading days, although this is only an average and there will be slight differences from year to year.) Using our fitted model we would estimate a return level by

$$\hat{r}_{n,k} = H^{-1}_{\hat{\xi},\hat{\mu},\hat{\sigma}}\left(1 - \frac{1}{k}\right) = \hat{\mu} + \frac{\hat{\sigma}}{\hat{\xi}}\left(\left(-\ln\left(1 - \frac{1}{k}\right)\right)^{-\hat{\xi}} - 1\right). \tag{5.7}$$

Definition 5.14 (return period). Let H denote the df of the true distribution of the n-block maximum. The return period of the event $\{M_n > u\}$ is given by $k_{n,u} = 1/\bar{H}(u)$.

Observe that the return period $k_{n,u}$ is defined in such a way that the $k_{n,u}$ n-block return level is u. In other words, in $k_{n,u}$ n-blocks we would expect to observe a single block in which the level u was exceeded. If there was a strong tendency for the extreme values to cluster, we might expect to see multiple exceedances of the level within that block. Assuming that H is the df of a GEV distribution and using our fitted model, we would estimate the return period by $\hat{k}_{n,u} = 1/\bar{H}_{\hat{\xi},\hat{\mu},\hat{\sigma}}(u)$.

Note that both $\hat{r}_{n,k}$ and $\hat{k}_{n,u}$ are simple functionals of the estimated parameters of the GEV distribution. As well as calculating point estimates for these quantities we should give confidence intervals that reflect the error in the parameter estimates of the GEV distribution. A good method is to base such confidence intervals on the likelihood ratio statistic, as described in Section A.3.5. To do this we reparametrize the GEV distribution in terms of the quantity of interest. For example, in the case of return level, let $\phi = H^{-1}_{\xi,\mu,\sigma}(1 - (1/k))$ and parametrize the GEV distribution by $\boldsymbol{\theta} = (\phi, \xi, \sigma)'$ rather than $\boldsymbol{\theta} = (\xi, \mu, \sigma)'$. The maximum likelihood estimate of ϕ is the estimate (5.7), and a confidence interval can be constructed according to the method in Section A.3.5 (see (A.28) in particular).

Example 5.15 (stress losses for S&P return data). We continue Example 5.12 by estimating the ten-year return level and the 20-semester return level based on data up to 16 October 1987, using (5.7) for the point estimate and the likelihood ratio method as described above to get confidence intervals. The point estimate of the ten-year return level is 4.3% with a 95% confidence interval of (3.4, 7.3); the point estimate of the 20-semester return level is 4.5% with a 95% confidence interval of (3.5, 7.1). Clearly, there is some uncertainty about the size of events of this frequency even with 28 years or 56 semesters of data.

The day after the end of our data set, 19 October 1987, was Black Monday. The index fell by the unprecedented amount of 20.5% in one day. This event is well outside our confidence interval for a ten-year loss. If we were to estimate a 50-year return level (an event beyond our experience if we have 28 years of data), then our point estimate would be 7.2 with a confidence interval of (4.8, 23.4), so the 1987 crash lies close to the upper boundary of our confidence interval for a much rarer event. But the 28 maxima are really too few to get a reliable estimate for an event as rare as the 50-year event.

If we turn the problem around and attempt to estimate the return period of a 20.5% loss, the point estimate is 1631 years (i.e. almost a two-millennium event) but the 95% confidence interval encompasses everything from 42 years to essentially never! The analysis of semesterly maxima gives only moderately more informative results: the point estimate is 1950 semesters; the confidence interval runs from 121 semesters to 3.0×10^6 semesters. In summary, on 16 October 1987 we simply did not have the data to say anything meaningful about an event of this magnitude. This illustrates the inherent difficulties of attempting to quantify events beyond our empirical experience.

Notes and Comments

The main source for this chapter is Embrechts, Klüppelberg and Mikosch (1997) (EKM). Further important texts on EVT include Gumbel (1958), Leadbetter, Lindgren and Rootzén (1983), Galambos (1987), Resnick (2008), Falk, Hüsler and Reiss (1994), Reiss and Thomas (1997), de Haan and Ferreira (2000), Coles (2001), Beirlant et al. (2004) and Resnick (2007).

The forms of the limit law for maxima were first studied by Fisher and Tippett (1928). The subject was brought to full mathematical fruition in the fundamental papers of Gnedenko (1941, 1943). The concept of the extremal index, which appears in the theory of maxima of stationary series, has a long history. The first mathematically precise definition seems to have been given by Leadbetter (1983). See also Leadbetter, Lindgren and Rootzén (1983) and Smith and Weissman (1994) for more details. The theory required to calculate the extremal index of an ARCH(1) process (as in Table 5.1) is found in de Haan et al. (1989) and also in EKM (pp. 473–480). For the GARCH(1, 1) process, consult Mikosch and Stărică (2000).

A further difficult task is the statistical estimation of the extremal index from time-series data under the assumption that these data do indeed come from a process with an extremal index. Two general methods known as the *blocks* and *runs* methods are described in Section 8.1.3 of EKM; these methods go back to work of Hsing (1991) and Smith and Weissman (1994). Although the estimators have been used in real-world data analyses (see, for example, Davison and Smith 1990), it remains true that the extremal index is a very difficult parameter to estimate accurately.

The maximum likelihood fitting of the GEV distribution is described by Hosking (1985) and Hosking, Wallis and Wood (1985). Consistency and asymptotic normality can be demonstrated for the case $\xi > -0.5$ using results in Smith (1985). An alternative method known as probability-weighted moments (PWM) has been proposed by Hosking, Wallis and Wood (1985) (see also pp. 321–323 of EKM). The analysis of block maxima in Examples 5.12 and 5.15 is based on McNeil (1998). Analyses of financial data using the block maxima method may also be found in Longin (1996), one of the earliest papers to apply EVT methodology to financial data.

5.2 Threshold Exceedances

The block maxima method discussed in Section 5.1.4 has the major defect that it is very wasteful of data; to perform our analyses we retain only the maximum losses in large blocks. For this reason it has been largely superseded in practice by methods based on threshold exceedances, where we use all the data that exceed a particular designated high level.

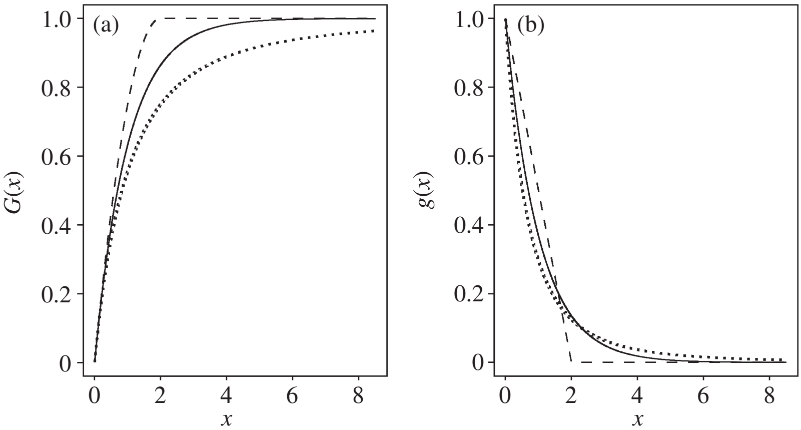

Figure 5.3. (a) Distribution function of the GPD in three cases: the solid line corresponds to $\xi = 0$ (exponential); the dotted line to $\xi = 0.5$ (a Pareto distribution); and the dashed line to $\xi = -0.5$ (Pareto type II). The scale parameter β is equal to 1 in all cases. (b) Corresponding densities.

5.2.1 Generalized Pareto Distribution

The main distributional model for exceedances over thresholds is the generalized Pareto distribution (GPD).

Definition 5.16 (GPD). The df of the GPD is given by

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \xi x/\beta)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-x/\beta), & \xi = 0, \end{cases} \tag{5.8}$$

where $\beta > 0$, and $x \geqslant 0$ when $\xi \geqslant 0$ and $0 \leqslant x \leqslant -\beta/\xi$ when $\xi < 0$. The parameters ξ and β are referred to, respectively, as the *shape* and *scale* parameters.

Like the GEV distribution in Definition 5.1, the GPD is generalized in the sense that it contains a number of special cases: when $\xi > 0$ the df $G_{\xi,\beta}$ is that of an ordinary Pareto distribution with $\alpha = 1/\xi$ and $\kappa = \beta/\xi$ (see Section A.2.8); when $\xi = 0$ we have an exponential distribution; when $\xi < 0$ we have a short-tailed, Pareto type II distribution. Moreover, as in the case of the GEV distribution, for fixed x the parametric form is continuous in ξ, so $\lim_{\xi \to 0} G_{\xi,\beta}(x) = G_{0,\beta}(x)$. The df and density of the GPD for various values of ξ and $\beta = 1$ are shown in Figure 5.3.

In terms of domains of attraction we have that $G_{\xi,\beta} \in \mathrm{MDA}(H_\xi)$ for all $\xi \in \mathbb{R}$. Note that, for $\xi > 0$ and $\xi < 0$, this assertion follows easily from the characterizations in Theorems 5.8 and 5.10. In the heavy-tailed case, $\xi > 0$, it may be easily verified that $E(X^k) = \infty$ for $k \geqslant 1/\xi$. The mean of the GPD is defined provided $\xi < 1$ and is

$$E(X) = \beta/(1 - \xi). \tag{5.9}$$

The role of the GPD in EVT is as a natural model for the *excess distribution* over a high threshold. We define this concept along with the *mean excess function*, which will also play an important role in the theory.

Definition 5.17 (excess distribution over threshold u). Let X be an rv with df F. The excess distribution over the threshold u has df

$$F_u(x) = P(X - u \leqslant x \mid X > u) = \frac{F(x+u) - F(u)}{1 - F(u)} \tag{5.10}$$

for $0 \leqslant x < x_F - u$, where $x_F \leqslant \infty$ is the right endpoint of F.

Definition 5.18 (mean excess function). The mean excess function of an rv X with finite mean is given by

$$e(u) = E(X - u \mid X > u). \tag{5.11}$$

The excess df F_u describes the distribution of the excess loss over the threshold u, given that u is exceeded. The mean excess function $e(u)$ expresses the mean of F_u as a function of u. In survival analysis the excess df is more commonly known as the residual life df—it expresses the probability that, say, an electrical component that has functioned for u units of time fails in the time period $(u, u + x]$. The mean excess function is known as the mean residual life function and gives the expected residual lifetime for components with different ages. For the special case of the GPD, the excess df and mean excess function are easily calculated.

Example 5.19 (excess distribution of exponential and GPD). If F is the df of an exponential rv, then it is easily verified that $F_u(x) = F(x)$ for all x, which is the famous *lack-of-memory* property of the exponential distribution—the residual lifetime of the aforementioned electrical component would be independent of the amount of time that component has already survived. More generally, if X has df $F = G_{\xi,\beta}$, then, using (5.10), the excess df is easily calculated to be

$$F_u(x) = G_{\xi,\beta(u)}(x), \quad \beta(u) = \beta + \xi u, \tag{5.12}$$

where $0 \leqslant x < \infty$ if $\xi \geqslant 0$ and $0 \leqslant x \leqslant -(\beta/\xi) - u$ if $\xi < 0$. The excess distribution remains a GPD with the same shape parameter ξ but with a scaling that grows linearly with the threshold u. The mean excess function of the GPD is easily calculated from (5.12) and (5.9) to be

$$e(u) = \frac{\beta(u)}{1 - \xi} = \frac{\beta + \xi u}{1 - \xi}, \tag{5.13}$$

where $0 \leqslant u < \infty$ if $0 \leqslant \xi < 1$ and $0 \leqslant u \leqslant -\beta/\xi$ if $\xi < 0$. It may be observed that the mean excess function is *linear in the threshold* u, which is a characterizing property of the GPD.

Example 5.19 shows that the GPD has a kind of stability property under the operation of calculating excess distributions. We now give a mathematical result that shows that the GPD is, in fact, a natural limiting excess distribution for many underlying loss distributions. The result can also be viewed as a characterization theorem for the maximum domain of attraction of the GEV distribution. In Section 5.1.2 we looked separately at characterizations for each of the three cases $\xi > 0$, $\xi = 0$ and $\xi < 0$; the following result offers a global characterization of MDA(H_ξ) for all ξ in terms of the limiting behaviour of excess distributions over thresholds.

Theorem 5.20 (Pickands–Balkema–de Haan). *We can find a (positive-measurable function) $\beta(u)$ such that*

$$\lim_{u \to x_F} \sup_{0 \leqslant x < x_F - u} |F_u(x) - G_{\xi,\beta(u)}(x)| = 0$$

if and only if $F \in \mathrm{MDA}(H_\xi)$, $\xi \in \mathbb{R}$.

Thus the distributions for which normalized maxima converge to a GEV distribution constitute a set of distributions for which the excess distribution converges to the GPD as the threshold is raised; moreover, the shape parameter of the limiting GPD for the excesses is the same as the shape parameter of the limiting GEV distribution for the maxima. We have already stated in Section 5.1.2 that essentially all the commonly used continuous distributions of statistics are in $\mathrm{MDA}(H_\xi)$ for some ξ, so Theorem 5.20 proves to be a very widely applicable result that essentially says that the GPD is *the canonical distribution* for modelling excess losses over high thresholds.

5.2.2 Modelling Excess Losses

We exploit Theorem 5.20 by assuming that we are dealing with a loss distribution $F \in \mathrm{MDA}(H_\xi)$ so that, for some suitably chosen high threshold u, we can model F_u by a generalized Pareto distribution. We formalize this with the following assumption.

Assumption 5.21. *Let F be a loss distribution with right endpoint x_F and assume that for some high threshold u we have $F_u(x) = G_{\xi,\beta}(x)$ for $0 \leqslant x < x_F - u$ and some $\xi \in \mathbb{R}$ and $\beta > 0$.*

This is clearly an idealization, since in practice the excess distribution will generally not be *exactly* GPD, but we use Assumption 5.21 to make a number of calculations in the following sections.

The method. Given loss data $X_1, \dots, X_n$ from F, a random number N_u will exceed our threshold u; it will be convenient to relabel these data $\tilde{X}_1, \dots, \tilde{X}_{N_u}$. For each of these exceedances we calculate the amount $Y_j = \tilde{X}_j - u$ of the excess loss. We wish to estimate the parameters of a GPD model by fitting this distribution to the N_u excess losses. There are various ways of fitting the GPD, including maximum likelihood (ML) and probability-weighted moments (PWM). The former method is more commonly used and is easy to implement if the excess data can be assumed to be realizations of independent rvs, since the joint density will then be a product of marginal GPD densities.

Writing $g_{\xi,\beta}$ for the density of the GPD, the log-likelihood may be easily calculated to be

$$\begin{aligned} \ln L(\xi, \beta; Y_1, \dots, Y_{N_u}) &= \sum_{j=1}^{N_u} \ln g_{\xi,\beta}(Y_j) \\ &= -N_u \ln \beta - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{N_u} \ln \left(1 + \xi \frac{Y_j}{\beta}\right), \qquad (5.14) \end{aligned}$$

which must be maximized subject to the parameter constraints that $\beta > 0$ and $1 + \xi Y_j/\beta > 0$ for all j. Solving the maximization problem yields a GPD model $G_{\hat{\xi},\hat{\beta}}$ for the excess distribution F_u.

Non-iid data. For insurance or operational risk data the iid assumption is often unproblematic, but this is clearly not true for time series of financial returns. If the data are serially dependent but show no tendency to give clusters of extreme values, then this might suggest that the underlying process has extremal index $\theta = 1$. In this case, asymptotic theory that we summarize in Section 5.3 suggests a limiting model for high-level threshold exceedances, in which exceedances occur according to a Poisson process and the excess loss amounts are iid generalized Pareto distributed. If extremal clustering is present, suggesting an extremal index $\theta < 1$ (as would be consistent with an underlying GARCH process), the assumption of independent excess losses is less satisfactory. The easiest approach is to neglect this problem and to consider the ML method to be a QML method, where the likelihood is misspecified with respect to the serial dependence structure of the data; we follow this course in this section. The point estimates should still be reasonable, although standard errors may be too small. In Section 5.3 we discuss threshold exceedances in non-iid data in more detail.

Excesses over higher thresholds. From the model we have fitted to the excess distribution over u, we can easily infer a model for the excess distribution over any higher threshold. We have the following lemma.

Lemma 5.22. *Under Assumption 5.21 it follows that $F_v(x) = G_{\xi,\beta+\xi(v-u)}(x)$ for any higher threshold $v \geqslant u$.*

Proof. We use (5.10) and the df of the GPD in (5.8) to infer that

$$\begin{aligned}\bar{F}_v(x) &= \frac{\bar{F}(v+x)}{\bar{F}(v)} = \frac{\bar{F}(u+(x+v-u))}{\bar{F}(u)} \frac{\bar{F}(u)}{\bar{F}(u+(v-u))} \\ &= \frac{\bar{F}_u(x+v-u)}{\bar{F}_u(v-u)} = \frac{\bar{G}_{\xi,\beta}(x+v-u)}{\bar{G}_{\xi,\beta}(v-u)} \\ &= \bar{G}_{\xi,\beta+\xi(v-u)}(x).\end{aligned}$$

□

Thus the excess distribution over higher thresholds remains a GPD with the same ξ parameter but a scaling that grows linearly with the threshold v. Provided that $\xi < 1$, the mean excess function is given by

$$e(v) = \frac{\beta + \xi(v-u)}{1-\xi} = \frac{\xi v}{1-\xi} + \frac{\beta - \xi u}{1-\xi}, \tag{5.15}$$

where $u \leqslant v < \infty$ if $0 \leqslant \xi < 1$ and $u \leqslant v \leqslant u - \beta/\xi$ if $\xi < 0$.

The linearity of the mean excess function (5.15) in v is commonly used as a diagnostic for data admitting a GPD model for the excess distribution. It forms the basis for the following simple graphical method for choosing an appropriate threshold.

Sample mean excess plot. For positive-valued loss data $X_1, \dots, X_n$ we define the *sample mean excess function* to be an empirical estimator of the mean excess function in Definition 5.18. The estimator is given by

$$e_n(v) = \frac{\sum_{i=1}^{n}(X_i - v)I_{\{X_i > v\}}}{\sum_{i=1}^{n} I_{\{X_i > v\}}}. \tag{5.16}$$

To study the sample mean excess function we generally construct the mean excess plot $\{(X_{i,n}, e_n(X_{i,n}))\colon 2 \leqslant i \leqslant n\}$, where $X_{i,n}$ denotes the upper (or descending) ith order statistic. If the data support a GPD model over a high threshold, then (5.15) suggests that this plot should become increasingly "linear" for higher values of v. A linear upward trend indicates a GPD model with positive shape parameter ξ; a plot tending towards the horizontal indicates a GPD with approximately zero shape parameter, or, in other words, an exponential excess distribution; a linear downward trend indicates a GPD with negative shape parameter.

These are the ideal situations, but in practice some experience is required to read mean excess plots. Even for data that are genuinely generalized Pareto distributed, the sample mean excess plot is seldom perfectly linear, particularly towards the right-hand end, where we are averaging a small number of large excesses. In fact, we often omit the final few points from consideration, as they can severely distort the picture. If we do see visual evidence that the mean excess plot becomes linear, then we might select as our threshold u a value towards the beginning of the linear section of the plot (see, in particular, Example 5.24).

Example 5.23 (Danish fire loss data). The Danish fire insurance data are a well-studied set of financial losses that neatly illustrate the basic ideas behind modelling observations that seem consistent with an iid model. The data set consists of 2156 fire insurance losses over 1 000 000 Danish kroner from 1980 to 1990 inclusive, expressed in units of 1 000 000 kroner. The loss figure represents a combined loss for a building and its contents, as well as in some cases a loss of business earnings; the losses are inflation adjusted to reflect 1985 values and are shown in Figure 5.4 (a).

The sample mean excess plot in Figure 5.4 (b) is in fact fairly "linear" over the entire range of the losses, and its upward slope leads us to expect that a GPD with positive shape parameter ξ could be fitted to the entire data set. However, there is some evidence of a "kink" in the plot below the value 10 and a "straightening out" of the plot above this value, so we have chosen to set our threshold at $u = 10$ and fit a GPD to excess losses above this threshold, in the hope of obtaining a model that is a good fit to the largest of the losses. The ML parameter estimates are $\hat{\xi} = 0.50$ and $\hat{\beta} = 7.0$ with standard errors 0.14 and 1.1, respectively. Thus the model we have fitted is essentially a very heavy-tailed, infinite-variance model. A picture of the fitted GPD model for the excess distribution $\hat{F}_u(x - u)$ is also given in Figure 5.4 (c), superimposed on points plotted at empirical estimates of the excess probabilities for each loss; note the good correspondence between the empirical estimates and the GPD curve.

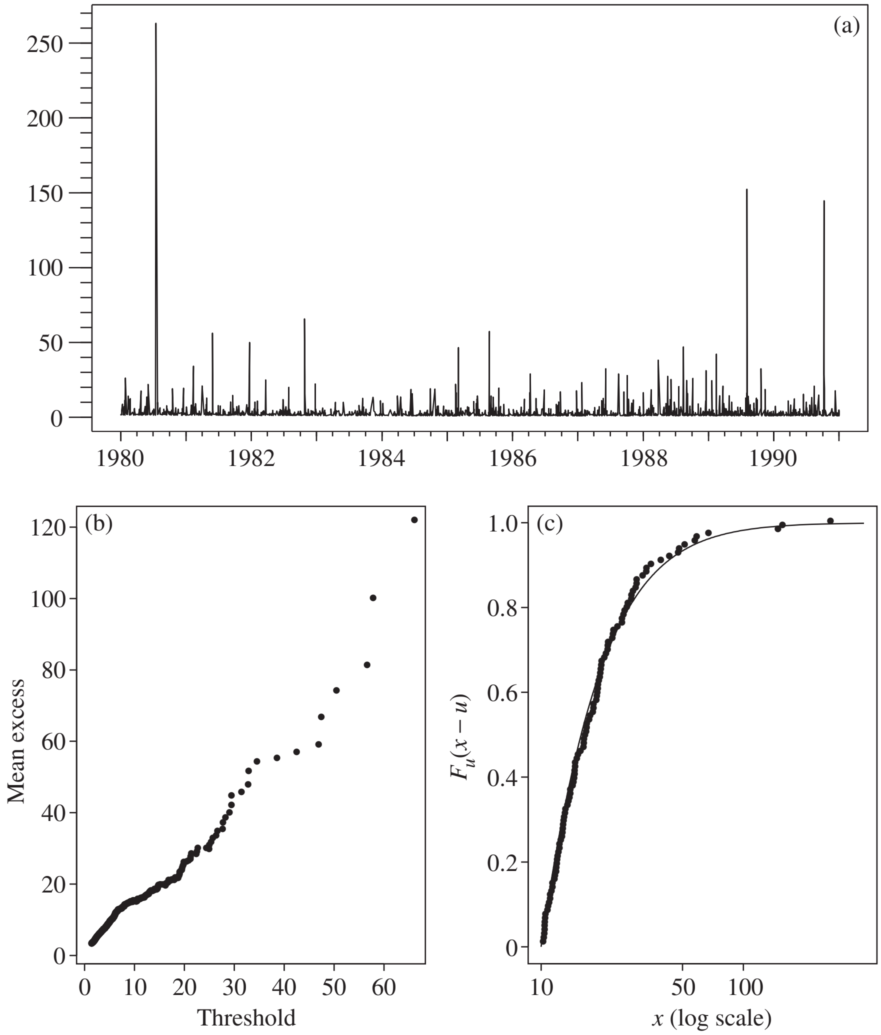

Figure 5.4. (a) Time-series plot of the Danish data. (b) Sample mean excess plot. (c) Empirical distribution of excesses and fitted GPD. See Example 5.23 for full details.

In insurance we might use the model to estimate the expected size of the insurance loss, given that it enters a given insurance *layer*. Thus we can estimate the expected loss size given exceedance of the threshold of 10 000 000 kroner or of any other higher threshold by using (5.15) with the appropriate parameter estimates.

Example 5.24 (AT&T weekly loss data). Suppose we have an investment in AT&T stock and want to model weekly losses in value using an unconditional approach. If X_t denotes the weekly log-return, then the percentage loss in value of our position over a week is given by $L_t = 100(1 - \exp(X_t))$, and data on this loss for the 521 complete weeks in the period 1991–2000 are shown in Figure 5.5 (a).

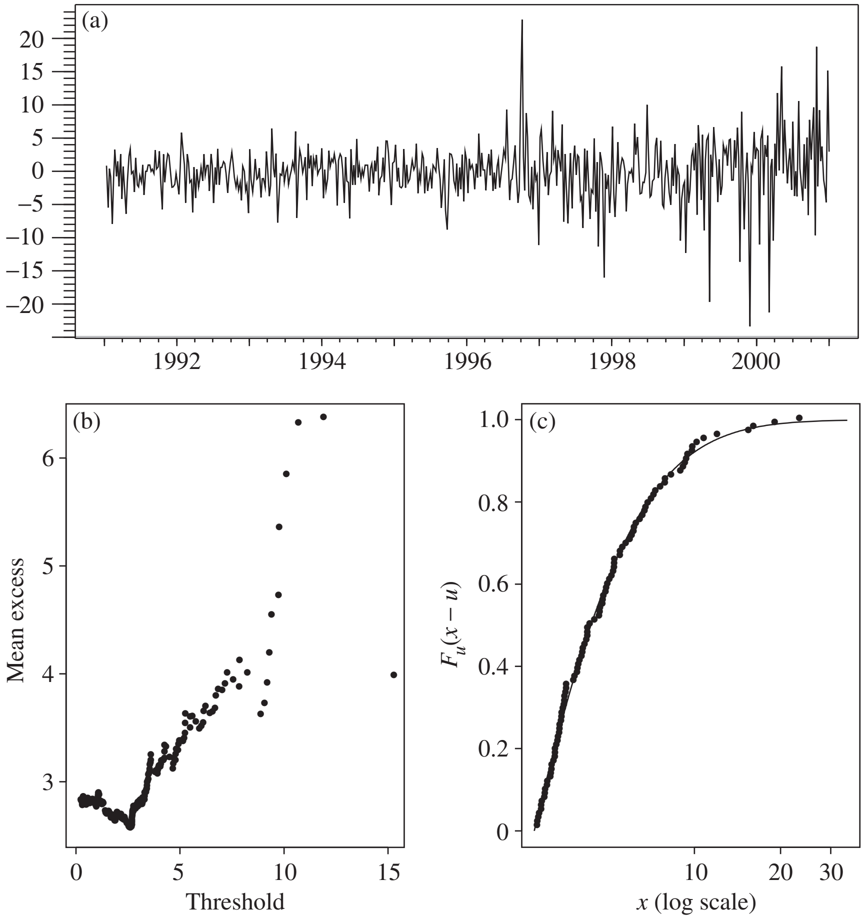

Figure 5.5. (a) Time-series plot of AT&T weekly percentage loss data. (b) Sample mean excess plot. (c) Empirical distribution of excesses and fitted GPD. See Example 5.24 for full details.

A sample mean excess plot of the positive loss values is shown in Figure 5.5 (b) and this suggests that a threshold can be found above which a GPD approximation to the excess distribution should be possible. We have chosen to position the threshold at a loss value of 2.75% and this gives 102 exceedances.

We observed in Section 3.1 that monthly AT&T return data over the period 1993–2000 do not appear consistent with a strict white noise hypothesis, so the issue of whether excess losses can be modelled as independent is relevant. This issue is taken up in Section 5.3 but for the time being we ignore it and implement a standard ML approach to estimating the parameters of a GPD model for the excess distribution; we obtain the estimates $\hat{\xi} = 0.22$ and $\hat{\beta} = 2.1$ with standard errors 0.13 and 0.34, respectively. Thus the model we have fitted is a model that is close to having an infinite fourth moment. A picture of the fitted GPD model for the excess distribution $\hat{F}_u(x-u)$ is also given in Figure 5.5 (c), superimposed on points plotted at empirical estimates of the excess probabilities for each loss.

5.2.3 Modelling Tails and Measures of Tail Risk

In this section we describe how the GPD model for the excess losses is used to estimate the tail of the underlying loss distribution F and associated risk measures. To make the necessary theoretical calculations we again make Assumption 5.21.

Tail probabilities and risk measures. We observe firstly that under Assumption 5.21 we have, for $x \geqslant u$,

$$\begin{aligned}\bar{F}(x) &= P(X > u)P(X > x \mid X > u)\\ &= \bar{F}(u)P(X - u > x - u \mid X > u)\\ &= \bar{F}(u)\bar{F}_u(x-u)\\ &= \bar{F}(u)\biggl(1 + \xi\frac{x-u}{\beta}\biggr)^{-1/\xi},\end{aligned} \tag{5.17}$$

which, if we know $F(u)$, gives us a formula for tail probabilities. This formula may be inverted to obtain a high quantile of the underlying distribution, which we interpret as a VaR. For $\alpha \geqslant F(u)$ we have that VaR is equal to

$$\mathrm{VaR}_\alpha = q_\alpha(F) = u + \frac{\beta}{\xi}\biggl(\biggl(\frac{1-\alpha}{\bar{F}(u)}\biggr)^{-\xi} - 1\biggr). \tag{5.18}$$

Assuming that $\xi < 1$, the associated expected shortfall can be calculated easily from (2.22) and (5.18). We obtain

$$\mathrm{ES}_\alpha = \frac{1}{1-\alpha}\int_\alpha^1 q_x(F)\,\mathrm{d}x = \frac{\mathrm{VaR}_\alpha}{1-\xi} + \frac{\beta - \xi u}{1-\xi}. \tag{5.19}$$

Note that Assumption 5.21 and Lemma 5.22 imply that excess losses above VaR_α have a GPD distribution satisfying $F_{\mathrm{VaR}_\alpha} = G_{\xi,\beta+\xi(\mathrm{VaR}_\alpha - u)}$. The expected shortfall estimator in (5.19) can also be obtained by adding the mean of this distribution to VaR_α, i.e. $\mathrm{ES}_\alpha = \mathrm{VaR}_\alpha + e(\mathrm{VaR}_\alpha)$, where $e(\mathrm{VaR}_\alpha)$ is given in (5.15). It is interesting to look at how the *ratio* of the two risk measures behaves for large values of the quantile probability α. It is easily calculated from (5.18) and (5.19) that

$$\lim_{\alpha\to 1}\frac{\mathrm{ES}_\alpha}{\mathrm{VaR}_\alpha} = \begin{cases}(1-\xi)^{-1}, & 0 \leqslant \xi < 1,\\ 1, & \xi < 0,\end{cases} \tag{5.20}$$

so the shape parameter ξ of the GPD effectively determines the ratio when we go far enough out into the tail.

Estimation in practice. We note that, under Assumption 5.21, tail probabilities, VaRs and expected shortfalls are all given by formulas of the form $g(\xi, \beta, \bar{F}(u))$. Assuming that we have fitted a GPD to excess losses over a threshold u, as described in Section 5.2.2, we estimate these quantities by first replacing ξ and β in formulas (5.17)–(5.19) by their estimates. Of course, we also require an estimate of $\bar{F}(u)$ and here we take the simple empirical estimator N_u/n. In doing this, we are implicitly assuming that there is a sufficient proportion of sample values above the threshold u to estimate $\bar{F}(u)$ reliably. However, we hope to gain over the empirical method by using a kind of extrapolation based on the GPD for more extreme tail probabilities and risk measures. For tail probabilities we obtain an estimator, first proposed by Smith (1987), of the form

$$\hat{\bar{F}}(x) = \frac{N_u}{n}\left(1 + \hat{\xi}\frac{x-u}{\hat{\beta}}\right)^{-1/\hat{\xi}}, \tag{5.21}$$

which we stress is only valid for $x \geqslant u$. For $\alpha \geqslant 1 - N_u/n$ we obtain analogous point estimators of VaR_α and ES_α from (5.18) and (5.19).

Of course, we would also like to obtain confidence intervals. If we have taken the likelihood approach to estimating ξ and β, then it is quite easy to give confidence intervals for $g(\hat{\xi}, \hat{\beta}, N_u/n)$ that take into account the uncertainty in $\hat{\xi}$ and $\hat{\beta}$, but neglect the uncertainty in N_u/n as an estimator of $\bar{F}(u)$. We use the approach described at the end of Section 5.1.4 for return levels, whereby the GPD model is reparametrized in terms of $\phi = g(\xi, \beta, N_u/n)$, and a confidence interval for $\hat{\phi}$ is constructed based on the likelihood ratio test as in Section A.3.5.

Example 5.25 (risk measures for AT&T loss data). Suppose we have fitted a GPD model to excess weekly losses above the threshold $u = 2.75\%$, as in Example 5.24. We use this model to obtain estimates of the 99% VaR and expected shortfall of the underlying weekly loss distribution. The essence of the method is displayed in Figure 5.6; this is a plot of estimated tail probabilities on logarithmic axes, with various dotted lines superimposed to indicate the estimation of risk measures and associated confidence intervals. The points on the graph are the 102 threshold exceedances and are plotted at y-values corresponding to the tail of the empirical distribution function; the smooth curve running through the points is the tail estimator (5.21).

Estimation of the 99% quantile amounts to determining the point of intersection of the tail estimation curve and the horizontal line $\bar{F}(x) = 0.01$ (not marked on the graph); the first vertical dotted line shows the quantile estimate. The horizontal dotted line aids in the visualization of a 95% confidence interval for the VaR estimate; the degree of confidence is shown on the alternative y-axis to the right of the plot. The boundaries of a 95% confidence interval are obtained by determining the two points of intersection of this horizontal line with the dotted curve, which is a profile likelihood curve for the VaR as a parameter of the GPD model and is constructed using likelihood ratio test arguments as in Section A.3.5. Dropping the horizontal line to the 99% mark would correspond to constructing a 99% confidence interval for the estimate of the 99% VaR. The point estimate and the 95% confidence interval for the 99% quantile are estimated to be 11.7% and (9.6, 16.1).

The second vertical line on the plot shows the point estimate of the 99% expected shortfall. A 95% confidence interval is determined from the dotted horizontal line and its points of intersection with the second dotted curve. The point estimate and the 95% confidence interval are 17.0% and (12.7, 33.6). Note that if we take the ratio of the point estimates of the shortfall and the VaR, we get $17/11.7 \approx 1.45$, which is larger than the asymptotic ratio $(1 - \hat{\xi})^{-1} = 1.29$ suggested by (5.20); this is generally the case at finite levels and is explained by the second term in (5.19) being a non-negligible positive quantity.

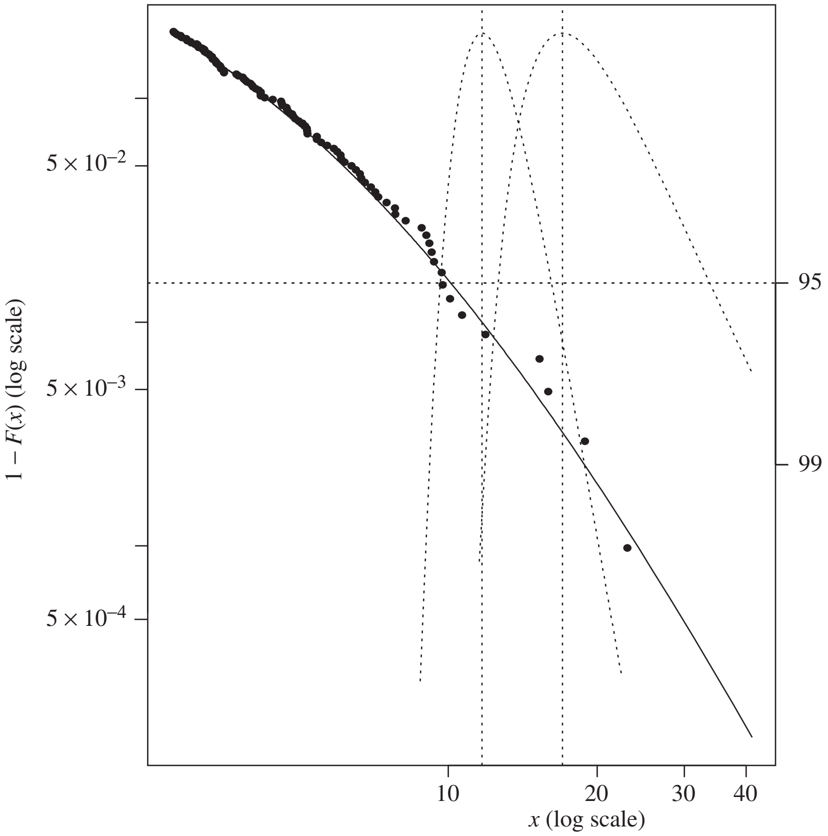

Figure 5.6. The smooth curve through the points shows the estimated tail of the AT&T weekly percentage loss data using the estimator (5.21). Points are plotted at empirical tail probabilities calculated from empirical df. The vertical dotted lines show estimates of 99% VaR and expected shortfall. The other curves are used in the construction of confidence intervals. See Example 5.25 for full details.

Before leaving the topic of GPD tail modelling it is clearly important to see how sensitive our risk-measure estimates are to the choice of the threshold. Hitherto, we have considered single choices of threshold u and looked at a series of incremental calculations that always build on the same GPD model for excesses over that threshold. We would hope that there is some robustness to our inference for different choices of threshold.

Example 5.26 (varying the threshold). In the case of the AT&T weekly loss data the influence of different thresholds is investigated in Figure 5.7. Given the importance of the ξ parameter in determining the weight of the tail and the relationship between quantiles and expected shortfalls, we first show how estimates of ξ vary as we consider a series of thresholds that give us between 20 and 150 exceedances. In fact, the estimates remain fairly constant around a value of approximately 0.2; a symmetric 95% confidence interval constructed from the standard error estimate is also shown, and it indicates how the uncertainty about the parameter

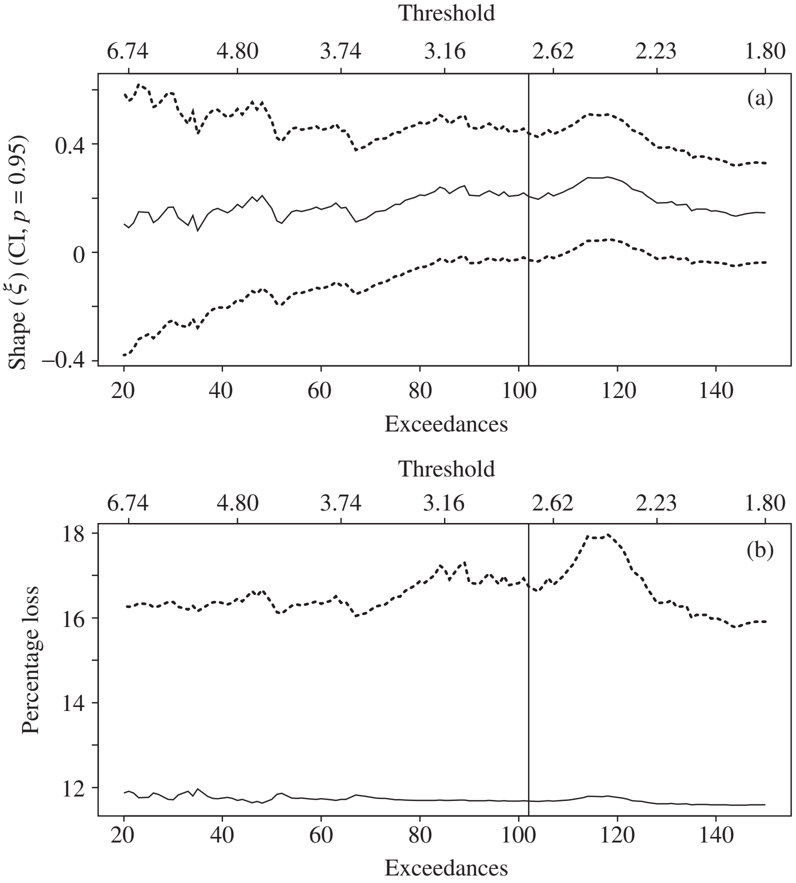

Figure 5.7. (a) Estimate of ξ for different thresholds u and numbers of exceedances N_u, together with a 95% confidence interval based on the standard error. (b) Associated point estimates of the 99% VaR (solid line) and the expected shortfall (dotted line). See Example 5.26 for commentary.

value decreases as the threshold is lowered or the number of threshold exceedances is increased.

Point estimates of the 99% VaR and expected shortfall estimates are also shown. The former remain remarkably constant around 12%, while the latter show modest variability that essentially tracks the variability of the ξ estimate. These pictures provide some reassurance that different thresholds do not lead to drastically different conclusions. We return to the issue of threshold choice again in Section 5.2.5.

5.2.4 The Hill Method

The GPD method is not the only way to estimate the tail of a distribution and, as an alternative, we describe in this section the well-known Hill approach to modelling the tails of heavy-tailed distributions.

Estimating the tail index. For this method we assume that the underlying loss distribution is in the maximum domain of attraction of the Fréchet distribution so that, by Theorem 5.8, it has a tail of the form

$$\bar{F}(x) = x^{-\alpha} L(x) \tag{5.22}$$

for a slowly varying function L (see Definition 5.7) and a positive parameter α. Traditionally, in the Hill approach, interest centres on the *tail index* α, rather than its reciprocal ξ, which appears in (5.3). The goal is to find an estimator of α based on identically distributed data $X_1, \dots, X_n$.

The Hill estimator can be derived in various ways (see EKM, pp. 330–336). Perhaps the most elegant is to consider the mean excess function of the generic logarithmic loss $\ln X$, where X is an rv with df (5.22). Writing e^* for the mean excess function of $\ln X$ and using integration by parts we find that

$$\begin{aligned} e^*(\ln u) &= E(\ln X - \ln u \mid \ln X > \ln u) \\ &= \frac{1}{\bar{F}(u)} \int_u^\infty (\ln x - \ln u)\,\mathrm{d}F(x) \\ &= \frac{1}{\bar{F}(u)} \int_u^\infty \frac{\bar{F}(x)}{x}\,\mathrm{d}x \\ &= \frac{1}{\bar{F}(u)} \int_u^\infty L(x) x^{-(\alpha+1)}\,\mathrm{d}x. \end{aligned}$$

For u sufficiently large, the slowly varying function $L(x)$ for $x \geqslant u$ can essentially be treated as a constant and taken outside the integral. More formally, using Karamata's Theorem (see Section A.1.4), we get, for $u \to \infty$,

$$e^*(\ln u) \sim \frac{L(u) u^{-\alpha} \alpha^{-1}}{\bar{F}(u)} = \alpha^{-1},$$

so $\lim_{u\to\infty} \alpha e^*(\ln u) = 1$. We expect to see similar tail behaviour in the sample mean excess function e_n^* (see (5.16)) constructed from the log observations. That is, we expect that $e_n^*(\ln X_{k,n}) \approx \alpha^{-1}$ for n large and k sufficiently small, where $X_{n,n} \leqslant \cdots \leqslant X_{1,n}$ are the order statistics as usual. Evaluating $e_n^*(\ln X_{k,n})$ gives us the estimator $\hat{\alpha}^{-1} = ((k-1)^{-1} \sum_{j=1}^{k-1} \ln X_{j,n} - \ln X_{k,n})$. The standard form of the Hill estimator is obtained by a minor modification:

$$\hat{\alpha}_{k,n}^{(\mathrm{H})} = \left(\frac{1}{k} \sum_{j=1}^{k} \ln X_{j,n} - \ln X_{k,n} \right)^{-1}, \quad 2 \leqslant k \leqslant n. \tag{5.23}$$

The Hill estimator is one of the best-studied estimators in the EVT literature. The asymptotic properties (consistency, asymptotic normality) of this estimator (as sample size $n \to \infty$, number of extremes $k \to \infty$ and the so-called tail-fraction $k/n \to 0$) have been extensively investigated under various assumed models for the data, including ARCH and GARCH (see Notes and Comments). We concentrate on the use of the estimator in practice and, in particular, on its performance relative to the GPD estimation approach.

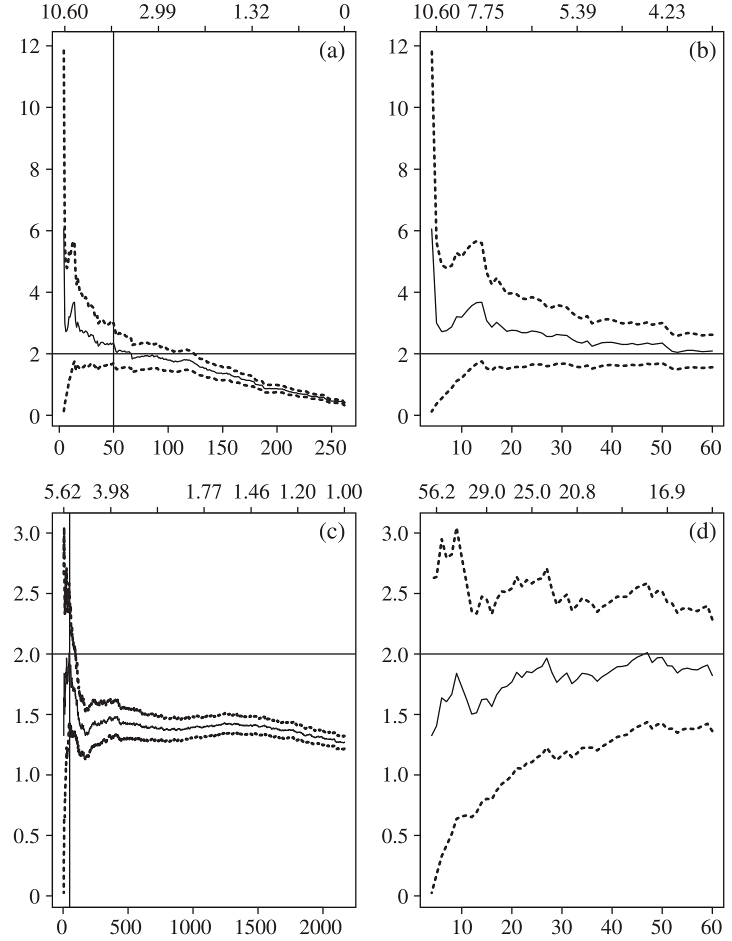

Figure 5.8. Hill plots showing estimates of the tail index $\alpha = 1/\xi$ for (a), (b) the AT&T weekly percentages losses and (c), (d) the Danish fire loss data. Parts (b) and (d) are expanded versions of sections of (a) and (c) showing Hill estimates based on up to 60 order statistics.

When the data are from a distribution with a tail that is close to a perfect power function, the Hill estimator is often a good estimator of α, or its reciprocal ξ. In practice, the general strategy is to plot Hill estimates for various values of k. This gives the Hill plot $\{(k, \hat{\alpha}^{(\mathrm{H})}_{k,n}) : k = 2, \dots, n\}$. We hope to find a stable region in the Hill plot where estimates constructed from different numbers of order statistics are quite similar.

Example 5.27 (Hill plots). We construct Hill plots for the Danish fire data of Example 5.23 and the weekly percentage loss data (positive values only) of Example 5.24 (shown in Figure 5.8).

It is very easy to construct the Hill plot for all possible values of k, but it can be misleading to do so; practical experience (see Example 5.28) suggests that the best choices of k are relatively small: say, 10–50 order statistics in a sample of size 1000.

For this reason we have enlarged sections of the Hill plots showing the estimates obtained for values of k less than 60.

For the Danish data, the estimates of α obtained are between 1.5 and 2, suggesting ξ estimates between 0.5 and 0.67, all of which correspond to infinite-variance models for these data. Recall that the estimate derived from our GPD model in Example 5.23 was $\hat{\xi} = 0.50$. For the AT&T data, there is no particularly stable region in the plot. The α estimates based on $k = 2, \dots, 60$ order statistics mostly range from 2 to 4, suggesting a ξ value in the range 0.25–0.5, which is larger than the values estimated in Example 5.26 with a GPD model.

Example 5.27 shows that the interpretation of Hill plots can be difficult. In practice, various deviations from the ideal situation can occur. If the data do not come from a distribution with a regularly varying tail, the Hill method is really not appropriate and Hill plots can be very misleading. Serial dependence in the data can also spoil the performance of the estimator, although this is also true for the GPD estimator. EKM contains a number of Hill "horror plots" based on simulated data illustrating the issues that arise (see Notes and Comments).

Hill-based tail estimates. For the risk-management applications of this book we are less concerned with estimating the tail index of heavy-tailed data and more concerned with tail and risk-measure estimates. We give a heuristic argument for a standard tail estimator based on the Hill approach. We assume a tail model of the form $\bar{F}(x) = Cx^{-\alpha}$, $x \geqslant u > 0$, for some high threshold u; in other words, we replace the slowly varying function by a constant for sufficiently large x. For an appropriate value of k the tail index α is estimated by $\hat{\alpha}^{(\mathrm{H})}_{k,n}$ and the threshold u is replaced by $X_{k,n}$ (or $X_{(k+1),n}$ in some versions); it remains to estimate C. Since C can be written as $C = u^{\alpha}\bar{F}(u)$, this is equivalent to estimating $\bar{F}(u)$, and the obvious empirical estimator is k/n (or $(k-1)/n$ in some versions). Putting these ideas together gives us the *Hill tail estimator* in its standard form:

$$\hat{\bar{F}}(x) = \frac{k}{n}\left(\frac{x}{X_{k,n}}\right)^{-\hat{\alpha}^{(\mathrm{H})}_{k,n}}, \quad x \geqslant X_{k,n}. \tag{5.24}$$

Writing the estimator in this way emphasizes the way it is treated mathematically. For any pair k and n, both the Hill estimator and the associated tail estimator are treated as functions of the k upper-order statistics from the sample of size n. Obviously, it is possible to invert this estimator to get a quantile estimator and it is also possible to devise an estimator of expected shortfall using arguments about regularly varying tails.

The GPD-based tail estimator (5.21) is usually treated as a function of a random number N_u of upper-order statistics for a fixed threshold u. The different presentation of these estimators in the literature is a matter of convention and we can easily recast both estimators in a similar form. Suppose we rewrite (5.24) in the notation of (5.21) by substituting $\hat{\xi}^{(\mathrm{H})}$, u and N_u for $1/\hat{\alpha}^{(\mathrm{H})}_{k,n}$, $X_{k,n}$ and k, respectively. We get

$$\hat{\bar{F}}(x) = \frac{N_u}{n}\left(1 + \hat{\xi}^{(\mathrm{H})}\frac{x-u}{\hat{\xi}^{(\mathrm{H})}u}\right)^{-1/\hat{\xi}^{(\mathrm{H})}}, \quad x \geqslant u.$$

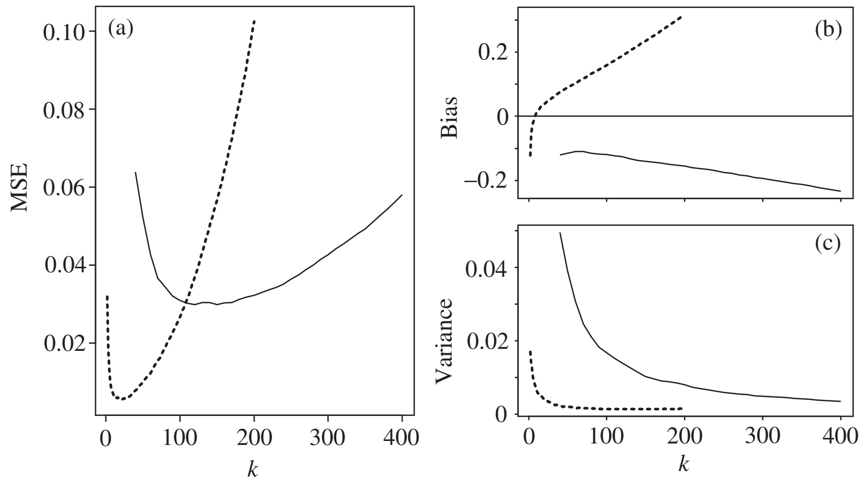

Figure 5.9. Comparison of (a) estimated MSE, (b) bias and (c) variance for the Hill (dotted line) and GPD (solid line) estimators of ξ, the reciprocal of the tail index, as a function of k (or N_u), the number of upper-order statistics from a sample of 1000 t-distributed data with four degrees of freedom. See Example 5.28 for details.

This estimator lacks the additional scaling parameter β in (5.21) and tends not to perform as well, as is shown in simulated examples in the next section.

5.2.5 Simulation Study of EVT Quantile Estimators

First we consider estimation of ξ and then estimation of the high quantile VaR_α. In both cases estimators are compared using mean squared errors (MSEs); we recall that the MSE of an estimator $\hat{\theta}$ of a parameter θ is given by $\mathrm{MSE}(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = (E(\hat{\theta} - \theta))^2 + \mathrm{var}(\hat{\theta})$, and thus has the well-known decomposition into *squared bias* plus *variance*. A good estimator should keep both the bias term $E(\hat{\theta} - \theta)$ and the variance term $\mathrm{var}(\hat{\theta})$ small.

Since analytical evaluation of bias and variance is not possible, we calculate Monte Carlo estimates by simulating 1000 data sets in each experiment. The parameters of the GPD are determined in all cases by ML; PWM, the main alternative, gives slightly different results, but the conclusions are similar.

We calculate estimates using the Hill method and the GPD method based on different numbers of upper-order statistics (or differing thresholds) and try to determine the choice of k (or N_u) that is most appropriate for a sample of size n. In the case of estimating VaR we also compare the EVT estimators with the simple empirical quantile estimator.

Example 5.28 (Monte Carlo experiment). We assume that we have a sample of 1000 iid data from a t distribution with four degrees of freedom and want to estimate ξ, the reciprocal of the tail index, which in this case has the true value 0.25. (This is demonstrated in Example 16.1.) The Hill estimate is constructed for k values in the range $\{2, \dots, 200\}$, and the GPD estimate is constructed for k (or N_u) values in $\{30, 40, 50, \dots, 400\}$. The results are shown in Figure 5.9.

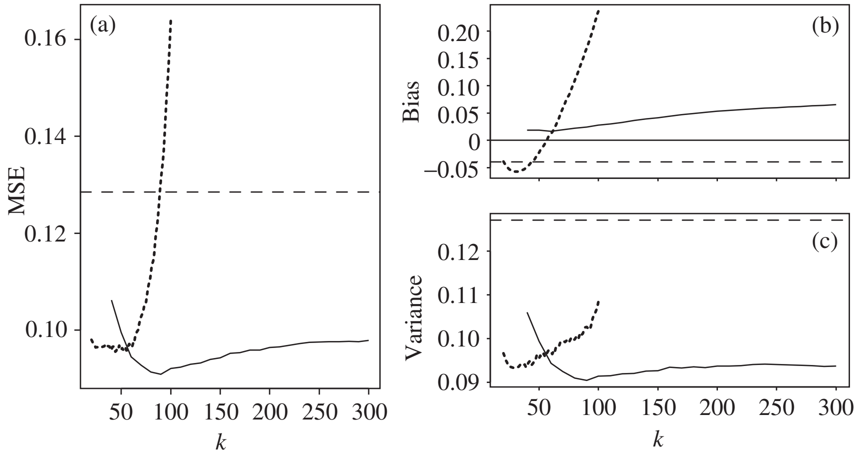

Figure 5.10. Comparison of (a) estimated MSE, (b) bias and (c) variance for the Hill (dotted line) and GPD (solid line) estimators of $\mathrm{VaR}_{0.99}$, as a function of k (or N_u), the number of upper-order statistics from a sample of 1000 t-distributed data with four degrees of freedom. The dashed line also shows results for the (threshold-independent) empirical quantile estimator. See Example 5.28 for details.

The t distribution has a well-behaved regularly varying tail and the Hill estimator gives better estimates of ξ than the GPD method, with an optimal value of k around 20–30. The variance plot shows where the Hill method gains over the GPD method; the variance of the GPD-based estimator is much higher than that of the Hill estimator for small numbers of order statistics. The magnitudes of the biases are closer together, with the Hill method tending to overestimate ξ and the GPD method tending to underestimate it. If we were to use the GPD method, the optimal choice of threshold would be one giving 100–150 exceedances.

The conclusions change when we attempt to estimate the 99% VaR; the results are shown in Figure 5.10. The Hill method has a negative bias for low values of k but a rapidly growing positive bias for larger values of k; the GPD estimator has a positive bias that grows much more slowly; the empirical method has a negative bias. The GPD attains its lowest MSE value for a value of k around 100, but, more importantly, the MSE is very robust to the choice of k because of the slow growth of the bias. The Hill method performs well for $20 \leqslant k \leqslant 75$ (we only use k values that lead to a quantile estimate beyond the effective threshold $X_{k,n}$) but then deteriorates rapidly. Both EVT methods obviously outperform the empirical quantile estimator. Given the relative robustness of the GPD-based tail estimator to changes in k, the issue of threshold choice for this estimator seems less critical than for the Hill method.

5.2.6 Conditional EVT for Financial Time Series

The GPD method when applied to threshold exceedances in a financial return series (as in Examples 5.24 and 5.25) gives us risk-measure estimates for the stationary (or unconditional) distribution of the underlying time series. We now consider a simple adaptation of the GPD method that allows us to obtain risk-measure estimates for the conditional (or one-step-ahead forecast) distribution of a time series. This adaptation uses the GARCH model and related ideas in Chapter 4 and will be applied in the market-risk context in Chapter 9.

We use the notation developed for prediction or forecasting in Sections 4.1.5 and 4.2.5. Let $X_{t-n+1}, \dots, X_t$ denote a series of *negative* log-returns and assume that these come from a strictly stationary time-series process (X_t). Assume further that the process satisfies equations of the form $X_t = \mu_t + \sigma_t Z_t$, where μ_t and σ_t are $\mathcal{F}_{t-1}$-measurable and (Z_t) are iid innovations with some unknown df F_Z; an example would be an ARMA model with GARCH errors.

We want to obtain VaR and expected shortfall estimates for the conditional distribution $F_{X_{t+1}|\mathcal{F}_t}$, and in Section 4.2.5 we showed that these risk measures are given by the equations

$$\mathrm{VaR}_\alpha^t = \mu_{t+1} + \sigma_{t+1} q_\alpha(Z), \qquad \mathrm{ES}_\alpha^t = \mu_{t+1} + \sigma_{t+1}\,\mathrm{ES}_\alpha(Z),$$

where we write Z for a generic rv with df F_Z.

These equations suggest an estimation method as follows. We first fit an ARMA–GARCH model by the QML procedure of Section 4.2.4 (since we do not wish to assume a particular innovation distribution) and use this to estimate μ_{t+1} and σ_{t+1}. As an alternative we could use EWMA volatility forecasting. To estimate $q_\alpha(Z)$ and $\mathrm{ES}_\alpha(Z)$ we essentially apply the GPD tail estimation procedure to the innovation distribution F_Z. To get round the problem that we do not observe data directly from the innovation distribution, we treat the residuals from the GARCH analysis as our data and apply the GPD tail estimation method of Section 5.2.3 to the residuals. In particular, we estimate $q_\alpha(Z)$ and $\mathrm{ES}_\alpha(Z)$ using the VaR and expected shortfall formulas in (5.18) and (5.19).

Notes and Comments

The ideas behind the important Theorem 5.20, which underlies GPD modelling, may be found in Pickands (1975) and Balkema and de Haan (1974). Important papers developing the technique in the statistical literature are Davison (1984) and Davison and Smith (1990). The estimation of the parameters of the GPD, both by ML and by the method of probability-weighted moments, is discussed in Hosking and Wallis (1987). The tail estimation formula (5.21) was suggested by Smith (1987), and the theoretical properties of this estimator for iid data in the domain of attraction of an extreme value distribution are extensively investigated in that paper. The Danish fire loss example is taken from McNeil (1997).

The Hill estimator goes back to Hill (1975) (see also Hall 1982). The theoretical properties for dependent data, including linear processes with heavy-tailed innovations and ARCH and GARCH processes, were investigated by Resnick and Stărică (1995, 1996). The idea of smoothing the estimator is examined in Resnick and Stărică (1997) and Resnick (1997). For Hill “horror plots”, showing situations when the Hill estimator delivers particularly poor estimates of the tail index, see pp. 194, 270 and 343 of EKM.

Alternative estimators based on order statistics include the estimator of Pickands (1975), which is also discussed in Dekkers and de Haan (1989), and the DEdH estimator of Dekkers, Einmahl and de Haan (1989). This latter estimator is used as the basis of a quantile estimator in de Haan and Rootzén (1993). Both the Pickands and DEdH estimators are designed to estimate general ξ in the extreme value limit (in contrast to the Hill estimator, which is designed for positive ξ); in empirical studies the DEdH estimator seems to work better than the Pickands estimator. The issue of the optimal number of order statistics in such estimators is taken up in a series of papers by Dekkers and de Haan (1993) and Daníelsson et al. (2001a). A method is proposed that is essentially based on the bootstrap approach to estimating mean squared error discussed in Hall (1990). A review paper relevant for applications to insurance and finance is Matthys and Beirlant (2000).

Analyses of the tails of financial data using methods based on the Hill estimator can be found in Koedijk, Schafgans and de Vries (1990), Lux (1996) and various papers by Daníelsson and de Vries (1997a,b,c). The conditional EVT method was developed in McNeil and Frey (2000); a Monte Carlo method using the GPD model to estimate risk measures for the h-day loss distribution is also described. See also Gençay, Selçuk and Ulugülyağci (2003), Gençay and Selçuk (2004) and Chavez-Demoulin, Embrechts and Sardy (2014) for interesting applications of EVT methodology to financial time series and VaR estimation.

5.3 Point Process Models

In our discussion of threshold models in Section 5.2 we considered only the magnitude of excess losses over high thresholds. In this section we consider exceedances of thresholds as events in time and use a point process approach to model the occurrence of these events. We begin by looking at the case of regularly spaced iid data and discuss the well-known POT model for the occurrence of extremes in such data; this model elegantly subsumes the models for maxima and the GPD models for excess losses that we have so far described.

However, the assumptions of the standard POT model are typically violated by financial return series, because of the kind of serial dependence that volatility clustering generates in such data. Our ultimate aim is to find more general point process models to describe the occurrence of extreme values in financial time series, and we find suitable candidates in the class of self-exciting point processes. These models are of a dynamic nature and can be used to estimate conditional VaRs; they offer an interesting alternative to the conditional EVT approach of Section 5.2.6, with the advantage that no pre-whitening of data with GARCH processes is required.

The following section gives an idea of the theory behind the POT model, but it may be skipped by readers who are content to go directly to a description of the standard POT model in Section 5.3.2.

5.3.1 Threshold Exceedances for Strict White Noise

Consider a strict white noise process $(X_i)_{i \in \mathbb{N}}$ representing financial losses. While we discuss the theory for iid variables for simplicity, the results we describe also hold for dependent processes with extremal index $\theta = 1$, i.e. processes where extreme values show no tendency to cluster (see Section 5.1.3 for examples of such processes).

Throughout this section we assume that the common loss distribution is in the maximum domain of attraction of an extreme value distribution (MDA(H_ξ)) so that (5.1) holds for the non-degenerate limiting distribution H_ξ and normalizing sequences c_n and d_n. From (5.1) it follows, by taking logarithms, that for any fixed x we have

$$\lim_{n\to\infty} n \ln F(c_n x + d_n) = \ln H_\xi(x). \tag{5.25}$$

Throughout this section we also consider a sequence of thresholds $(u_n(x))$ defined by $u_n(x) := c_n x + d_n$ for some fixed value of x. By noting that $-\ln y \sim 1 - y$ as $y \to 1$, we can infer from (5.25) that $n\bar{F}(u_n(x)) \sim -n \ln F(u_n(x)) \to -\ln H_\xi(x)$ as $n \to \infty$ for this sequence of thresholds.

The number of losses in the sample $X_1, \dots, X_n$ exceeding the threshold $u_n(x)$ is a binomial rv, $N_{u_n(x)} \sim B(n, \bar{F}(u_n(x)))$, with expectation $n\bar{F}(u_n(x))$. Since (5.25) holds, the standard Poisson limit result implies that, as $n \to \infty$, the number of exceedances $N_{u_n(x)}$ converges to a Poisson rv with mean $\lambda(x) = -\ln H_\xi(x)$, depending on the particular value of x chosen.

The theory goes further. Not only is the number of exceedances asymptotically Poisson, these exceedances occur according to a Poisson point process. To state the result it is useful to give a brief summary of some ideas concerning point processes.

On point processes. Suppose we have a sequence of rvs or vectors $Y_1, \dots, Y_n$ taking values in some *state space* $\mathcal{X}$ (for example, $\mathbb{R}$ or $\mathbb{R}^2$) and we define, for any set $A \subset \mathcal{X}$, the rv

$$N(A) = \sum_{i=1}^{n} I_{\{Y_i \in A\}}, \tag{5.26}$$

which counts the random number of Y_i in the set A. Under some technical conditions (see EKM, pp. 220–223), (5.26) is said to define a point process $N(\cdot)$. An example of a point process is the Poisson point process.

Definition 5.29 (Poisson point process). The point process $N(\cdot)$ is called a *Poisson point process* (or Poisson random measure) on $\mathcal{X}$ with *intensity measure* Λ if the following two conditions are satisfied.

(a) For $A \subset \mathcal{X}$ and $k \geqslant 0$,

$$P(N(A) = k) = \begin{cases} \mathrm{e}^{-\Lambda(A)} \dfrac{\Lambda(A)^k}{k!}, & \Lambda(A) < \infty, \\ 0, & \Lambda(A) = \infty. \end{cases}$$

(b) For any $m \geqslant 1$, if $A_1, \dots, A_m$ are mutually disjoint subsets of $\mathcal{X}$, then the rvs $N(A_1), \dots, N(A_m)$ are independent.

The intensity measure $\Lambda(\cdot)$ of $N(\cdot)$ is also known as the *mean measure* because $E(N(A)) = \Lambda(A)$. We also speak of the intensity function (or simply intensity) of the process, which is the derivative $\lambda(\boldsymbol{x})$ of the measure satisfying $\Lambda(A) = \int_A \lambda(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$.

Asymptotic behaviour of the point process of exceedances. Consider again the strict white noise $(X_i)_{i\in\mathbb{N}}$ and the sequence of thresholds $u_n(x) = c_n x + d_n$ for some fixed x. For $n \in \mathbb{N}$ and $1 \leqslant i \leqslant n$ let $Y_{i,n} = (i/n)I_{\{X_i > u_n(x)\}}$ and observe that $Y_{i,n}$ can be thought of as returning either the normalized "time" i/n of an exceedance, or zero. The point process of exceedances of the threshold u_n is the process $N_n(\cdot)$ with state space $\mathcal{X} = (0, 1]$, which is given by

$$N_n(A) = \sum_{i=1}^{n} I_{\{Y_{i,n}\in A\}} \tag{5.27}$$

for $A \subset \mathcal{X}$. As the notation indicates, we consider this process to be an element in a sequence of point processes indexed by n. The point process (5.27) counts the exceedances with time of occurrence in the set A, and we are interested in the behaviour of this process as $n \to \infty$.

It may be shown (see Theorem 5.3.2 in EKM) that $N_n(\cdot)$ converges in distribution on $\mathcal{X}$ to a Poisson point process $N(\cdot)$ with intensity measure $\Lambda(\cdot)$ satisfying $\Lambda(A) = (t_2 - t_1)\lambda(x)$ for $A = (t_1, t_2) \subset \mathcal{X}$, where $\lambda(x) = -\ln H_\xi(x)$ as before. This implies, in particular, that $E(N_n(A)) \to E(N(A)) = \Lambda(A) = (t_2 - t_1)\lambda(x)$. Clearly, the intensity does not depend on time and takes the constant value $\lambda := \lambda(x)$; we refer to the limiting process as a *homogeneous Poisson process* with intensity or rate λ.

Application of the result in practice. We give a heuristic argument explaining how this limiting result is used in practice. We consider a fixed large sample size n and a fixed high threshold u, which we assume satisfies $u = c_n y + d_n$ for some value y. We expect that the number of threshold exceedances can be approximated by a Poisson rv and that the point process of exceedances of u can be approximated by a homogeneous Poisson process with rate $\lambda = -\ln H_\xi(y) = -\ln H_\xi((u - d_n)/c_n)$. If we replace the normalizing constants c_n and d_n by $\sigma > 0$ and μ, we have a Poisson process with rate $-\ln H_{\xi,\mu,\sigma}(u)$. Clearly, we could repeat the same argument with any high threshold so that, for example, we would expect it to be approximately true that exceedances of the level $x \geqslant u$ occur according to a Poisson process with rate $-\ln H_{\xi,\mu,\sigma}(x)$.

We therefore have an intimate relationship between the GEV model for block maxima and a Poisson model for the occurrence in time of exceedances of a high threshold. The arguments of this section therefore provide theoretical support for the observation in Figure 3.3: that exceedances for simulated iid t data are separated by waiting times that behave like iid exponential observations.

5.3.2 The POT Model

The theory of the previous section combined with the theory of Section 5.2 suggests an asymptotic model for threshold exceedances in regularly spaced iid data (or data from a process with extremal index $\theta = 1$). The so-called POT model makes the following assumptions.

- Exceedances occur according to a homogeneous Poisson process in time.
- Excess amounts above the threshold are iid and independent of exceedance times.
- The distribution of excess amounts is generalized Pareto.

There are various alternative ways of describing this model. It might also be called a *marked Poisson* point process, where the exceedance times constitute the points and the GPD-distributed excesses are the marks. It can also be described as a (non-homogeneous) *two-dimensional Poisson* point process, where points (t, x) in two-dimensional space record times and magnitudes of exceedances. The latter representation is particularly powerful, as we now discuss.

Two-dimensional Poisson formulation of POT model. Assume that we have regularly spaced random losses $X_1, \dots, X_n$ and that we set a high threshold u. We *assume* that, on the state space $\mathcal{X} = (0, 1] \times (u, \infty)$, the point process defined by $N(A) = \sum_{i=1}^{n} I_{\{(i/n, X_i) \in A\}}$ is a Poisson process with intensity at a point (t, x) given by

$$\lambda(t, x) = \frac{1}{\sigma}\left(1 + \xi\frac{x - \mu}{\sigma}\right)^{-1/\xi - 1} \tag{5.28}$$

provided $(1+\xi(x-\mu)/\sigma) > 0$, and by $\lambda(t, x) = 0$ otherwise. Note that this intensity does not depend on t but does depend on x, and hence the two-dimensional Poisson process is non-homogeneous; we simplify the notation to $\lambda(x) := \lambda(t, x)$. For a set of the form $A = (t_1, t_2) \times (x, \infty) \subset \mathcal{X}$, the intensity measure is

$$\Lambda(A) = \int_{t_1}^{t_2} \int_{x}^{\infty} \lambda(y)\,\mathrm{d}y\,\mathrm{d}t = -(t_2 - t_1) \ln H_{\xi,\mu,\sigma}(x). \tag{5.29}$$

It follows from (5.29) that for any $x \geqslant u$, the implied one-dimensional process of exceedances of the level x is a homogeneous Poisson process with rate $\tau(x) := -\ln H_{\xi,\mu,\sigma}(x)$. Now consider the excess amounts over the threshold u. The tail of the excess df over the threshold u, denoted by $\bar{F}_u(x)$ before, can be calculated as the ratio of the rates of exceeding the levels $u + x$ and u. We obtain

$$\bar{F}_u(x) = \frac{\tau(u + x)}{\tau(u)} = \left(1 + \frac{\xi x}{\sigma + \xi(u - \mu)}\right)^{-1/\xi} = \bar{G}_{\xi,\beta}(x)$$

for a positive scaling parameter $\beta = \sigma + \xi(u - \mu)$. This is precisely the tail of the GPD model for excesses over the threshold u used in Section 5.2.2. Thus this seemingly complicated model is indeed the POT model described informally at the beginning of this section.

Note also that the model implies the GEV distributional model for maxima. To see this, consider the event that $\{M_n \leqslant x\}$ for some value $x \geqslant u$. This may be expressed in point process language as the event that there are no points in the set $A = (0, 1] \times (x, \infty)$. The probability of this event is calculated to be $P(M_n \leqslant x) = P(N(A) = 0) = \mathrm{e}^{-\Lambda(A)} = H_{\xi,\mu,\sigma}(x)$, $x \geqslant u$, which is precisely the GEV model for maxima of n-blocks used in Section 5.1.4.

Statistical estimation of the POT model. The most elegant way of fitting the POT model to data is to fit the point process with intensity (5.28) to the exceedance data in one step. Given the exceedance data $\{\tilde{X}_j : j = 1, \dots, N_u\}$, the likelihood can be written as

$$L(\xi, \sigma, \mu; \tilde{X}_1, \dots, \tilde{X}_{N_u}) = \mathrm{e}^{-\tau(u)} \prod_{j=1}^{N_u} \lambda(\tilde{X}_j). \tag{5.30}$$

Parameter estimates of ξ, σ and μ are obtained by maximizing this expression, which is easily accomplished by numerical means. For literature on the derivation of this likelihood, see Notes and Comments.

There are, however, simpler ways of getting the same parameter estimates. Suppose we reparametrize the POT model in terms of $\tau := \tau(u) = -\ln H_{\xi,\mu,\sigma}(u)$, the rate of the one-dimensional Poisson process of exceedances of the level u, and $\beta = \sigma + \xi(u - \mu)$, the scaling parameter of the implied GPD for the excess losses over u. The intensity in (5.28) can then be rewritten as

$$\lambda(x) = \lambda(t, x) = \frac{\tau}{\beta}\left(1 + \xi\frac{x - u}{\beta}\right)^{-1/\xi-1}, \tag{5.31}$$

where $\xi \in \mathbb{R}$ and $\tau, \beta > 0$. Using this parametrization it is easily verified that the log of the likelihood in (5.30) becomes

$$\ln L(\xi, \sigma, \mu; \tilde{X}_1, \dots, \tilde{X}_{N_u}) = \ln L_1(\xi, \beta; \tilde{X}_1 - u, \dots, \tilde{X}_{N_u} - u) + \ln L_2(\tau; N_u),$$

where L_1 is precisely the likelihood for fitting a GPD to excess losses given in (5.14), and $\ln L_2(\tau; N_u) = -\tau + N_u \ln \tau$, which is the log-likelihood for a one-dimensional homogeneous Poisson process with rate τ. Such a partition of a log-likelihood into a sum of two terms involving two different sets of parameters means that we can make separate inferences about the two sets of parameters; we can estimate ξ and β in a GPD analysis and then estimate τ by its MLE N_u and use these to infer estimates of μ and σ.

Advantages of the POT model formulation. One might ask what the advantages of approaching the modelling of extremes through the two-dimensional Poisson point process model described by the intensity (5.28) could be? One advantage is the fact that the parameters ξ, μ and σ in the Poisson point process model do not have any theoretical dependence on the threshold chosen, unlike the parameter β in the GPD model, which appears in the theory as a function of the threshold u. In practice, we would expect the estimated parameters of the Poisson model to be roughly stable over a range of high thresholds, whereas the estimated β parameter varies with threshold choice.

For this reason the intensity (5.28) is a framework that is often used to introduce covariate effects into extreme value modelling. One method of doing this is to replace the parameters μ and σ in (5.28) by parameters that vary over time as a function of deterministic covariates. For example, we might have $\mu(t) = \alpha + \boldsymbol{\gamma}'\boldsymbol{y}(t)$, where $\boldsymbol{y}(t)$ represents a vector of covariate values at time t. This would give us Poisson processes that are also non-homogeneous in time.

Applicability of the POT model to return series data. We now turn to the use of the POT model with financial return data. An initial comment is that returns do not really form genuine point events in time, in contrast to recorded water levels or wind speeds, for example. Returns are discrete-time measurements that describe changes in value taking place over the course of, say, a day or a week. Nonetheless, we assume that if we take a longer-term perspective, such data can be approximated by point events in time.

In Section 3.1 and in Figure 3.3 in particular, we saw evidence that, in contrast to iid data, exceedances of a high threshold for daily financial return series do not necessarily occur according to a homogeneous Poisson process. They tend instead to form clusters corresponding to episodes of high volatility. Thus the standard POT model is not directly applicable to financial return data.

Theory suggests that for stochastic processes with extremal index $\theta < 1$, such as GARCH processes, the extremal clusters themselves should occur according to a homogeneous Poisson process in time, so that the individual exceedances occur according to a *Poisson cluster process* (see, for example, Leadbetter 1991). A suitable model for the occurrence and magnitude of exceedances in a financial return series might therefore be some form of marked Poisson cluster process.

Rather than attempting to specify the mechanics of cluster formation, it is quite common to try to circumvent the problem by *declustering* financial return data: we attempt to formally identify clusters of exceedances and then we apply the POT model to cluster maxima only. This method is obviously somewhat ad hoc, as there is usually no clear way of deciding where one cluster ends and another begins. A possible declustering algorithm is given by the *runs method*. In this method a run size r is fixed and two successive exceedances are said to belong to two different clusters if they are separated by a run of at least r values below the threshold (see EKM, pp. 422–424). In Figure 5.11 the DAX daily negative returns of Figure 3.3 have been declustered with a run length of ten trading days; this reduces the 100 exceedances to 42 cluster maxima.

However, it is not clear that applying the POT model to declustered data gives us a particularly useful model. We can estimate the rate of occurrence of clusters of extremes and say something about average cluster size; we can also derive a GPD model for excess losses over thresholds for cluster maxima (where standard errors for parameters may be more realistic than if we fitted the GPD to the dependent sample of all threshold exceedances). However, by neglecting the modelling of cluster formation, we cannot make more dynamic statements about the intensity of occurrence of threshold exceedances at any point in time. In Section 16.2 we describe self-exciting point process models, which do attempt to model the dynamics of cluster formation.

Example 5.30 (POT analysis of AT&T weekly losses). We close this section with an example of a standard POT model applied to extremes in financial return data. To mitigate the clustering phenomenon discussed above we use weekly return data, as previously analysed in Examples 5.24 and 5.25. Recall that these yield 102 weekly percentage losses for the AT&T stock price exceeding a threshold of 2.75%. The

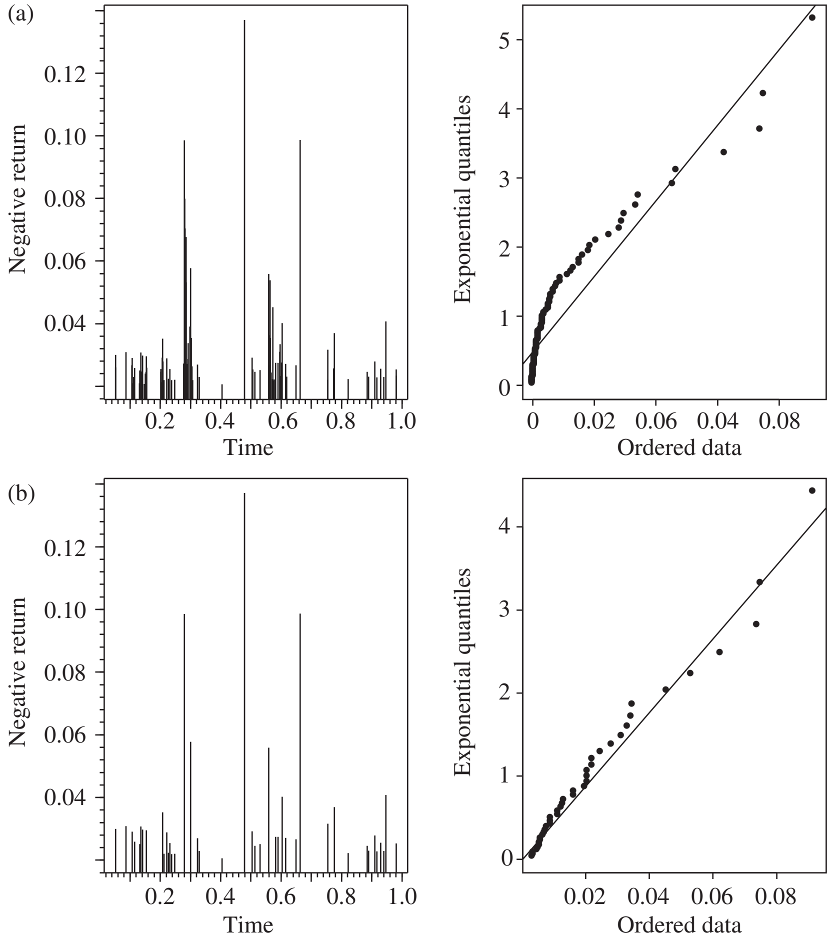

Figure 5.11. (a) DAX daily negative returns and a Q–Q plot of their spacings as in Figure 3.3. (b) Data have been declustered with the runs method using a run length of ten trading days. The spacings of the 42 cluster maxima are more consistent with a Poisson model.

data are shown in Figure 5.12, where we observe that the inter-exceedance times seem to have a roughly exponential distribution, although the discrete nature of the times and the relatively low value of n means that there are some tied values for the spacings, which makes the plot look a little granular. Another noticeable feature is that the exceedances of the threshold appear to become more frequent over time, which might be taken as evidence against the homogeneous Poisson assumption for threshold exceedances and against the implicit assumption that the underlying data form a realization from a stationary time series. It would be possible to consider a POT model incorporating a trend of increasingly frequent exceedances, but we will not go this far.

We fit the standard two-dimensional Poisson model to the 102 exceedances of the threshold 2.75% using the likelihood in (5.30) and obtain parameter estimates

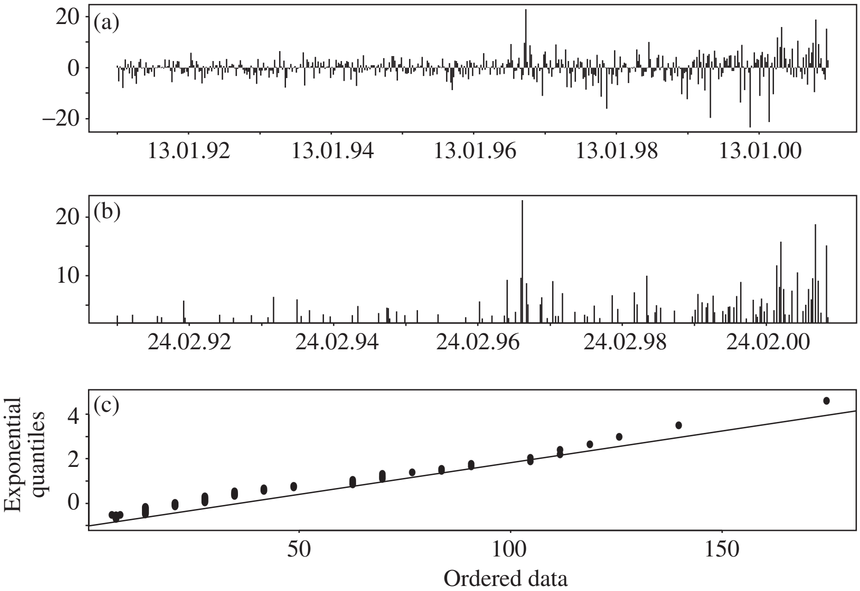

Figure 5.12. (a) Time series of AT&T weekly percentage losses from 1991 to 2000. (b) Corresponding realization of the marked point process of exceedances of the threshold 2.75%. (c) Q–Q plot of inter-exceedance times against an exponential reference distribution. See Example 5.30 for details.

$\hat{\xi} = 0.22$, $\hat{\mu} = 19.9$ and $\hat{\sigma} = 5.95$. The implied GPD scale parameter for the distribution of excess losses over the threshold u is $\hat{\beta} = \hat{\sigma} + \hat{\xi}(u - \hat{\mu}) = 2.1$, so we have exactly the same estimates of ξ and β as in Example 5.24.

The estimated exceedance rate for the threshold $u = 2.75$ is given by $\hat{\tau}(u) = -\ln H_{\hat{\xi},\hat{\mu},\hat{\sigma}}(u) = 102$, which is precisely the number of exceedances of that threshold, as theory suggests. It is of more interest to look at estimated exceedance rates for higher thresholds. For example, we get $\hat{\tau}(15) = 2.50$, which implies that losses exceeding 15% occur as a Poisson process with rate 2.5 losses per ten-year period, so that such a loss is, roughly speaking, a four-year event. Thus the Poisson model gives us an alternative method of defining the return period of a stress event and a more powerful way of calculating such a risk measure. Similarly, we can invert the problem to estimate return levels: suppose we define the ten-year return level as that level which is exceeded according to a Poisson process with rate one loss per ten years, then we can easily estimate the level in our model by calculating

$$H^{-1}_{\hat{\xi},\hat{\mu},\hat{\sigma}}(\mathrm{e}^{-1}) = 19.9,$$

so the ten-year event is a weekly loss of roughly 20%. Using the profile likelihood method in Section A.3.5 we could also give confidence intervals for such estimates.

Notes and Comments

For more information about point processes consult EKM, Cox and Isham (1980), Kallenberg (1983) and Resnick (2008). The point process approach to extremes dates back to Pickands (1971) and is also discussed in Leadbetter, Lindgren and Rootzén (1983), Leadbetter (1991) and Falk, Hüsler and Reiss (1994).

The two-dimensional Poisson point process model was first used in practice by Smith (1989) and may also be found in Smith and Shively (1995); both these papers discuss the adaptation of the point process model to incorporate covariates or time trends in the context of environmental data. An insurance application is treated in Smith and Goodman (2000), which also treats the point process model from a Bayesian perspective. An interesting application to wind storm losses is Rootzén and Tajvidi (1997). A further application of the bivariate point process framework to the modeling of insurance loss data, showing trends in both intensity and severity of occurrence, is found in McNeil and Saladin (2000). For further applications to insurance and finance, see Chavez-Demoulin and Embrechts (2004) and Chavez-Demoulin, Embrechts and Hofert (2014). An excellent overview of statistical approaches to the GPD and point process models is found in Coles (2001).

The derivation of likelihoods for point process models is beyond the scope of this book and we have simply recorded the likelihoods to be maximized without further justification. See Daley and Vere-Jones (2003, Chapter 7) for more details on this subject; see also Coles (2001, p. 127) for a good intuitive account in the Poisson case.

6
Multivariate Models

Financial risk models, whether for market or credit risks, are inherently multivariate. The value change of a portfolio of traded instruments over a fixed time horizon depends on a random vector of risk-factor changes or returns. The loss incurred by a credit portfolio depends on a random vector of losses for the individual counterparties in the portfolio.

This chapter is the first of two successive ones that focus on models for random vectors. The emphasis in this chapter is on tractable models that describe both the individual behaviour of the components of a random vector and their joint behaviour or dependence structure. We consider a number of distributions that extend the multivariate normal but provide more realistic models for many kinds of financial data.

In Chapter 7 we focus explicitly on modelling the dependence structure of a random vector and largely ignore marginal behaviour. We introduce copula models of dependence and study a number of dependence measures and concepts related to copulas. Both Chapter 6 and Chapter 7 take a static, distributional view of multivariate modelling; for multivariate time-series models, see Chapter 14.

Section 6.1 reviews basic ideas in multivariate statistics and discusses the multivariate normal (or Gaussian) distribution and its deficiencies as a model for empirical return data.

In Section 6.2 we consider a generalization of the multivariate normal distribution known as a multivariate normal mixture distribution, which shares much of the structure of the multivariate normal and retains many of its properties. We treat both variance mixtures, which belong to the wider class of elliptical distributions, and mean–variance mixtures, which allow asymmetry. Concrete examples include t distributions and generalized hyperbolic distributions, and we show in empirical examples that these models provide a better fit than a Gaussian distribution to asset return data. In some cases, multivariate return data are not strongly asymmetric and models from the class of elliptical distributions are good enough; in Section 6.3 we investigate the elegant properties of these distributions.

In Section 6.4 we discuss the important issue of dimension-reduction techniques for reducing large sets of risk factors to smaller subsets of essential risk drivers. The key idea here is that of a factor model, and we also review the principal components method of constructing factors.

6.1 Basics of Multivariate Modelling

This first section reviews important basic material from multivariate statistics, which will be known to many readers. The main topic of the section is the multivariate normal distribution and its properties; this distribution is central to much of classical multivariate analysis and was the starting point for attempts to model market risk (the variance–covariance method of Section 9.2.2).

6.1.1 Random Vectors and Their Distributions

Joint and marginal distributions. Consider a general d-dimensional random vector of risk-factor changes (or so-called returns) $\boldsymbol{X} = (X_1, \dots, X_d)'$. The distribution of $\boldsymbol{X}$ is completely described by the joint distribution function (df)

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{\boldsymbol{X}}(x_1, \dots, x_d) = P(\boldsymbol{X} \leqslant \boldsymbol{x}) = P(X_1 \leqslant x_1, \dots, X_d \leqslant x_d).$$

Where no ambiguity arises we simply write F, omitting the subscript.

The *marginal* df of X_i, written F_{X_i} or often simply F_i, is the df of that risk factor considered individually and is easily calculated from the joint df. For all i we have

$$F_i(x_i) = P(X_i \leqslant x_i) = F(\infty, \dots, \infty, x_i, \infty, \dots, \infty). \tag{6.1}$$

If the marginal df $F_i(x)$ is absolutely continuous, then we refer to its derivative $f_i(x)$ as the marginal density of X_i. It is also possible to define k-dimensional marginal distributions of $\boldsymbol{X}$ for $2 \leqslant k \leqslant d-1$. Suppose we partition $\boldsymbol{X}$ into $(\boldsymbol{X}_1', \boldsymbol{X}_2')'$, where $\boldsymbol{X}_1 = (X_1, \dots, X_k)'$ and $\boldsymbol{X}_2 = (X_{k+1}, \dots, X_d)'$, then the marginal df of $\boldsymbol{X}_1$ is

$$F_{\boldsymbol{X}_1}(\boldsymbol{x}_1) = P(\boldsymbol{X}_1 \leqslant \boldsymbol{x}_1) = F(x_1, \dots, x_k, \infty, \dots, \infty).$$

For bivariate and other low-dimensional margins it is convenient to have a simpler alternative notation in which, for example, $F_{ij}(x_i, x_j)$ stands for the marginal distribution of the components X_i and X_j.

The df of a random vector $\boldsymbol{X}$ is said to be absolutely continuous if

$$F(x_1, \dots, x_d) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_d} f(u_1, \dots, u_d)\, \mathrm{d}u_1 \cdots \mathrm{d}u_d$$

for some non-negative function f, which is then known as the *joint density* of $\boldsymbol{X}$. Note that the existence of a joint density implies the existence of marginal densities for all k-dimensional marginals. However, the existence of a joint density is not necessarily implied by the existence of marginal densities (counterexamples can be found in Chapter 7 on copulas).

In some situations it is convenient to work with the *survival function* of $\boldsymbol{X}$, defined by

$$\bar{F}_{\boldsymbol{X}}(\boldsymbol{x}) = \bar{F}_{\boldsymbol{X}}(x_1, \dots, x_d) = P(\boldsymbol{X} > \boldsymbol{x}) = P(X_1 > x_1, \dots, X_d > x_d)$$

and written simply as $\bar{F}$ when no ambiguity arises. The marginal survival function of X_i, written $\bar{F}_{X_i}$ or often simply $\bar{F}_i$, is given by

$$\bar{F}_i(x_i) = P(X_i > x_i) = \bar{F}(-\infty, \dots, -\infty, x_i, -\infty, \dots, -\infty).$$

Conditional distributions and independence. If we have a multivariate model for risks in the form of a joint df, survival function or density, then we have implicitly described the *dependence structure* of the risks. We can make conditional probability statements about the probability that certain components take certain values given that other components take other values. For example, consider again our partition of $\boldsymbol{X}$ into $(\boldsymbol{X}_1', \boldsymbol{X}_2')'$ and assume absolute continuity of the df of $\boldsymbol{X}$. Let $f_{\boldsymbol{X}_1}$ denote the joint density of the k-dimensional marginal distribution $F_{\boldsymbol{X}_1}$. Then the conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$ has density

$$f_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) = \frac{f(\boldsymbol{x}_1, \boldsymbol{x}_2)}{f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)}, \tag{6.2}$$

and the corresponding df is

$$\begin{aligned} &F_{\boldsymbol{X}_2|\boldsymbol{X}_1}(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) \\ &\quad = \int_{u_{k+1}=-\infty}^{x_{k+1}} \cdots \int_{u_d=-\infty}^{x_d} \frac{f(x_1, \dots, x_k, u_{k+1}, \dots, u_d)}{f_{\boldsymbol{X}_1}(\boldsymbol{x}_1)} \, \mathrm{d}u_{k+1} \cdots \mathrm{d}u_d. \end{aligned}$$

If the joint density of $\boldsymbol{X}$ factorizes into $f(\boldsymbol{x}) = f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)$, then the conditional distribution and density of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$ are identical to the marginal distribution and density of $\boldsymbol{X}_2$: in other words, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent. We recall that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are independent if and only if

$$F(\boldsymbol{x}) = F_{\boldsymbol{X}_1}(\boldsymbol{x}_1) F_{\boldsymbol{X}_2}(\boldsymbol{x}_2), \quad \forall \boldsymbol{x},$$

or, in the case where $\boldsymbol{X}$ possesses a joint density, $f(\boldsymbol{x}) = f_{\boldsymbol{X}_1}(\boldsymbol{x}_1) f_{\boldsymbol{X}_2}(\boldsymbol{x}_2)$.

The components of $\boldsymbol{X}$ are *mutually* independent if and only if $F(\boldsymbol{x}) = \prod_{i=1}^d F_i(x_i)$ for all $\boldsymbol{x} \in \mathbb{R}^d$ or, in the case where $\boldsymbol{X}$ possesses a density, $f(\boldsymbol{x}) = \prod_{i=1}^d f_i(x_i)$.

Moments and characteristic function. The *mean vector* of $\boldsymbol{X}$, when it exists, is given by

$$E(\boldsymbol{X}) := (E(X_1), \dots, E(X_d))'.$$

The *covariance matrix*, when it exists, is the matrix $\operatorname{cov}(\boldsymbol{X})$ defined by

$$\operatorname{cov}(\boldsymbol{X}) := E((\boldsymbol{X} - E(\boldsymbol{X}))(\boldsymbol{X} - E(\boldsymbol{X}))'),$$

where the expectation operator acts componentwise on matrices. If we write Σ for $\operatorname{cov}(\boldsymbol{X})$, then the (i, j)th element of this matrix is

$$\sigma_{ij} = \operatorname{cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j),$$

the ordinary pairwise covariance between X_i and X_j. The diagonal elements $\sigma_{11}, \dots, \sigma_{dd}$ are the variances of the components of $\boldsymbol{X}$.

The *correlation matrix* of $\boldsymbol{X}$, denoted by $\rho(\boldsymbol{X})$, can be defined by introducing a standardized vector $\boldsymbol{Y}$ such that $Y_i = X_i/\sqrt{\operatorname{var}(X_i)}$ for all i and taking $\rho(\boldsymbol{X}) := \operatorname{cov}(\boldsymbol{Y})$. If we write P for $\rho(\boldsymbol{X})$, then the (i, j)th element of this matrix is

$$\rho_{ij} = \rho(X_i, X_j) = \frac{\operatorname{cov}(X_i, X_j)}{\sqrt{\operatorname{var}(X_i) \operatorname{var}(X_j)}}, \tag{6.3}$$

the ordinary pairwise linear correlation of X_i and X_j. To express the relationship between correlation and covariance matrices in matrix form, it is useful to introduce operators on a covariance matrix Σ as follows:

$$\Delta(\Sigma) := \operatorname{diag}(\sqrt{\sigma_{11}}, \dots, \sqrt{\sigma_{dd}}), \tag{6.4}$$

$$\wp(\Sigma) := (\Delta(\Sigma))^{-1}\Sigma(\Delta(\Sigma))^{-1}. \tag{6.5}$$

Thus $\Delta(\Sigma)$ extracts from Σ a diagonal matrix of standard deviations, and $\wp(\Sigma)$ extracts a correlation matrix. The covariance and correlation matrices Σ and P of $\boldsymbol{X}$ are related by

$$P = \wp(\Sigma). \tag{6.6}$$

Mean vectors and covariance matrices are manipulated extremely easily under linear operations on the vector $\boldsymbol{X}$. For any matrix $B \in \mathbb{R}^{k \times d}$ and vector $\boldsymbol{b} \in \mathbb{R}^k$ we have

$$E(B\boldsymbol{X} + \boldsymbol{b}) = BE(\boldsymbol{X}) + \boldsymbol{b}, \tag{6.7}$$

$$\operatorname{cov}(B\boldsymbol{X} + \boldsymbol{b}) = B\operatorname{cov}(\boldsymbol{X})B'. \tag{6.8}$$

Covariance matrices (and hence correlation matrices) are therefore *positive semidefinite*; writing Σ for $\operatorname{cov}(\boldsymbol{X})$ we see that (6.8) implies that $\operatorname{var}(\boldsymbol{a}'\boldsymbol{X}) = \boldsymbol{a}'\Sigma\boldsymbol{a} \geqslant 0$ for any $\boldsymbol{a} \in \mathbb{R}^d$. If we have that $\boldsymbol{a}'\Sigma\boldsymbol{a} > 0$ for any $\boldsymbol{a} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, we say that Σ is *positive definite*; in this case the matrix is invertible. We will make use of the well-known *Cholesky factorization* of positive-definite covariance matrices at many points; it is well known that such a matrix can be written as $\Sigma = AA'$ for a lower triangular matrix A with positive diagonal elements. The matrix A is known as the Cholesky factor. It will be convenient to denote this factor by $\Sigma^{1/2}$ and its inverse by $\Sigma^{-1/2}$. Note that there are other ways of defining the "square root" of a symmetric positive-definite matrix (such as the symmetric decomposition), but we will always use $\Sigma^{1/2}$ to denote the Cholesky factor.

In this chapter many properties of the multivariate distribution of a vector $\boldsymbol{X}$ are demonstrated using the characteristic function, which is given by

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}), \quad \boldsymbol{t} \in \mathbb{R}^d.$$

6.1.2 Standard Estimators of Covariance and Correlation

Suppose we have n observations of a d-dimensional risk-factor return vector denoted by $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. Typically, these would be daily, weekly, monthly or yearly observations forming a multivariate time series. We will assume throughout this chapter that the observations are *identically distributed* in the window of observation and either independent or at least serially *uncorrelated* (also known as multivariate white noise). As discussed in Chapter 3, the assumption of independence may be roughly tenable for longer time intervals such as months or years. For shorter time intervals independence may be a less appropriate assumption (due to a phenomenon known as volatility clustering, discussed in Section 3.1.1), but serial correlation of returns is often quite weak.

We assume that the observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ come from a distribution with mean vector $\boldsymbol{\mu}$, finite covariance matrix Σ and correlation matrix P. We now briefly review the standard estimators of these vector and matrix parameters.

Standard method-of-moments estimators of $\boldsymbol{\mu}$ and Σ are given by the *sample mean vector* $\bar{\boldsymbol{X}}$ and the *sample covariance matrix* S. These are defined by

$$\bar{\boldsymbol{X}} := \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i, \qquad S := \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})', \tag{6.9}$$

where arithmetic operations on vectors and matrices are performed componentwise. $\bar{\boldsymbol{X}}$ is an unbiased estimator but S is biased; an unbiased version may be obtained by taking $S_u := nS/(n-1)$, as may be seen by calculating

$$\begin{aligned} nE(S) &= E\biggl(\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu})(\boldsymbol{X}_i - \boldsymbol{\mu})' - n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})(\bar{\boldsymbol{X}} - \boldsymbol{\mu})'\biggr) \\ &= \sum_{i=1}^{n} \operatorname{cov}(\boldsymbol{X}_i) - n \operatorname{cov}(\bar{\boldsymbol{X}}) = n\Sigma - \Sigma, \end{aligned}$$

since $\operatorname{cov}(\bar{\boldsymbol{X}}) = n^{-1}\Sigma$ when the data vectors are iid, or identically distributed and uncorrelated.

The *sample correlation matrix* R may be easily calculated from the sample covariance matrix; its (j,k)th element is given by $r_{jk} = s_{jk}/\sqrt{s_{jj}s_{kk}}$, where s_{jk} denotes the (j,k)th element of S. Or, using the notation introduced in (6.5), we have

$$R = \wp(S),$$

which is the analogous equation to (6.6) for estimators.

Further properties of the estimators $\bar{\boldsymbol{X}}$, S and R will very much depend on the *true multivariate distribution* of the observations. These quantities are not necessarily the best estimators of the corresponding theoretical quantities in all situations. This point is often forgotten in financial risk management, where sample covariance and correlation matrices are routinely calculated and interpreted with little critical consideration of underlying models.

If our data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are iid multivariate normal, then $\bar{\boldsymbol{X}}$ and S are the *maximum likelihood estimators* (MLEs) of the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Their behaviour as estimators is well understood, and statistical inference for the model parameters is described in all standard texts on multivariate analysis.

However, the multivariate normal is certainly not a good description of financial risk-factor returns over short time intervals, such as daily data, and is often not good over longer time intervals either. Under these circumstances the behaviour of the standard estimators in (6.9) is often less well understood, and other estimators of the true mean vector $\boldsymbol{\mu}$ and covariance matrix Σ may perform better in terms of *efficiency* and *robustness*. Roughly speaking, by a more efficient estimator we mean an estimator with a smaller expected estimation error; by a more robust estimator we mean an estimator whose performance is not so susceptible to the presence of outlying data values.

6.1.3 The Multivariate Normal Distribution

Definition 6.1. $\boldsymbol{X} = (X_1, \dots, X_d)'$ has a multivariate normal or *Gaussian* distribution if

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + A\boldsymbol{Z},$$

where $\boldsymbol{Z} = (Z_1, \dots, Z_k)'$ is a vector of iid univariate *standard* normal rvs (mean 0 and variance 1), and $A \in \mathbb{R}^{d \times k}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ are a matrix and a vector of constants, respectively.

It is easy to verify, using (6.7) and (6.8), that the mean vector of this distribution is $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and the covariance matrix is $\operatorname{cov}(\boldsymbol{X}) = \Sigma$, where $\Sigma = AA'$ is a positive-semidefinite matrix. Moreover, using the fact that the characteristic function of a standard univariate normal variate Z is $\phi_Z(t) = \mathrm{e}^{-t^2/2}$, the characteristic function of $\boldsymbol{X}$ may be calculated to be

$$\phi_X(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}), \quad \boldsymbol{t} \in \mathbb{R}^d. \tag{6.10}$$

Clearly, the distribution is characterized by its mean vector and covariance matrix, and hence a standard notation is $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$. Note that the components of $\boldsymbol{X}$ are mutually independent if and only if Σ is diagonal. For example, $\boldsymbol{X} \sim N_d(\boldsymbol{0}, I_d)$ if and only if $X_1, \dots, X_d$ are iid $N(0, 1)$, the standard univariate normal distribution.

We concentrate on the *non-singular case* of the multivariate normal when $\operatorname{rank}(A) = d \leqslant k$. In this case the covariance matrix Σ has full rank d and is therefore invertible (non-singular) and positive definite. Moreover, $\boldsymbol{X}$ has an absolutely continuous distribution function with joint density given by

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\{-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}, \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{6.11}$$

where $|\Sigma|$ denotes the determinant of Σ.

The form of the density clearly shows that points with equal density lie on *ellipsoids* determined by equations of the form $(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c$, for constants $c > 0$. In two dimensions the contours of equal density are ellipses, as illustrated in Figure 6.1. Whenever a multivariate density $f(\boldsymbol{x})$ depends on $\boldsymbol{x}$ only through the quadratic form $(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})$, it is the density of a so-called elliptical distribution, as discussed in more detail in Section 6.3.

Definition 6.1 is essentially a simulation recipe for the multivariate normal distribution. To be explicit, if we wished to generate a vector $\boldsymbol{X}$ with distribution $N_d(\boldsymbol{\mu}, \Sigma)$, where Σ is positive definite, we would use the following algorithm.

Algorithm 6.2 (simulation of multivariate normal distribution).

(1) Perform a Cholesky decomposition of Σ (see, for example, Press et al. 1992) to obtain the Cholesky factor $\Sigma^{1/2}$.

(2) Generate a vector $\boldsymbol{Z} = (Z_1, \dots, Z_d)'$ of independent standard normal variates.

(3) Set $\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{Z}$.

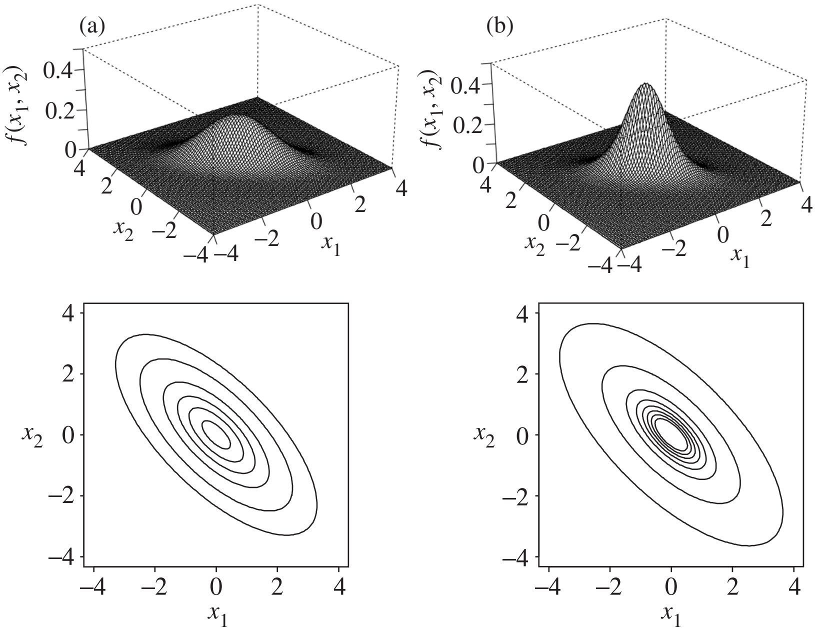

Figure 6.1. (a) Perspective and contour plots for the density of a bivariate normal distribution with standard normal margins and correlation -70%. (b) Corresponding plots for a bivariate t density with four degrees of freedom (see Example 6.7 for details) and the *same mean vector and covariance matrix* as the normal distribution. Contour lines are plotted at the same heights for both densities.

We now summarize further useful properties of the multivariate normal. These properties underline the attractiveness of the multivariate normal for computational work in risk management. Note, however, that many of them are in fact shared by the broader classes of normal mixture distributions and elliptical distributions (see Section 6.3.3 for properties of the latter).

Linear combinations. If we take linear combinations of multivariate normal random vectors, then these remain multivariate normal. Let $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ and take any $B \in \mathbb{R}^{k \times d}$ and $\boldsymbol{b} \in \mathbb{R}^k$. Then it is easily shown (e.g. using the characteristic function (6.10)) that

$$B\boldsymbol{X} + \boldsymbol{b} \sim N_k(B\boldsymbol{\mu} + \boldsymbol{b}, B\Sigma B'). \tag{6.12}$$

As a special case, if $\boldsymbol{a} \in \mathbb{R}^d$, then

$$\boldsymbol{a}'\boldsymbol{X} \sim N(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\Sigma\boldsymbol{a}), \tag{6.13}$$

and this fact is used routinely in the variance–covariance approach to risk management, as discussed in Section 9.2.2.

In this context it is interesting to note the following elegant characterization of multivariate normality. It is easily shown using characteristic functions that $\boldsymbol{X}$ is multivariate normal *if and only if* $\boldsymbol{a}'\boldsymbol{X}$ is univariate normal for all vectors $\boldsymbol{a} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$.

Marginal distributions. It is clear from (6.13) that the univariate marginal distributions of $\boldsymbol{X}$ must be univariate normal. More generally, using the $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ notation from Section 6.1.1 and extending this notation naturally to $\boldsymbol{\mu}$ and Σ,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

property (6.12) implies that the marginal distributions of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are also multivariate normal and are given by $\boldsymbol{X}_1 \sim N_k(\boldsymbol{\mu}_1, \Sigma_{11})$ and $\boldsymbol{X}_2 \sim N_{d-k}(\boldsymbol{\mu}_2, \Sigma_{22})$.

Conditional distributions. Assuming that Σ is positive definite, the conditional distributions of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1$ and of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2$ may also be shown to be multivariate normal. For example, $\boldsymbol{X}_2 \mid \boldsymbol{X}_1 = \boldsymbol{x}_1 \sim N_{d-k}(\boldsymbol{\mu}_{2.1}, \Sigma_{22.1})$, where

$$\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) \quad \text{and} \quad \Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$$

are the conditional mean vector and covariance matrix.

Quadratic forms. If $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ with Σ positive definite, then

$$(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi_d^2, \tag{6.14}$$

a chi-squared distribution with d degrees of freedom. This is seen by observing that $\boldsymbol{Z} = \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim N_d(\boldsymbol{0}, I_d)$ and $(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) = \boldsymbol{Z}'\boldsymbol{Z} \sim \chi_d^2$. This property (6.14) is useful for checking multivariate normality (see Section 6.1.4).

Convolutions. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent d-dimensional random vectors satisfying $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{Y} \sim N_d(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma})$, then we may take the product of characteristic functions to show that $\boldsymbol{X} + \boldsymbol{Y} \sim N_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma + \tilde{\Sigma})$.

6.1.4 Testing Multivariate Normality

We now consider the issue of testing whether the data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are observations from a multivariate normal distribution.

Univariate tests. If $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are iid multivariate normal, then for $1 \leqslant j \leqslant d$ the univariate sample $X_{1,j}, \dots, X_{n,j}$ consisting of the observations of the jth component must be iid univariate normal; in fact, any univariate sample constructed from a linear combination of the data of the form $\boldsymbol{a}'\boldsymbol{X}_1, \dots, \boldsymbol{a}'\boldsymbol{X}_n$ must be iid univariate normal. This can be assessed graphically with a Q–Q plot against a standard normal reference distribution, or it can be tested formally using one of the many numerical tests of normality (see Section 3.1.2 for more details of univariate tests of normality).

Multivariate tests. To test for multivariate normality it is not sufficient to test that the univariate margins of the distribution are normal. We will see in Chapter 7 that it is possible to have multivariate distributions with normal margins that are not themselves multivariate normal distributions. Thus we also need to be able to test *joint normality*, and a simple way of doing this is to exploit the fact that the quadratic form in (6.14) has a chi-squared distribution. Suppose we estimate $\boldsymbol{\mu}$ and Σ using the standard estimators in (6.9) and construct the data

$$\{D_i^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})' S^{-1} (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) \colon i = 1, \dots, n\}. \tag{6.15}$$

Because the estimates of the mean vector and the covariance matrix are used in the construction of each D_i^2, these data are not independent, even if the original $\boldsymbol{X}_i$ data were. Moreover, the marginal distribution of D_i^2 under the null hypothesis is not exactly chi-squared; in fact, we have that $n(n-1)^{-2} D_i^2 \sim \text{Beta}(\frac{1}{2}d, \frac{1}{2}(n-d-1))$, so that the true distribution is a scaled beta distribution, although it turns out to be very close to chi-squared for large n. We expect $D_1^2, \dots, D_n^2$ to behave roughly like an iid sample from a χ_d^2 distribution, and for simplicity we construct Q–Q plots against this distribution. (It is also possible to make Q–Q plots against the beta reference distribution, and these look very similar.)

Numerical tests of multivariate normality based on multivariate measures of skewness and kurtosis are also possible. Suppose we define, in analogy to (3.1),

$$b_d = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}^3, \qquad k_d = \frac{1}{n} \sum_{i=1}^{n} D_i^4, \tag{6.16}$$

where D_i is given in (6.15) and is known as the *Mahalanobis distance* between $\boldsymbol{X}_i$ and $\bar{\boldsymbol{X}}$, and $D_{ij} = (\boldsymbol{X}_i - \bar{\boldsymbol{X}}) S^{-1} (\boldsymbol{X}_j - \bar{\boldsymbol{X}})$ is known as the *Mahalanobis angle* between $\boldsymbol{X}_i - \bar{\boldsymbol{X}}$ and $\boldsymbol{X}_j - \bar{\boldsymbol{X}}$. Under the null hypothesis of multivariate normality the asymptotic distributions of these statistics as $n \to \infty$ are

$$\tfrac{1}{6} n b_d \sim \chi^2_{d(d+1)(d+2)/6}, \qquad \frac{k_d - d(d+2)}{\sqrt{8d(d+2)/n}} \sim N(0,1). \tag{6.17}$$

Mardia's test of multinormality involves comparing the skewness and kurtosis statistics with the above theoretical reference distributions. Since large values of the statistics cast doubt on the multivariate normal model, one-sided tests are generally performed. Usually, the tests of kurtosis and skewness are performed separately, although there are also a number of joint (or so-called omnibus) tests (see Notes and Comments).

Example 6.3 (on the normality of returns on Dow Jones 30 stocks). In Section 3.1.2 we applied univariate tests of normality to an arbitrary subgroup of ten stocks from the Dow Jones index. We took eight years of data spanning the period 1993–2000 and formed daily, weekly, monthly and quarterly logarithmic returns. In this example we apply Mardia's tests of multinormality based on both multivariate skewness and kurtosis to the multivariate data for all ten stocks. The results are shown in Table 6.1. We also compare the D_i^2 data (6.15) to a χ_{10}^2 distribution using a Q–Q plot (see Figure 6.2).

The daily, weekly and monthly return data fail the multivariate tests of normality. For quarterly return data the multivariate kurtosis test does not reject the null hypothesis but the skewness test does; the Q–Q plot in Figure 6.2 (d) looks slightly

Table 6.1. Mardia's tests of multivariate normality based on the multivariate measures of skewness and kurtosis in (6.16) and the asymptotic distributions in (6.17) (see Example 6.3 for details).

	Daily	Weekly	Monthly	Quarterly
n	2020	416	96	32
b_{10}	9.31	9.91	21.10	50.10
p-value	0.00	0.00	0.00	0.02
k_{10}	242.45	177.04	142.65	120.83
p-value	0.00	0.00	0.00	0.44

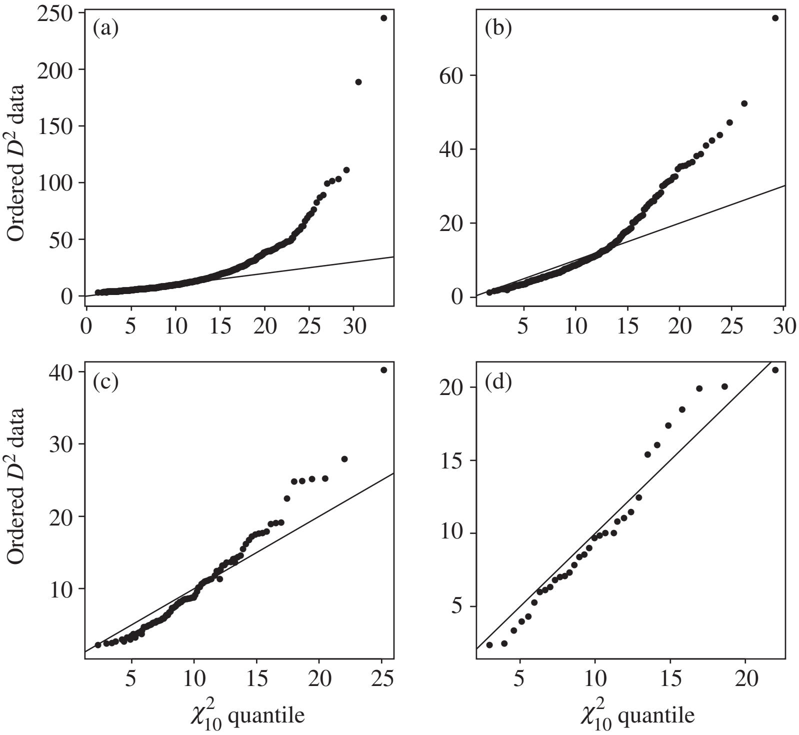

Figure 6.2. Q–Q plot of the D_i^2 data in (6.15) against a χ^2_{10} distribution for the data sets of Example 6.3: (a) daily analysis, (b) weekly analysis, (c) monthly analysis and (d) quarterly analysis. Under the null hypothesis of *multivariate* normality these should be roughly linear.

more linear. There is therefore some evidence that returns over a quarter year are close to being normally distributed, which might indicate a central limit theorem effect taking place, although the sample size is too small to reach any more reliable conclusion. The evidence against the multivariate normal distribution is certainly overwhelming for daily, weekly and monthly data.

The results in Example 6.3 are fairly typical for financial return data. This suggests that in many risk-management applications the multivariate normal distribution is not a good description of reality. It has three main defects, all of which are discussed at various points in this book.

(1) The tails of its univariate marginal distributions are too thin; they do not assign enough weight to *extreme* events (see also Section 3.1.2).

(2) The joint tails of the distribution do not assign enough weight to *joint extreme* outcomes (see also Section 7.3.1).

(3) The distribution has a strong form of symmetry, known as elliptical symmetry.

In the next section we look at models that address some of these defects. We consider normal variance mixture models, which share the elliptical symmetry of the multivariate normal but have the flexibility to address (1) and (2) above; we also look at normal mean–variance mixture models, which introduce some asymmetry and thus address (3).

Notes and Comments

Much of the material covered briefly in Section 6.1 can be found in greater detail in standard texts on multivariate statistical analysis such as Mardia, Kent and Bibby (1979), Seber (1984), Giri (1996) and Johnson and Wichern (2002).

The true distribution of $D_i^2 = (\boldsymbol{X}_i - \bar{\boldsymbol{X}})S^{-1}(\boldsymbol{X}_i - \bar{\boldsymbol{X}})$ for iid Gaussian data was shown by Gnanadesikan and Kettenring (1972) to be a scaled beta distribution (see also Gnanadesikan 1997). The implications of this fact for the construction of Q–Q plots in small samples are considered by Small (1978). References for multivariate measures of skewness and kurtosis and Mardia's test of multinormality are Mardia (1970, 1974, 1975). See also Mardia, Kent and Bibby (1979), the entry on "multivariate normality, testing for" in Volume 6 of the *Encyclopedia of Statistical Sciences* (Kotz, Johnson and Read 1985), and the entry on "Mardia's test of multinormality" in Volume 5 of the same publication. A paper that compares the performance of different goodness-of-fit tests for the multivariate normal distribution implemented in R is Joenssen and Vogel (2014).

6.2 Normal Mixture Distributions

In this section we generalize the multivariate normal to obtain multivariate normal mixture distributions. The crucial idea is the introduction of randomness into first the covariance matrix and then the mean vector of a multivariate normal distribution via a positive mixing variable, which will be known throughout as W.

6.2.1 Normal Variance Mixtures

Definition 6.4. The random vector $\boldsymbol{X}$ is said to have a (multivariate) normal variance mixture distribution if

$$\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + \sqrt{W} A \boldsymbol{Z}, \tag{6.18}$$

where

(i) $\boldsymbol{Z} \sim N_k(\boldsymbol{0}, I_k)$;

(ii) $W \geqslant 0$ is a non-negative, scalar-valued rv that is independent of $\boldsymbol{Z}$, and

(iii) $A \in \mathbb{R}^{d \times k}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ are a matrix and a vector of constants, respectively.

Such distributions are known as variance mixtures, since if we condition on the rv W, we observe that $\boldsymbol{X} \mid W = w \sim N_d(\boldsymbol{\mu}, w\Sigma)$, where $\Sigma = AA'$. The distribution of $\boldsymbol{X}$ can be thought of as a composite distribution constructed by taking a set of multivariate normal distributions with the same mean vector and with the same covariance matrix up to a multiplicative constant w. The mixture distribution is constructed by drawing randomly from this set of component multivariate normals according to a set of "weights" determined by the distribution of W; the resulting mixture is not itself a multivariate normal distribution. In the context of modelling risk-factor returns, the mixing variable W could be interpreted as a *shock* that arises from new information and impacts the volatilities of all risk factors.

As for the multivariate normal, we are most interested in the case where $\operatorname{rank}(A) = d \leqslant k$ and Σ is a full-rank, positive-definite matrix; this will give us a non-singular normal variance mixture.

Provided that W has a finite expectation, we may easily calculate that

$$E(\boldsymbol{X}) = E(\boldsymbol{\mu} + \sqrt{W}A\boldsymbol{Z}) = \boldsymbol{\mu} + E(\sqrt{W})AE(\boldsymbol{Z}) = \boldsymbol{\mu}$$

and that

$$\operatorname{cov}(\boldsymbol{X}) = E((\sqrt{W}A\boldsymbol{Z})(\sqrt{W}A\boldsymbol{Z})') = E(W)AE(\boldsymbol{Z}\boldsymbol{Z}')A' = E(W)\Sigma. \tag{6.19}$$

We generally refer to $\boldsymbol{\mu}$ and Σ as the *location vector* and the *dispersion matrix* of the distribution. Note that Σ (the covariance matrix of $A\boldsymbol{Z}$) is only the covariance matrix of $\boldsymbol{X}$ if $E(W) = 1$, and that $\boldsymbol{\mu}$ is only the mean vector when $E(\boldsymbol{X})$ is defined, which requires $E(W^{1/2}) < \infty$. The correlation matrices of $\boldsymbol{X}$ and $A\boldsymbol{Z}$ are the same when $E(W) < \infty$. Note also that these distributions provide good examples of models where a lack of correlation does not necessarily imply independence of the components of $\boldsymbol{X}$; indeed, we have the following simple result.

Lemma 6.5. *Let (X_1, X_2) have a normal mixture distribution with $A = I_2$ and $E(W) < \infty$ so that $\operatorname{cov}(X_1, X_2) = 0$. Then X_1 and X_2 are independent if and only if W is almost surely constant, i.e. (X_1, X_2) are normally distributed.*

Proof. If W is almost surely a constant, then (X_1, X_2) have a bivariate normal distribution and are independent. Conversely, if (X_1, X_2) are independent, then we must have $E(|X_1||X_2|) = E(|X_1|)E(|X_2|)$. We calculate that

$$\begin{aligned} E(|X_1||X_2|) &= E(W|Z_1||Z_2|) = E(W)E(|Z_1|)E(|Z_2|) \\ &\geqslant (E(\sqrt{W}))^2 E(|Z_1|)E(|Z_2|) = E(|X_1|)E(|X_2|), \end{aligned}$$

and we can only have equality throughout when W is a constant. □

Using (6.10), we can calculate that the characteristic function of a normal variance mixture is given by

$$\begin{aligned}\phi_{X}(\boldsymbol{t}) = E(E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}} \mid W)) &= E(\exp(\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu} - \tfrac{1}{2}W\boldsymbol{t}'\Sigma\boldsymbol{t})) \\ &= \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu}}\hat{H}(\tfrac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}),\end{aligned} \tag{6.20}$$

where $\hat{H}(\theta) = \int_0^\infty \mathrm{e}^{-\theta v}\,\mathrm{d}H(v)$ is the Laplace–Stieltjes transform of the df H of W. Based on (6.20) we use the notation $\boldsymbol{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{H})$ for normal variance mixtures.

Assuming that Σ is positive definite and that the distribution of W has no point mass at zero, we may derive the joint density of a normal variance mixture distribution. Writing $f_{X|W}$ for the (Gaussian) conditional density of $\boldsymbol{X}$ given W, the density of $\boldsymbol{X}$ is given by

$$\begin{aligned}f(\boldsymbol{x}) &= \int f_{X|W}(\boldsymbol{x} \mid w)\,\mathrm{d}H(w) \\ &= \int \frac{w^{-d/2}}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2w}\right\}\mathrm{d}H(w),\end{aligned} \tag{6.21}$$

in terms of the Lebesgue–Stieltjes integral; when H has density h we simply mean the Riemann integral $\int_0^\infty f_{X|W}(\boldsymbol{x} \mid w)h(w)\,\mathrm{d}w$. All such densities will depend on $\boldsymbol{x}$ only through the quadratic form $(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, and this means they are the densities of elliptical distributions, as will be discussed in Section 6.3.

Example 6.6 (multivariate two-point normal mixture distribution). Simple examples of normal mixtures are obtained when W is a discrete rv. For example, the two-point normal mixture model is obtained by taking W in (6.18) to be a discrete rv that assumes the distinct positive values k_1 and k_2 with probabilities p and $1-p$, respectively. By setting k_2 large relative to k_1 and choosing p large, this distribution might be used to define two regimes: an *ordinary* regime that holds most of the time and a *stress* regime that occurs with small probability $1-p$. Obviously this idea extends to k-point mixture models.

Example 6.7 (multivariate *t* distribution). If we take W in (6.18) to be an rv with an inverse gamma distribution $W \sim \mathrm{Ig}(\frac{1}{2}\nu, \frac{1}{2}\nu)$ (which is equivalent to saying that $\nu/W \sim \chi^2_\nu$), then $\boldsymbol{X}$ has a multivariate t distribution with ν degrees of freedom (see Section A.2.6 for more details concerning the inverse gamma distribution). Our notation for the multivariate t is $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$, and we note that Σ is not the covariance matrix of $\boldsymbol{X}$ in this definition of the multivariate t. Since $E(W) = \nu/(\nu-2)$ we have $\operatorname{cov}(\boldsymbol{X}) = (\nu/(\nu-2))\Sigma$, and the covariance matrix (and correlation matrix) of this distribution is only defined if $\nu > 2$.

Using (6.21), the density can be calculated to be

$$f(\boldsymbol{x}) = \frac{\Gamma(\frac{1}{2}(\nu+d))}{\Gamma(\frac{1}{2}\nu)(\pi\nu)^{d/2}|\Sigma|^{1/2}}\left(1 + \frac{(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{\nu}\right)^{-(\nu+d)/2}. \tag{6.22}$$

Clearly, the locus of points with equal density is again an ellipsoid with equation $(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = c$ for some $c > 0$. A bivariate example with four degrees of freedom is given in Figure 6.1. In comparison with the multivariate normal, the contours of equal density rise more quickly in the centre of the distribution and decay more gradually on the "lower slopes" of the distribution. In comparison with the multivariate normal, the multivariate t has heavier marginal tails (as discussed in Section 5.1.2) and a more pronounced tendency to generate simultaneous extreme values (see also Section 7.3.1).

Example 6.8 (symmetric generalized hyperbolic distribution). A flexible family of normal variance mixtures is obtained by taking W in (6.18) to have a generalized inverse Gaussian (GIG) distribution, $W \sim N^{-}(\lambda, \chi, \psi)$ (see Section A.2.5). Using (6.21), it can be shown that a normal variance mixture constructed with this mixing density has the joint density

$$f(\boldsymbol{x}) = \frac{(\sqrt{\chi\psi})^{-\lambda}\psi^{d/2}}{(2\pi)^{d/2}|\Sigma|^{1/2}K_{\lambda}(\sqrt{\chi\psi})} \frac{K_{\lambda-(d/2)}(\sqrt{(\chi + (\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))\psi})}{(\sqrt{(\chi + (\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))\psi})^{(d/2)-\lambda}}, \tag{6.23}$$

where K_{λ} denotes a modified Bessel function of the third kind (see Section A.2.5 for more details). This distribution is a special case of the more general family of multivariate generalized hyperbolic distributions, which we will discuss in greater detail in Section 6.2.2. The more general family can be obtained as *mean–variance* mixtures of normals, which are not necessarily elliptical distributions.

The GIG mixing distribution is very flexible and contains the gamma and inverse gamma distributions as special boundary cases (corresponding, respectively, to $\lambda > 0$, $\chi = 0$ and to $\lambda < 0$, $\psi = 0$). In these cases the density in (6.23) should be interpreted as a limit as $\chi \to 0$ or as $\psi \to 0$. (Information on the limits of Bessel functions is found in Section A.2.5.) The gamma mixing distribution yields Laplace distributions or so-called symmetric variance-gamma (VG) models, and the inverse gamma yields the t as in Example 6.7; to be precise, the t corresponds to the case when $\lambda = -\nu/2$ and $\chi = \nu$. The special cases $\lambda = -0.5$ and $\lambda = 1$ have also attracted attention in financial modelling. The former gives rise to the symmetric normal inverse Gaussian (NIG) distribution; the latter gives rise to a symmetric multivariate distribution whose one-dimensional margins are known simply as hyperbolic distributions.

To calculate the covariance matrix of distributions in the symmetric generalized hyperbolic family, we require the mean of the GIG distribution, which is given in (A.15) for the case $\chi > 0$ and $\psi > 0$. The covariance matrix of the multivariate distribution in (6.23) follows from (6.19).

Normal variance mixture distributions are easy to work with under linear operations, as shown in the following simple proposition.

Proposition 6.9. *If $\boldsymbol{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{H})$ and $\boldsymbol{Y} = B\boldsymbol{X} + \boldsymbol{b}$, where $B \in \mathbb{R}^{k\times d}$ and $\boldsymbol{b} \in \mathbb{R}^k$, then $\boldsymbol{Y} \sim M_k(B\boldsymbol{\mu} + \boldsymbol{b}, B\Sigma B', \hat{H})$.*

Proof. The characteristic function in (6.20) may be used to show that

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'(B\boldsymbol{X}+\boldsymbol{b})}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{b}}\phi_{\boldsymbol{X}}(B'\boldsymbol{t}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'(B\boldsymbol{\mu}+\boldsymbol{b})}\hat{H}(\tfrac{1}{2}\boldsymbol{t}'B\Sigma B'\boldsymbol{t}).$$

□

The subclass of mixture distributions specified by $\hat{H}$ is therefore closed under linear transformations. For example, if $\boldsymbol{X}$ has a multivariate t distribution with ν degrees of freedom, then so does any linear transformation of $\boldsymbol{X}$; the linear combination $\boldsymbol{a}'\boldsymbol{X}$ would have a univariate t distribution with ν degrees of freedom (more precisely, the distribution $\boldsymbol{a}'\boldsymbol{X} \sim t_1(\nu, \boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\Sigma\boldsymbol{a})$).

Normal variance mixture distributions (and the mean–variance mixtures considered later in Section 6.2.2) are easily simulated, the method being obvious from Definition 6.4. To generate a variate $\boldsymbol{X} \sim M_d(\boldsymbol{\mu}, \Sigma, \hat{H})$ with Σ positive definite, we use the following algorithm.

Algorithm 6.10 (simulation of normal variance mixtures).

(1) Generate $\boldsymbol{Z} \sim N_d(\boldsymbol{0}, \Sigma)$ using Algorithm 6.2.

(2) Generate independently a positive mixing variable W with df H (corresponding to the Laplace–Stieltjes transform $\hat{H}$).

(3) Set $\boldsymbol{X} = \boldsymbol{\mu} + \sqrt{W}\boldsymbol{Z}$.

To generate $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$, the mixing variable W should have an $\mathrm{Ig}(\frac{1}{2}\nu, \frac{1}{2}\nu)$ distribution; it is helpful to note that in this case $\nu/W \sim \chi^2_\nu$, a chi-squared distribution with ν degrees of freedom. Sampling from a generalized hyperbolic distribution with density (6.23) requires us to generate $W \sim N^-(\lambda, \chi, \psi)$. Sampling from the GIG distribution can be accomplished using a rejection algorithm proposed by Atkinson (1982).

6.2.2 Normal Mean–Variance Mixtures

All of the multivariate distributions we have considered so far have elliptical symmetry (see Section 6.3.2 for explanation) and this may well be an oversimplified model for real risk-factor return data. Among other things, elliptical symmetry implies that all one-dimensional marginal distributions are rigidly symmetric, which contradicts the frequent observation for stock returns that negative returns (losses) have heavier tails than positive returns (gains). The models we now introduce attempt to add some asymmetry to the class of normal mixtures by mixing normal distributions with different means as well as different variances; this yields the class of multivariate normal mean–variance mixtures.

Definition 6.11. The random vector $\boldsymbol{X}$ is said to have a (multivariate) normal mean–variance mixture distribution if

$$\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{m}(W) + \sqrt{W}A\boldsymbol{Z}, \tag{6.24}$$

where

(i) $\boldsymbol{Z} \sim N_k(\boldsymbol{0}, I_k)$,

(ii) $W \geqslant 0$ is a non-negative, scalar-valued rv which is independent of $\boldsymbol{Z}$,

(iii) $A \in \mathbb{R}^{d\times k}$ is a matrix, and

(iv) $\boldsymbol{m}\colon [0,\infty) \to \mathbb{R}^d$ is a measurable function.

In this case we have that

$$\boldsymbol{X} \mid W = w \sim N_d(\boldsymbol{m}(w), w\Sigma), \tag{6.25}$$

where $\Sigma = AA'$ and it is clear why such distributions are known as mean–variance mixtures of normals. In general, such distributions are not elliptical.

A possible concrete specification for the function $\boldsymbol{m}(W)$ in (6.25) is

$$\boldsymbol{m}(W) = \boldsymbol{\mu} + W\boldsymbol{\gamma}, \tag{6.26}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\gamma}$ are parameter vectors in $\mathbb{R}^d$. Since $E(\boldsymbol{X} \mid W) = \boldsymbol{\mu} + W\boldsymbol{\gamma}$ and $\operatorname{cov}(\boldsymbol{X} \mid W) = W\Sigma$, it follows in this case by simple calculations that

$$E(\boldsymbol{X}) = E(E(\boldsymbol{X} \mid W)) = \boldsymbol{\mu} + E(W)\boldsymbol{\gamma}, \tag{6.27}$$

$$\begin{aligned}\operatorname{cov}(\boldsymbol{X}) &= E(\operatorname{cov}(\boldsymbol{X} \mid W)) + \operatorname{cov}(E(\boldsymbol{X} \mid W)) \\ &= E(W)\Sigma + \operatorname{var}(W)\boldsymbol{\gamma}\boldsymbol{\gamma}'\end{aligned} \tag{6.28}$$

when the mixing variable W has finite variance. We observe from (6.27) and (6.28) that the parameters $\boldsymbol{\mu}$ and Σ are not, in general, the mean vector and covariance matrix of $\boldsymbol{X}$ (or a multiple thereof). This is only the case when $\boldsymbol{\gamma} = \boldsymbol{0}$, so that the distribution is a normal variance mixture and the simpler moment formulas given in (6.19) apply.

6.2.3 Generalized Hyperbolic Distributions

In Example 6.8 we looked at the special subclass of the generalized hyperbolic (GH) distributions consisting of the elliptically symmetric normal variance mixture distributions. The full GH family is obtained using the mean–variance mixture construction (6.24) and the conditional mean specification (6.26). For the mixing distribution we assume that $W \sim N^-(\lambda, \chi, \psi)$, a GIG distribution with density (A.14).

Remark 6.12. This class of distributions has received a lot of attention in the financial-modelling literature, particularly in the univariate case. An important reason for this attention is their link to Lévy processes, i.e. processes with independent and stationary increments (like Brownian motion or the compound Poisson distribution) that are used to model price processes in continuous time. For every GH distribution it is possible to construct a Lévy process so that the value of the increment of the process over a fixed time interval has that distribution; this is only possible because the GH law is a so-called infinitely divisible distribution, a property that it inherits from the GIG mixing distribution of W.

The joint density in the non-singular case ($\varSigma$ has rank d) is

$$f(\boldsymbol{x}) = \int_0^\infty \frac{\mathrm{e}^{(\boldsymbol{x}-\boldsymbol{\mu})'\varSigma^{-1}\boldsymbol{\gamma}}}{(2\pi)^{d/2}|\varSigma|^{1/2}w^{d/2}} \times \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})'\varSigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2w} - \frac{\boldsymbol{\gamma}'\varSigma^{-1}\boldsymbol{\gamma}}{2/w}\right\} h(w)\,\mathrm{d}w,$$

where $h(w)$ is the density of W. Evaluation of this integral gives the GH density

$$f(\boldsymbol{x}) = c\frac{K_{\lambda-(d/2)}(\sqrt{(\chi + (\boldsymbol{x}-\boldsymbol{\mu})'\varSigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))(\psi + \boldsymbol{\gamma}'\varSigma^{-1}\boldsymbol{\gamma})})\mathrm{e}^{(\boldsymbol{x}-\boldsymbol{\mu})'\varSigma^{-1}\boldsymbol{\gamma}}}{(\sqrt{(\chi + (\boldsymbol{x}-\boldsymbol{\mu})'\varSigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))(\psi + \boldsymbol{\gamma}'\varSigma^{-1}\boldsymbol{\gamma})})^{(d/2)-\lambda}}, \tag{6.29}$$

where the normalizing constant is

$$c = \frac{(\sqrt{\chi\psi})^{-\lambda}\psi^{\lambda}(\psi + \boldsymbol{\gamma}'\varSigma^{-1}\boldsymbol{\gamma})^{(d/2)-\lambda}}{(2\pi)^{d/2}|\varSigma|^{1/2}K_\lambda(\sqrt{\chi\psi})}.$$

Clearly, if $\boldsymbol{\gamma} = \boldsymbol{0}$, the distribution reduces to the symmetric GH special case of Example 6.8. In general, we have a non-elliptical distribution with asymmetric margins. The mean vector and covariance matrix of the distribution are easily calculated from (6.27) and (6.28) using the information on the GIG and its moments given in Section A.2.5. The characteristic function of the GH distribution may be calculated using the same approach as in (6.20) to yield

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu}}\hat{H}(\tfrac{1}{2}\boldsymbol{t}'\varSigma\boldsymbol{t} - \mathrm{i}\boldsymbol{t}'\boldsymbol{\gamma}), \tag{6.30}$$

where $\hat{H}$ is the Laplace–Stieltjes transform of the GIG distribution.

We adopt the notation $\boldsymbol{X} \sim \mathrm{GH}_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \varSigma, \boldsymbol{\gamma})$. Note that the distributions $\mathrm{GH}_d(\lambda, \chi/k, k\psi, \boldsymbol{\mu}, k\varSigma, k\boldsymbol{\gamma})$ and $\mathrm{GH}_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \varSigma, \boldsymbol{\gamma})$ are identical for any $k > 0$, which causes an *identifiability problem* when we attempt to estimate the parameters in practice. This can be solved by constraining the determinant $|\varSigma|$ to be a particular value (such as one) when fitting. Note that, while such a constraint will have an effect on the values of χ and ψ that we estimate, it will not have an effect on the value of $\chi\psi$, so this product is a useful summary parameter for the GH distribution.

Linear combinations. The GH class is closed under linear operations.

Proposition 6.13. *If* $\boldsymbol{X} \sim \mathrm{GH}_d(\lambda, \chi, \psi, \boldsymbol{\mu}, \varSigma, \boldsymbol{\gamma})$ *and* $\boldsymbol{Y} = B\boldsymbol{X} + \boldsymbol{b}$, *where* $B \in \mathbb{R}^{k\times d}$ *and* $\boldsymbol{b} \in \mathbb{R}^k$, *then* $\boldsymbol{Y} \sim \mathrm{GH}_k(\lambda, \chi, \psi, B\boldsymbol{\mu} + \boldsymbol{b}, B\varSigma B', B\boldsymbol{\gamma})$.

Proof. We calculate, using (6.30) and a similar method to Proposition 6.9, that

$$\phi_{\boldsymbol{Y}}(\boldsymbol{t}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'(B\boldsymbol{\mu}+\boldsymbol{b})}\hat{H}(\tfrac{1}{2}\boldsymbol{t}'B\varSigma B'\boldsymbol{t} - \mathrm{i}\boldsymbol{t}'B\boldsymbol{\gamma}).$$

□

The parameters inherited from the GIG mixing distribution therefore remain unchanged under linear operations. This means, for example, that margins of $\boldsymbol{X}$ are easy to calculate; we have that $X_i \sim \mathrm{GH}_1(\lambda, \chi, \psi, \mu_i, \varSigma_{ii}, \gamma_i)$.

Parametrizations. There is a bewildering array of alternative parametrizations for the GH distribution in the literature, and it is more common to meet this distribution in a reparametrized form. In one common version the dispersion matrix we call Σ is renamed Δ and the constraint that $|\Delta| = 1$ is imposed; this addresses the identifiability problem mentioned above. The skewness parameters $\boldsymbol{\gamma}$ are replaced by parameters $\boldsymbol{\beta}$, and the non-negative parameters χ and ψ are replaced by the non-negative parameters δ and α according to

$$\boldsymbol{\beta} = \Delta^{-1}\boldsymbol{\gamma}, \qquad \delta = \sqrt{\chi}, \qquad \alpha = \sqrt{\psi + \boldsymbol{\gamma}'\Delta^{-1}\boldsymbol{\gamma}}.$$

These parameters must satisfy the constraints $\delta \geqslant 0$, $\alpha^2 > \boldsymbol{\beta}'\Delta\boldsymbol{\beta}$ if $\lambda > 0$; $\delta > 0$, $\alpha^2 > \boldsymbol{\beta}'\Delta\boldsymbol{\beta}$ if $\lambda = 0$; and $\delta > 0$, $\alpha^2 \geqslant \boldsymbol{\beta}'\Delta\boldsymbol{\beta}$ if $\lambda < 0$. Blæsild (1981) uses this parametrization to show that GH distributions form a closed class of distributions under linear operations and conditioning. However, the parametrization does have the problem that the important parameters α and δ are not generally invariant under either of these operations.

It is useful to be able to move easily between our χ–ψ–Σ–$\boldsymbol{\gamma}$ parametrization, as in (6.29), and the α–δ–Δ–$\boldsymbol{\beta}$ parametrization; λ and $\boldsymbol{\mu}$ are common to both parametrizations. If the χ–ψ–Σ–$\boldsymbol{\gamma}$ parametrization is used, then the formulas for obtaining the other parametrization are

$$\Delta = |\Sigma|^{-1/d}\Sigma, \qquad \boldsymbol{\beta} = \Sigma^{-1}\boldsymbol{\gamma},$$
$$\delta = \sqrt{\chi|\Sigma|^{1/d}}, \qquad \alpha = \sqrt{|\Sigma|^{-1/d}(\psi + \boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma})}.$$

If the α–δ–Δ–$\boldsymbol{\beta}$ form is used, then we can obtain our parametrization by setting

$$\Sigma = \Delta, \qquad \gamma = \Delta\boldsymbol{\beta}, \qquad \chi = \delta^2, \qquad \psi = \alpha^2 - \boldsymbol{\beta}'\Delta\boldsymbol{\beta}.$$

Special cases. The multivariate GH family is extremely flexible and, as we have mentioned, contains many special cases known by alternative names.

- If $\lambda = \frac{1}{2}(d+1)$, we drop the word "generalized" and refer to the distribution as a d-dimensional hyperbolic distribution. Note that the univariate margins of this distribution also have $\lambda = \frac{1}{2}(d+1)$ and are not one-dimensional hyperbolic distributions.
- If $\lambda = 1$, we get a multivariate distribution whose univariate margins are one-dimensional hyperbolic distributions. The one-dimensional hyperbolic distribution has been widely used in univariate analyses of financial return data (see Notes and Comments).
- If $\lambda = -\frac{1}{2}$, then the distribution is known as an NIG distribution. In the univariate case this model has also been used in analyses of return data; its functional form is similar to the hyperbolic distribution but with a slightly heavier tail. (Note that the NIG and the GIG are different distributions!)
- If $\lambda > 0$ and $\chi = 0$, we get a limiting case of the distribution known variously as a generalized Laplace, Bessel function or VG distribution.

- If $\lambda = -\frac{1}{2}\nu$, $\chi = \nu$ and $\psi = 0$, we get another limiting case that seems to have been less well studied; it could be called an asymmetric or skewed t distribution. Evaluating the limit of (6.29) as $\psi \to 0$ yields the multivariate density

$$f(\boldsymbol{x}) = c\frac{K_{(\nu+d)/2}(\sqrt{(\nu + Q(\boldsymbol{x}))\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma}})\exp((\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}\boldsymbol{\gamma})}{(\sqrt{(\nu + Q(\boldsymbol{x}))\boldsymbol{\gamma}'\Sigma^{-1}\boldsymbol{\gamma}})^{-(\nu+d)/2}(1 + (Q(\boldsymbol{x})/\nu))^{(\nu+d)/2}}, \quad (6.31)$$

where $Q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ and the normalizing constant is

$$c = \frac{2^{1-(\nu+d)/2}}{\Gamma(\frac{1}{2}\nu)(\pi\nu)^{d/2}|\Sigma|^{1/2}}.$$

This density reduces to the standard multivariate t density in (6.22) as $\boldsymbol{\gamma} \to \mathbf{0}$.

6.2.4 Empirical Examples

In this section we fit the multivariate GH distribution to real data and examine which of the subclasses—such as t, hyperbolic or NIG—are most useful; we also explore whether the general mean–variance mixture models can be replaced by (elliptically symmetric) variance mixtures. Our first example prepares the ground for multivariate examples by looking briefly at univariate models. The univariate distributions are fitted by straightforward numerical maximization of the log-likelihood. The multivariate distributions are fitted by using a variant of the EM algorithm, as described in Section 15.1.1.

Example 6.14 (univariate stock returns). In the literature, the NIG, hyperbolic and t models have been particularly popular special cases. We fit symmetric and asymmetric cases of these distributions to the data used in Example 6.3, restricting attention to daily and weekly returns, where the data are more plentiful ($n = 2020$ and $n = 468$, respectively). Models are fitted using maximum likelihood under the simplifying assumption that returns form iid samples; a simple quasi-Newton method provides a viable alternative to the EM algorithm in the univariate case.

In the upper two panels of Table 6.2 we show results for symmetric models. The t, NIG and hyperbolic models may be compared directly using the log-likelihood at the maximum, since all have the same number of parameters: for daily data we find that eight out of ten stocks prefer the t distribution to the hyperbolic and NIG distributions; for weekly returns the t distribution is favoured in six out of ten cases. Overall, the second best model appears to be the NIG distribution. The mixture models fit much better than the Gaussian model in all cases, and it may be easily verified using the Akaike information criterion (AIC) that they are preferred to the Gaussian model in a formal comparison (see Section A.3.6 for more on the AIC).

For the asymmetric models, we only show cases where at least one of the asymmetric t, NIG or hyperbolic models offered a significant improvement ($p < 0.05$) on the corresponding symmetric model according to a likelihood ratio test. This occurred for weekly returns on Citigroup (C) and Intel (INTC) but for no daily returns. For Citigroup the p-values of the tests were, respectively, 0.06, 0.04 and

Table 6.2. Comparison of univariate models in the GH family, showing estimates of selected parameters and the value of the log-likelihood at the maximum; bold numbers indicate the models that give the largest values of the log-likelihood. See Example 6.14 for commentary.

	Gauss	t model		NIG model		Hyperbolic model	
Stock	$\ln L$	ν	$\ln L$	$\sqrt{\chi\psi}$	$\ln L$	$\sqrt{\chi\psi}$	$\ln L$
Daily returns: symmetric models							
AXP	4945.7	5.8	5001.8	1.6	**5002.4**	1.3	5002.1
EK	5112.9	3.8	**5396.2**	0.8	5382.5	0.6	5366.0
BA	5054.9	3.8	**5233.5**	0.8	5229.1	0.5	5221.2
C	4746.6	6.3	**4809.5**	1.9	4806.8	1.7	4805.0
KO	5319.6	5.1	**5411.0**	1.4	5407.3	1.3	5403.3
MSFT	4724.3	5.8	**4814.6**	1.6	4809.5	1.5	4806.4
HWP	4480.1	4.5	**4588.8**	1.1	4587.2	0.9	4583.4
INTC	4392.3	5.4	**4492.2**	1.5	4486.7	1.4	4482.4
JPM	4898.3	5.1	4967.8	1.3	4969.5	0.9	**4969.7**
DIS	5047.2	4.4	**5188.3**	1	5183.8	0.8	5177.6
Weekly returns: symmetric models							
AXP	719.9	8.8	724.2	3.0	**724.3**	2.8	724.3
EK	718.7	3.6	**765.6**	0.7	764.0	0.5	761.3
BA	732.4	4.4	**759.2**	1.0	758.3	0.8	757.2
C	656.0	5.7	**669.6**	1.6	669.3	1.3	669
KO	757.1	6.0	765.7	1.7	766.2	1.3	**766.3**
MSFT	671.5	6.3	**683.9**	1.9	683.2	1.8	682.9
HWP	627.1	6.0	637.3	1.8	**637.3**	1.5	637.1
INTC	595.8	5.2	**611.0**	1.5	610.6	1.3	610
JPM	681.7	5.9	**693.0**	1.7	692.9	1.5	692.6
DIS	734.1	6.4	742.7	1.9	**742.8**	1.7	742.7
Weekly returns: asymmetric models							
C	NA	6.1	**671.4**	1.7	671.3	1.3	671.2
INTC	NA	6.3	**614.2**	1.8	613.9	1.7	613.3

0.04 for the t, NIG and hyperbolic cases; for Intel the p-values were 0.01 in all cases, indicating quite strong asymmetry.

In the case of Intel we have superimposed the densities of various fitted asymmetric distributions on a histogram of the data in Figure 6.3. A plot of the log densities shown alongside reveals the differences between the distributions in the tail area. The left tail (corresponding to losses) appears to be heavier for these data, and the best-fitting distribution according to the likelihood comparison is the asymmetric t distribution.

Example 6.15 (multivariate stock returns). We fitted multivariate models to the full ten-dimensional data set of log-returns used in the previous example. The resulting values of the maximized log-likelihood are shown in Table 6.3 along with p-values for a likelihood ratio test of all special cases against the (asymmetric) GH model. The number of parameters in each model is also given; note that the general

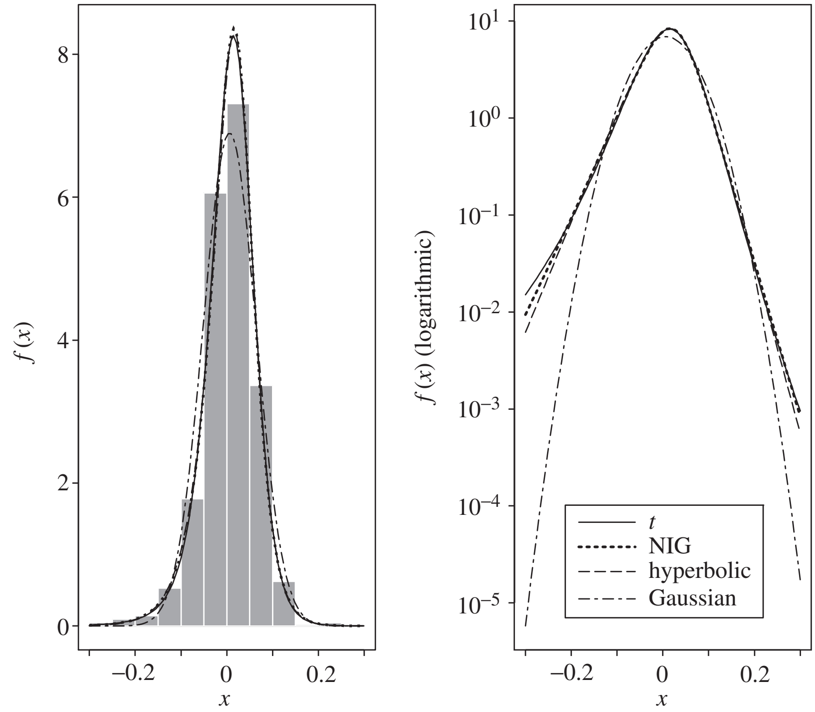

Figure 6.3. Models for weekly returns on Intel (INTC).

d-dimensional GH model has $\frac{1}{2}d(d+1)$ dispersion parameters, d location parameters, d skewness parameters and three parameters coming from the GIG mixing distribution, but is subject to one identifiability constraint; this gives $\frac{1}{2}(d(d+5)+4)$ free parameters.

For the daily data the best of the special cases is the skewed t distribution, which gives a value for the maximized likelihood that cannot be discernibly improved by the more general model with its additional parameter. All other non-elliptically symmetric submodels are rejected in a likelihood ratio test. Note, however, that the elliptically symmetric t distribution cannot be rejected when compared with the most general model, so that this seems to offer a simple parsimonious model for these data (the estimated degree of freedom is 6.0).

For the weekly data the best special case is the NIG distribution, followed closely by the skewed t; the hyperbolic and VG are rejected. The best elliptically symmetric special case seems to be the t distribution (the estimated degree of freedom this time being 6.2).

Example 6.16 (multivariate exchange-rate returns). We fitted the same multivariate models to a four-dimensional data set of exchange-rate log-returns, these being sterling, the euro, Japanese yen and Swiss franc against the US dollar for the period January 2000 to the end of March 2004 (1067 daily returns and 222 weekly returns). The resulting values of the maximized log-likelihood are shown in Table 6.4.

Table 6.3. A comparison of models in the GH family for ten-dimensional stock-return data. For each model, the table shows the value of the log-likelihood at the maximum ($\ln L$), the numbers of parameters ("# par.") and the p-value for a likelihood ratio test against the general GH model. The log-likelihood values for the general model, the best special case and the best elliptically symmetric special case are in bold type. See Example 6.15 for details.

	GH	NIG	Hyperbolic	t	VG	Gauss
Daily returns: asymmetric models						
$\ln L$	**52 174.62**	52 141.45	52 111.65	**52 174.62**	52 063.44	
# par.	77	76	76	76	76	
p-value		0.00	0.00	1.00	0.00	
Daily returns: symmetric models						
$\ln L$	52 170.14	52 136.55	52 106.34	**52 170.14**	52 057.38	50 805.28
# par.	67	66	66	66	66	65
p-value	0.54	0.00	0.00	0.63	0.00	0.00
Weekly returns: asymmetric models						
$\ln L$	**7 639.32**	**7 638.59**	7 636.49	7 638.56	7 631.33	
p-value		0.23	0.02	0.22	0.00	
Weekly returns: symmetric models						
$\ln L$	7 633.65	7 632.68	7 630.44	**7 633.11**	7 625.4	7 433.77
p-value	0.33	0.27	0.09	0.33	0.00	0.00

Table 6.4. A comparison of models in the GH family for four-dimensional exchange-rate return data. For each model, the table shows the value of the log-likelihood at the maximum ($\ln L$), the numbers of parameters ("# par.") and the p-value for a likelihood ratio test against the general GH model. The log-likelihood values for the general model, the best special case and the best elliptically symmetric special case are in bold type. See Example 6.16 for details.

	GH	NIG	Hyperbolic	t	VG	Gauss
Daily returns: asymmetric models						
$\ln L$	**17 306.44**	**17 306.43**	17 305.61	17 304.97	17 302.5	
# par.	20	19	19	19	19	
p-value		0.85	0.20	0.09	0.00	
Daily returns: symmetric models						
$\ln L$	17 303.10	**17 303.06**	17 302.15	17 301.85	17 299.15	17 144.38
# par.	16	15	15	15	15	14
p-value	0.15	0.24	0.13	0.10	0.01	0.00
Weekly returns: asymmetric models						
$\ln L$	**2 890.65**	2 889.90	2 889.65	**2 890.65**	2 888.98	
p-value		0.22	0.16	1.00	0.07	
Weekly returns: symmetric models						
$\ln L$	2 887.52	2 886.74	2 886.48	**2 887.52**	2 885.86	2 872.36
p-value	0.18	0.17	0.14	0.28	0.09	0.00

For the daily data the best of the special cases (both in general and if we restrict ourselves to symmetric models) is the NIG distribution, followed by the hyperbolic, *t* and VG distributions in that order. In a likelihood ratio test of the special cases against the general GH distribution, only the VG model is rejected at the 5% level; the skewed *t* model is rejected at the 10% level. When tested against the full model, certain elliptical models could not be rejected, the best of these being the NIG.

For the weekly data the best special case is the *t* distribution, followed by the NIG, hyperbolic and VG; none of the special cases can be rejected in a test at the 5% level, although the VG model is rejected at the 10% level. Among the elliptically symmetric distributions the Gauss distribution is clearly rejected, and the VG is again rejected at the 10% level, but otherwise the elliptical special cases are accepted; the best of these seems to be the *t* distribution, which has an estimated degrees-of-freedom parameter of 5.99.

Notes and Comments

Important early papers on multivariate normal mixtures are Kelker (1970) and Cambanis, Huang and Simons (1981). See also Bingham and Kiesel (2002), which contains an overview of the connections between the normal mixture, elliptical and hyperbolic models, and discusses their role in financial modelling. Fang, Kotz and Ng (1990) discuss the symmetric normal mixture models as special cases in their account of the more general family of spherical and elliptical distributions.

The GH distributions (univariate and multivariate) were introduced in Barndorff-Nielsen (1978) and further explored in Barndorff-Nielsen and Blæsild (1981). Useful references on the multivariate distribution are Blæsild (1981) and Blæsild and Jensen (1981). Generalized hyperbolic distributions (particularly in the univariate case) have been popularized as models for financial returns in recent papers by Eberlein and Keller (1995) and Eberlein, Keller and Prause (1998) (see also Bibby and Sørensen 2003). The PhD thesis of Prause (1999) is also a compendium of useful information in this context.

The reasons for their popularity in financial applications are both empirical and theoretical: they appear to provide a good fit to financial return data (again mostly in univariate investigations); they are consistent with continuous-time models, where logarithmic asset prices follow univariate or multivariate Lévy processes (thus generalizing the Black–Scholes model, where logarithmic prices follow Brownian motion); see Eberlein and Keller (1995) and Schoutens (2003).

For the NIG special case see Barndorff-Nielsen (1997), who discusses both univariate and multivariate cases and argues that the NIG is slightly superior to the hyperbolic as a univariate model for return data, a claim that our analyses support for stock-return data. Kotz, Kozubowski and Podgórski (2001) is a useful reference for the VG special case; the distribution appears here under the name generalized Laplace distribution and a (univariate or multivariate) Lévy process with VG-distributed increments is called a Laplace motion. The univariate Laplace motion is essentially the model proposed by Madan and Seneta (1990), who derived it as a Brownian motion under a stochastic time change and referred to it as the VG model (see also Madan, Carr and Chang 1998). The multivariate t distribution is discussed in Kotz and Nadarajah (2004); the asymmetric or skewed t distribution presented in this chapter is also discussed in Bibby and Sørensen (2003). For alternative skewed extensions of the multivariate t, see Kotz and Nadarajah (2004) and Genton (2004).

6.3 Spherical and Elliptical Distributions

In the previous section we observed that normal variance mixture distributions—particularly the multivariate t and symmetric multivariate NIG—provided models that were far superior to the multivariate normal for daily and weekly US stock-return data. The more general asymmetric mean–variance mixture distributions did not seem to offer much of an improvement on the symmetric variance mixture models. While this was a single example, other investigations suggest that multivariate return data for groups of returns of a similar type often show similar behaviour.

The normal variance mixture distributions are so-called elliptical distributions, and in this section we look more closely at the theory of elliptical distributions. To do this we begin with the special case of spherical distributions.

6.3.1 Spherical Distributions

The spherical family constitutes a large class of distributions for random vectors with *uncorrelated* components and identical, symmetric marginal distributions. It is important to note that within this class, $N_d(\mathbf{0}, I_d)$ is the only model for a vector of mutually independent components. Many of the properties of elliptical distributions can best be understood by beginning with spherical distributions.

Definition 6.17. A random vector $\boldsymbol{X} = (X_1, \dots, X_d)'$ has a spherical distribution if, for every orthogonal map $U \in \mathbb{R}^{d\times d}$ (i.e. maps satisfying $UU' = U'U = I_d$),

$$U\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{X}.$$

Thus spherical random vectors are distributionally invariant under rotations. There are a number of different ways of defining distributions with this property, as we demonstrate below.

Theorem 6.18. *The following are equivalent.*

(1) *$\boldsymbol{X}$ is spherical.*

(2) *There exists a function ψ of a scalar variable such that, for all $\boldsymbol{t} \in \mathbb{R}^d$,*

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \psi(\boldsymbol{t}'\boldsymbol{t}) = \psi(t_1^2 + \cdots + t_d^2). \tag{6.32}$$

(3) *For every $\boldsymbol{a} \in \mathbb{R}^d$,*

$$\boldsymbol{a}'\boldsymbol{X} \overset{\mathrm{d}}{=} \|\boldsymbol{a}\|X_1, \tag{6.33}$$

where $\|\boldsymbol{a}\|^2 = \boldsymbol{a}'\boldsymbol{a} = a_1^2 + \cdots + a_d^2$.

Proof. (1) ⇒ (2). If $\boldsymbol{X}$ is spherical, then for any orthogonal matrix U we have

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \phi_{U\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'U\boldsymbol{X}}) = \phi_{\boldsymbol{X}}(U'\boldsymbol{t}).$$

This can only be true if $\phi_{\boldsymbol{X}}(\boldsymbol{t})$ only depends on the length of $\boldsymbol{t}$, i.e. if $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \psi(\boldsymbol{t}'\boldsymbol{t})$ for some function ψ of a non-negative scalar variable.

(2) ⇒ (3). First observe that $\phi_{X_1}(t) = E(\mathrm{e}^{\mathrm{i}tX_1}) = \phi_{\boldsymbol{X}}(t\boldsymbol{e}_1) = \psi(t^2)$, where $\boldsymbol{e}_1$ denotes the first unit vector in $\mathbb{R}^d$. It follows that for any $\boldsymbol{a} \in \mathbb{R}^d$,

$$\phi_{\boldsymbol{a}'\boldsymbol{X}}(t) = \phi_{\boldsymbol{X}}(t\boldsymbol{a}) = \psi(t^2\boldsymbol{a}'\boldsymbol{a}) = \psi(t^2\|\boldsymbol{a}\|^2) = \phi_{X_1}(t\|\boldsymbol{a}\|) = \phi_{\|\boldsymbol{a}\|X_1}(t).$$

(3) ⇒ (1). For any orthogonal matrix U we have

$$\phi_{U\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}(U'\boldsymbol{t})'\boldsymbol{X}}) = E(\mathrm{e}^{\mathrm{i}\|U'\boldsymbol{t}\|X_1}) = E(\mathrm{e}^{\mathrm{i}\|\boldsymbol{t}\|X_1}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \phi_{\boldsymbol{X}}(\boldsymbol{t}).$$

□

Part (2) of Theorem 6.18 shows that the characteristic function of a spherically distributed random vector is fully described by a function ψ of a scalar variable. For this reason ψ is known as the *characteristic generator* of the spherical distribution and the notation $\boldsymbol{X} \sim S_d(\psi)$ is used. Part (3) of Theorem 6.18 shows that linear combinations of spherical random vectors always have a distribution of the same *type*, so that they have the same distribution up to changes of location and scale (see Section A.1.1). This important property will be used in Chapter 8 to prove the subadditivity of value-at-risk for linear portfolios of elliptically distributed risk factors. We now give examples of spherical distributions.

Example 6.19 (multivariate normal). A random vector $\boldsymbol{X}$ with the standard uncorrelated normal distribution $N_d(\boldsymbol{0}, I_d)$ is clearly spherical. The characteristic function is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = \mathrm{e}^{-\boldsymbol{t}'\boldsymbol{t}/2},$$

so that, using part (2) of Theorem 6.18, $\boldsymbol{X} \sim S_d(\psi)$ with characteristic generator $\psi(t) = \mathrm{e}^{-t/2}$.

Example 6.20 (normal variance mixtures). A random vector $\boldsymbol{X}$ with a standardized, uncorrelated normal variance mixture distribution $M_d(\boldsymbol{0}, I_d, \hat{H})$ also has a spherical distribution. Using (6.20), we see that $\phi_{\boldsymbol{X}}(\boldsymbol{t}) = \hat{H}(\frac{1}{2}\boldsymbol{t}'\boldsymbol{t})$, which obviously satisfies (6.32), and the characteristic generator of the spherical distribution is related to the Laplace–Stieltjes transform of the mixture distribution function of W by $\psi(t) = \hat{H}(\frac{1}{2}t)$. Thus $\boldsymbol{X} \sim M_d(\boldsymbol{0}, I_d, \hat{H}(\cdot))$ and $\boldsymbol{X} \sim S_d(\hat{H}(\frac{1}{2}\cdot))$ are two ways of writing the same mixture distribution.

A further, extremely important, way of characterizing spherical distributions is given by the following result.

Theorem 6.21. *$\boldsymbol{X}$ has a spherical distribution if and only if it has the stochastic representation*

$$\boldsymbol{X} \overset{\mathrm{d}}{=} R\boldsymbol{S}, \tag{6.34}$$

where $\boldsymbol{S}$ is uniformly distributed on the unit sphere $\mathcal{S}^{d-1} = \{\boldsymbol{s} \in \mathbb{R}^d : \boldsymbol{s}'\boldsymbol{s} = 1\}$ and $R \geqslant 0$ is a radial rv, independent of $\boldsymbol{S}$.

Proof. First we prove that if $\boldsymbol{S}$ is uniformly distributed on the unit sphere and $R \geqslant 0$ is an independent scalar variable, then $R\boldsymbol{S}$ has a spherical distribution. This is seen by considering the characteristic function

$$\phi_{R\boldsymbol{S}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}R\boldsymbol{t}'\boldsymbol{S}}) = E(E(\mathrm{e}^{\mathrm{i}R\boldsymbol{t}'\boldsymbol{S}} \mid R)).$$

Since $\boldsymbol{S}$ is itself spherically distributed, its characteristic function has a characteristic generator, which is usually given the special notation Ω_d. Thus, by Theorem 6.18 (2) we have that

$$\phi_{R\boldsymbol{S}}(\boldsymbol{t}) = E(\Omega_d(R^2\boldsymbol{t}'\boldsymbol{t})) = \int \Omega_d(r^2\boldsymbol{t}'\boldsymbol{t})\,\mathrm{d}F(r), \tag{6.35}$$

where F is the df of R. Since this is a function of $\boldsymbol{t}'\boldsymbol{t}$, it follows, again from Theorem 6.18 (2), that $R\boldsymbol{S}$ has a spherical distribution.

We now prove that if the random vector $\boldsymbol{X}$ is spherical, then it has the representation (6.34). For any arbitrary $\boldsymbol{s} \in \mathcal{S}^{d-1}$, the characteristic generator ψ of $\boldsymbol{X}$ must satisfy $\psi(\boldsymbol{t}'\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\boldsymbol{t}) = \phi_{\boldsymbol{X}}(\|\boldsymbol{t}\|\boldsymbol{s})$. It follows that, if we introduce a random vector $\boldsymbol{S}$ that is uniformly distributed on the sphere $\mathcal{S}^{d-1}$, we can write

$$\psi(\boldsymbol{t}'\boldsymbol{t}) = \int_{\mathcal{S}^{d-1}} \phi_{\boldsymbol{X}}(\|\boldsymbol{t}\|\boldsymbol{s})\,\mathrm{d}F_{\boldsymbol{S}}(\boldsymbol{s}) = \int_{\mathcal{S}^{d-1}} E(\mathrm{e}^{\mathrm{i}\|\boldsymbol{t}\|\boldsymbol{s}'\boldsymbol{X}})\,\mathrm{d}F_{\boldsymbol{S}}(\boldsymbol{s}).$$

Interchanging the order of integration and using the Ω_d notation for the characteristic generator of $\boldsymbol{S}$ we have

$$\psi(\boldsymbol{t}'\boldsymbol{t}) = E(\Omega_d(\|\boldsymbol{t}\|^2\|\boldsymbol{X}\|^2)) = \int \Omega_d(\boldsymbol{t}'\boldsymbol{t}r^2)\,\mathrm{d}F_{\|\boldsymbol{X}\|}(r), \tag{6.36}$$

where $F_{\|\boldsymbol{X}\|}$ is the df of $\|\boldsymbol{X}\|$. By comparison with (6.35) we see that (6.36) is the characteristic function of $R\boldsymbol{S}$, where R is an rv with df $F_{\|\boldsymbol{X}\|}$ that is independent of $\boldsymbol{S}$. □

We often exclude from consideration distributions that place point mass at the origin; that is, we consider spherical rvs $\boldsymbol{X}$ in the subclass $S_d^+(\psi)$ for which $P(\boldsymbol{X} = \boldsymbol{0}) = 0$. A particularly useful corollary of Theorem 6.21 is then the following result, which is used in Section 15.1.2 to devise tests for spherical and elliptical symmetry.

Corollary 6.22. *Suppose* $\boldsymbol{X} \overset{\mathrm{d}}{=} R\boldsymbol{S} \sim S_d^+(\psi)$. *Then*

$$\left(\|\boldsymbol{X}\|, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}\right) \overset{\mathrm{d}}{=} (R, \boldsymbol{S}). \tag{6.37}$$

Proof. Let $f_1(\boldsymbol{x}) = \|\boldsymbol{x}\|$ and $f_2(\boldsymbol{x}) = \boldsymbol{x}/\|\boldsymbol{x}\|$. It follows from (6.34) that

$$\left(\|\boldsymbol{X}\|, \frac{\boldsymbol{X}}{\|\boldsymbol{X}\|}\right) = (f_1(\boldsymbol{X}), f_2(\boldsymbol{X})) \overset{\mathrm{d}}{=} (f_1(R\boldsymbol{S}), f_2(R\boldsymbol{S})) = (R, \boldsymbol{S}).$$

□

Example 6.23 (working with $\boldsymbol{R}$ and $\boldsymbol{S}$). Suppose $\boldsymbol{X} \sim N_d(\mathbf{0}, I_d)$. Since $\boldsymbol{X}'\boldsymbol{X} \sim \chi_d^2$, a chi-squared distribution with d degrees of freedom, it follows from (6.37) that $R^2 \sim \chi_d^2$.

We can use this fact to calculate $E(\boldsymbol{S})$ and $\operatorname{cov}(\boldsymbol{S})$, the first two moments of a uniform distribution on the unit sphere. We have that

$$\begin{aligned} \mathbf{0} = E(\boldsymbol{X}) = E(R)E(\boldsymbol{S}) &\Rightarrow E(\boldsymbol{S}) = \mathbf{0}, \\ I_d = \operatorname{cov}(\boldsymbol{X}) = E(R^2)\operatorname{cov}(\boldsymbol{S}) &\Rightarrow \operatorname{cov}(\boldsymbol{S}) = I_d/d, \end{aligned} \tag{6.38}$$

since $E(R^2) = d$ when $R^2 \sim \chi_d^2$.

Now suppose that $\boldsymbol{X}$ has a spherical normal variance mixture distribution $\boldsymbol{X} \sim M_d(\mathbf{0}, I_d, \hat{H})$ and we wish to calculate the distribution of $R^2 \overset{\mathrm{d}}{=} \boldsymbol{X}'\boldsymbol{X}$ in this case. Since $\boldsymbol{X} \overset{\mathrm{d}}{=} \sqrt{W}\boldsymbol{Y}$, where $\boldsymbol{Y} \sim N_d(\mathbf{0}, I_d)$ and W is independent of $\boldsymbol{Y}$, it follows that $R^2 \overset{\mathrm{d}}{=} W\tilde{R}^2$, where $\tilde{R}^2 \sim \chi_d^2$ and W and $\tilde{R}$ are independent. If we can calculate the distribution of the product of W and an independent chi-squared variate, then we have the distribution of R^2.

For a concrete example suppose that $\boldsymbol{X} \sim t_d(\nu, \mathbf{0}, I_d)$. For a multivariate t distribution we know from Example 6.7 that $W \sim \operatorname{Ig}(\frac{1}{2}\nu, \frac{1}{2}\nu)$, which means that $\nu/W \sim \chi_\nu^2$. Using the fact that the ratio of independent chi-squared rvs divided by their degrees of freedom is F-distributed, it may be calculated that $R^2/d \sim F(d, \nu)$, the F distribution on d with ν degrees of freedom (see Section A.2.3). Since an $F(d, \nu)$ distribution has mean $\nu/(\nu-2)$, it follows from (6.38) that

$$\operatorname{cov}(\boldsymbol{X}) = E(\operatorname{cov}(R\boldsymbol{S} \mid R)) = E(R^2 I_d/d) = (\nu/(\nu-2))I_d.$$

The normal mixtures with $\boldsymbol{\mu} = \mathbf{0}$ and $\Sigma = I_d$ represent an easily understood subgroup of the spherical distributions. There are other spherical distributions that cannot be represented as normal variance mixtures; an example is the distribution of the uniform vector $\boldsymbol{S}$ on $\mathcal{S}^{d-1}$ itself. However, the normal mixtures have a special role in the spherical world, as summarized by the following theorem.

Theorem 6.24. *Denote by Ψ_∞ the set of characteristic generators that generate a d-dimensional spherical distribution for arbitrary $d \geqslant 1$. Then $\boldsymbol{X} \sim S_d(\psi)$ with $\psi \in \Psi_\infty$ if and only if $\boldsymbol{X} \overset{\mathrm{d}}{=} \sqrt{W}\boldsymbol{Z}$, where $\boldsymbol{Z} \sim N_d(\mathbf{0}, I_d)$ is independent of $W \geqslant 0$.*

Proof. This is proved in Fang, Kotz and Ng (1990, pp. 48–51). □

Thus, the characteristic generators of normal mixtures generate spherical distributions in arbitrary dimensions, while other spherical generators may only be used in certain dimensions. A concrete example is given by the uniform distribution on the unit sphere. Let Ω_d denote the characteristic generator of the uniform vector $\boldsymbol{S} = (S_1, \dots, S_d)'$ on $\mathcal{S}_{d-1}$. It can be shown that $\Omega_d((t_1, \dots, t_{d+1})'(t_1, \dots, t_{d+1}))$ is not the characteristic function of a spherical distribution in $\mathbb{R}^{d+1}$ (for more details see Fang, Kotz and Ng (1990, pp. 70–72)).

If a spherical distribution has a density f, then, by using the inversion formula

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^d}\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{i}\boldsymbol{t}'\boldsymbol{x}}\phi_{\boldsymbol{X}}(\boldsymbol{t})\,\mathrm{d}t_1 \cdots \mathrm{d}t_d,$$

it is easily inferred from Theorem 6.18 that $f(\boldsymbol{x}) = f(U\boldsymbol{x})$ for any orthogonal matrix U, so that the density must be of the form

$$f(\boldsymbol{x}) = g(\boldsymbol{x}'\boldsymbol{x}) = g(x_1^2 + \cdots + x_d^2) \tag{6.39}$$

for some function g of a scalar variable, which is referred to as the *density generator*. Clearly, the joint density is constant on hyperspheres $\{\boldsymbol{x} : x_1^2 + \cdots + x_d^2 = c\}$ in $\mathbb{R}^d$. To give a single example, the density generator of the multivariate t (i.e. the model $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{0}, I_d)$ of Example 6.7) is

$$g(x) = \frac{\Gamma(\frac{1}{2}(\nu + d))}{\Gamma(\frac{1}{2}\nu)(\pi\nu)^{d/2}} \left(1 + \frac{x}{\nu}\right)^{-(\nu+d)/2}.$$

6.3.2 Elliptical Distributions

Definition 6.25. $\boldsymbol{X}$ has an elliptical distribution if

$$\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + A\boldsymbol{Y},$$

where $\boldsymbol{Y} \sim S_k(\psi)$ and $A \in \mathbb{R}^{d\times k}$ and $\boldsymbol{\mu} \in \mathbb{R}^d$ are a matrix and vector of constants, respectively.

In other words, elliptical distributions are obtained by multivariate *affine* transformations of spherical distributions. Since the characteristic function is

$$\phi_{\boldsymbol{X}}(\boldsymbol{t}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}}) = E(\mathrm{e}^{\mathrm{i}\boldsymbol{t}'(\boldsymbol{\mu}+A\boldsymbol{Y})}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu}}E(\mathrm{e}^{\mathrm{i}(A'\boldsymbol{t})'\boldsymbol{Y}}) = \mathrm{e}^{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu}}\psi(\boldsymbol{t}'\Sigma\boldsymbol{t}),$$

where $\Sigma = AA'$, we denote the elliptical distributions by

$$\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$$

and refer to $\boldsymbol{\mu}$ as the location vector, Σ as the dispersion matrix and ψ as the characteristic generator of the distribution.

Remark 6.26. Knowledge of $\boldsymbol{X}$ does not uniquely determine its elliptical representation $E_d(\boldsymbol{\mu}, \Sigma, \psi)$. Although $\boldsymbol{\mu}$ is uniquely determined, Σ and ψ are only determined up to a positive constant. For example, the multivariate normal distribution $N_d(\boldsymbol{\mu}, \Sigma)$ can be written as $E_d(\boldsymbol{\mu}, \Sigma, \psi(\cdot))$ or $E_d(\boldsymbol{\mu}, c\Sigma, \psi(\cdot/c))$ for $\psi(u) = \mathrm{e}^{-u/2}$ and any $c > 0$. Provided that variances are finite, then an elliptical distribution is fully specified by its mean vector, covariance matrix and characteristic generator, and it is possible to find an elliptical representation $E_d(\boldsymbol{\mu}, \Sigma, \psi)$ such that Σ is the covariance matrix of $\boldsymbol{X}$, although this is not always the standard representation of the distribution.

We now give an alternative stochastic representation for the elliptical distributions that follows directly from Definition 6.25 and Theorem 6.21.

Proposition 6.27. *$\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ if and only if there exist $\boldsymbol{S}$, R and A satisfying*

$$\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{\mu} + RA\boldsymbol{S}, \tag{6.40}$$

with

(i) $\boldsymbol{S}$ *uniformly distributed on the unit sphere* $\mathcal{S}^{k-1} = \{\boldsymbol{s} \in \mathbb{R}^k : \boldsymbol{s}'\boldsymbol{s} = 1\}$,

(ii) $R \geqslant 0$, *a radial rv, independent of* $\boldsymbol{S}$, *and*

(iii) $A \in \mathbb{R}^{d \times k}$ *with* $AA' = \Sigma$.

For practical examples we are most interested in the case where Σ is positive definite. The relation between the elliptical and spherical cases is then clearly

$$\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi) \iff \Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu}) \sim S_d(\psi). \tag{6.41}$$

In this case, if the spherical vector $\boldsymbol{Y}$ has density generator g, then $\boldsymbol{X} = \boldsymbol{\mu} + \Sigma^{1/2}\boldsymbol{Y}$ has density

$$f(\boldsymbol{x}) = \frac{1}{|\Sigma|^{1/2}} g((\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})).$$

The joint density is always constant on sets of the form $\{\boldsymbol{x} : (\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = c\}$, which are ellipsoids in $\mathbb{R}^d$. Clearly, the full family of multivariate normal variance mixtures with general location and dispersion parameters $\boldsymbol{\mu}$ and Σ are elliptical, since they are obtained by affine transformations of the spherical special cases considered in the previous section.

It follows from (6.37) and (6.41) that for a non-singular elliptical variate $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ with no point mass at $\boldsymbol{\mu}$, we have

$$\left(\sqrt{(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})}, \frac{\Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})}{\sqrt{(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})}}\right) \stackrel{\mathrm{d}}{=} (R, \boldsymbol{S}), \tag{6.42}$$

where $\boldsymbol{S}$ is uniformly distributed on $\mathcal{S}^{d-1}$ and R is an independent scalar rv. This forms the basis of a test of elliptical symmetry described in Section 15.1.2.

The following proposition shows that a particular conditional distribution of an elliptically distributed random vector $\boldsymbol{X}$ has the same correlation matrix as $\boldsymbol{X}$ and can also be used to test for elliptical symmetry.

Proposition 6.28. *Let* $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ *and assume that* Σ *is positive definite and* $\operatorname{cov}(\boldsymbol{X})$ *is finite. For any* $c \geqslant 0$ *such that* $P((\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \geqslant c) > 0$, *we have*

$$\rho(\boldsymbol{X} \mid (\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \geqslant c) = \rho(\boldsymbol{X}). \tag{6.43}$$

Proof. It follows easily from (6.42) that

$$\boldsymbol{X} \mid (\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \geqslant c \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + R\Sigma^{1/2}\boldsymbol{S} \mid R^2 \geqslant c,$$

where $R \stackrel{\mathrm{d}}{=} \sqrt{(\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})}$ and $\boldsymbol{S}$ is independent of R and uniformly distributed on $\mathcal{S}^{d-1}$. Thus we have

$$\boldsymbol{X} \mid (\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \geqslant c \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \tilde{R}\Sigma^{1/2}\boldsymbol{S},$$

where $\tilde{R} \stackrel{\mathrm{d}}{=} R \mid R^2 \geqslant c$. It follows from Proposition 6.27 that the conditional distribution remains elliptical with dispersion matrix Σ and that (6.43) holds. □

6.3.3 Properties of Elliptical Distributions

We now summarize some of the properties of elliptical distributions in a format that allows their comparison with the properties of multivariate normal distributions in Section 6.1.3. Many properties carry over directly and others only need to be modified slightly. These parallels emphasize that it would be fairly easy to base many standard procedures in risk management on an assumption that risk-factor changes have an approximately elliptical distribution, rather than the patently false assumption that they are multivariate normal.

Linear combinations. If we take linear combinations of elliptical random vectors, then these remain elliptical with the same characteristic generator ψ. Let $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ and take any $B \in \mathbb{R}^{k \times d}$ and $\boldsymbol{b} \in \mathbb{R}^k$. Using a similar argument to that in Proposition 6.9 it is then easily shown that

$$B\boldsymbol{X} + \boldsymbol{b} \sim E_k(B\boldsymbol{\mu} + \boldsymbol{b}, B\Sigma B', \psi). \tag{6.44}$$

As a special case, if $\boldsymbol{a} \in \mathbb{R}^d$, then

$$\boldsymbol{a}'\boldsymbol{X} \sim E_1(\boldsymbol{a}'\boldsymbol{\mu}, \boldsymbol{a}'\Sigma\boldsymbol{a}, \psi). \tag{6.45}$$

Marginal distributions. It follows from (6.45) that marginal distributions of $\boldsymbol{X}$ must be elliptical distributions with the same characteristic generator. Using the $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2')'$ notation from Section 6.1.3 and again extending this notation naturally to $\boldsymbol{\mu}$ and Σ,

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

we have that $\boldsymbol{X}_1 \sim E_k(\boldsymbol{\mu}_1, \Sigma_{11}, \psi)$ and $\boldsymbol{X}_2 \sim E_{d-k}(\boldsymbol{\mu}_2, \Sigma_{22}, \psi)$.

Conditional distributions. The conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1$ may also be shown to be elliptical, although in general it will have a *different* characteristic generator $\tilde{\psi}$. For details of how the generator changes see Fang, Kotz and Ng (1990, pp. 45, 46). In the special case of multivariate normality the generator remains the same.

Quadratic forms. If $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ with Σ non-singular, then we observed in (6.42) that

$$Q := (\boldsymbol{X} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \overset{\mathrm{d}}{=} R^2, \tag{6.46}$$

where R is the radial rv in the stochastic representation (6.40). As we have seen in Example 6.23, for some particular cases the distribution of R^2 is well known: if $\boldsymbol{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, then $R^2 \sim \chi_d^2$; if $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{\mu}, \Sigma)$, then $R^2/d \sim F(d, \nu)$. For all elliptical distributions, Q must be independent of $\Sigma^{-1/2}(\boldsymbol{X} - \boldsymbol{\mu})/\sqrt{Q}$.

Convolutions. The convolution of two independent elliptical vectors with the *same dispersion matrix* Σ is also elliptical. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent d-dimensional random vectors satisfying $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ and $\boldsymbol{Y} \sim E_d(\tilde{\boldsymbol{\mu}}, \Sigma, \tilde{\psi})$, then we may take the product of characteristic functions to show that

$$\boldsymbol{X} + \boldsymbol{Y} \sim E_d(\boldsymbol{\mu} + \tilde{\boldsymbol{\mu}}, \Sigma, \bar{\psi}), \tag{6.47}$$

where $\bar{\psi}(u) = \psi(u)\tilde{\psi}(u)$.

If the dispersion matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ differ by more than a constant factor, then the convolution will not necessarily remain elliptical, even when the two generators ψ and $\tilde{\psi}$ are identical.

6.3.4 Estimating Dispersion and Correlation

Suppose we have risk-factor return data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ that we believe come from some elliptical distribution $E_d(\boldsymbol{\mu}, \Sigma, \psi)$ with heavier tails than the multivariate normal. We recall from Remark 6.26 that the dispersion matrix Σ is not uniquely determined, but rather is only fixed up to a constant of proportionality; when covariances are finite, the covariance matrix is proportional to Σ.

In this section we briefly consider the problem of estimating the location parameter $\boldsymbol{\mu}$, a dispersion matrix Σ and the correlation matrix P, assuming finiteness of second moments. We could use the standard estimators of Section 6.1.2. Under an assumption of iid or uncorrelated vector observations we observed that $\bar{\boldsymbol{X}}$ and S in (6.9) are unbiased estimators of the mean vector and the covariance matrix, respectively. They will also be consistent under quite weak assumptions. However, this does not necessarily mean they are the best estimators of location and dispersion for any given finite sample of elliptical data. There are many alternative estimators that may be more efficient for heavy-tailed data and may enjoy better robustness properties for contaminated data.

One strategy would be to fit a number of normal variance mixture models, such as the t and NIG, using the approach of Section 6.2.4. From the best-fitting model we would obtain an estimate of the mean vector and could easily calculate the implied estimates of the covariance and correlation matrices. In this section we give simpler, alternative methods that do not require a full fitting of a multivariate distribution; consult Notes and Comments for further references to robust dispersion estimation.

M-estimators. Maronna's M-estimators (Maronna 1976) of location and dispersion are a relatively old idea in robust statistics, but they have the virtue of being particularly simple to implement. Let $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ denote estimates of the mean vector and the dispersion matrix. Suppose for every observation $\boldsymbol{X}_i$ we calculate $D_i^2 = (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})' \hat{\Sigma}^{-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}})$. If we wanted to calculate improved estimates of location and dispersion, particularly for heavy-tailed data, it might be expected that this could be achieved by reducing the influence of observations for which D_i is large, since these are the observations that might tend to distort the parameter estimates most. M-estimation uses decreasing weight functions $w_j \colon \mathbb{R}^+ \to \mathbb{R}^+$, $j = 1, 2$, to reduce the weight of observations with large D_i values. This can be turned into an iterative procedure that converges to so-called M-estimates of location and dispersion; the dispersion matrix estimate is generally a biased estimate of the true covariance matrix.

Algorithm 6.29 (M-estimators of location and dispersion).

(1) As starting estimates take $\hat{\boldsymbol{\mu}}^{[1]} = \bar{\boldsymbol{X}}$ and $\hat{\Sigma}^{[1]} = S$, the standard estimators in (6.9). Set iteration count $k = 1$.

(2) For $i = 1, \dots, n$ set $D_i^2 = (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})' \hat{\Sigma}^{[k]-1} (\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})$.

(3) Update the location estimate using

$$\hat{\boldsymbol{\mu}}^{[k+1]} = \frac{\sum_{i=1}^{n} w_1(D_i)\boldsymbol{X}_i}{\sum_{i=1}^{n} w_1(D_i)},$$

where w_1 is a weight function, as discussed below.

(4) Update the dispersion matrix estimate using

$$\hat{\Sigma}^{[k+1]} = \frac{1}{n}\sum_{i=1}^{n} w_2(D_i^2)(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})(\boldsymbol{X}_i - \hat{\boldsymbol{\mu}}^{[k]})',$$

where w_2 is a weight function.

(5) Set $k = k + 1$ and repeat steps (2)–(4) until estimates converge.

Popular choices for the weight functions w_1 and w_2 are the decreasing functions $w_1(x) = (d + \nu)/(x^2 + \nu) = w_2(x^2)$ for some positive constant ν. Interestingly, use of these weight functions in Algorithm 6.29 exactly corresponds to fitting a multivariate $t_d(\nu, \boldsymbol{\mu}, \Sigma)$ distribution with known degrees of freedom ν using the EM algorithm (see, for example, Meng and van Dyk 1997).

There are many other possibilities for the weight functions. For example, the observations in the central part of the distribution could be given full weight and only the more outlying observations downweighted. This can be achieved by setting $w_1(x) = 1$ for $x \leqslant a$, $w_1(x) = a/x$ for $x > a$, for some value a, and $w_2(x^2) = (w_1(x))^2$.

Correlation estimates via Kendall's tau. A method for estimating correlation that is particularly easy to carry out is based on Kendall's rank correlation coefficient; this method will turn out to be related to a method in Chapter 7 that is used for estimating the parameters of certain copulas. The theoretical version of Kendall's rank correlation (also known as Kendall's tau) for two rvs X_1 and X_2 is denoted by $\rho_\tau(X_1, X_2)$ and is defined formally in Section 7.2.3; it is shown in Proposition 7.43 that if $(X_1, X_2) \sim E_2(\boldsymbol{\mu}, \Sigma, \psi)$, then

$$\rho_\tau(X_1, X_2) = \frac{2}{\pi}\arcsin(\rho), \tag{6.48}$$

where $\rho = \sigma_{12}/(\sigma_{11}\sigma_{22})^{1/2}$ is the *pseudo-correlation coefficient* of the elliptical distribution, which is always defined (even when correlation coefficients are undefined because variances are infinite). This relationship can be inverted to provide a method for estimating ρ from data; we simply replace the left-hand side of (6.48) by the standard textbook estimator of Kendall's tau, which is given in (7.52), to get an estimating equation that is solved for $\hat{\rho}$. This method estimates correlation by exploiting the geometry of an elliptical distribution and does not require us to estimate variances and covariances.

The method can be used to estimate a correlation matrix of a higher-dimensional elliptical distribution by applying the technique to each bivariate margin. This does, however, result in a matrix of pairwise correlation estimates that is not necessarily positive definite; this problem does not always arise, and if it does, a matrix

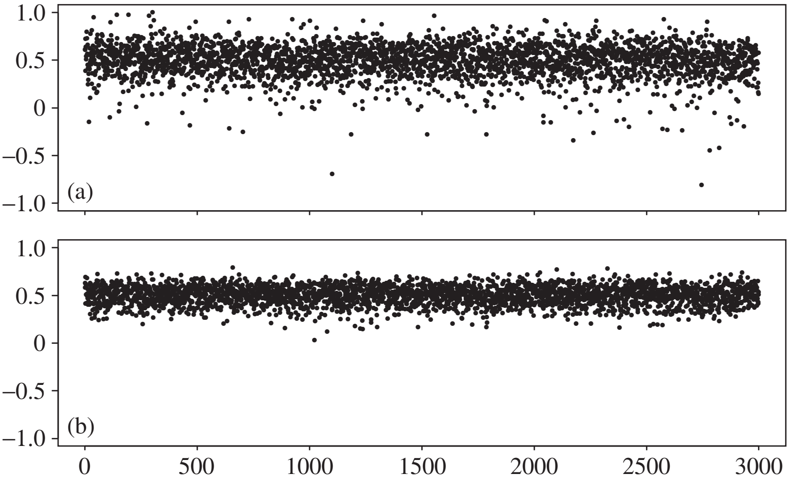

Figure 6.4. For 3000 independent samples of size 90 from a bivariate t distribution with three degrees of freedom and linear correlation 0.5: (a) the standard (Pearson) estimator of correlation; (b) the Kendall's tau transform estimator. See Example 6.30 for commentary.

adjustment method can be used, such as the eigenvalue method of Rousseeuw and Molenberghs (1993), which is given in Algorithm 7.57.

Note that to turn an estimate of a bivariate correlation matrix into a robust estimate of a dispersion matrix we could estimate the ratio of standard deviations $\lambda = (\sigma_{22}/\sigma_{11})^{1/2}$, e.g. by using a ratio of *trimmed* sample standard deviations; in other words, we leave out an equal number of outliers from each of the univariate data sets $X_{1,i}, \dots, X_{n,i}$ for $i = 1, 2$ and calculate the sample standard deviations with the remaining observations. This would give us the estimate

$$\hat{\Sigma} = \begin{pmatrix} 1 & \hat{\lambda}\hat{\rho} \\ \hat{\lambda}\hat{\rho} & \hat{\lambda}^2 \end{pmatrix}. \tag{6.49}$$

Example 6.30 (efficient correlation estimation for heavy-tailed data). Suppose we calculate correlations of asset or risk-factor returns based on 90 days (somewhat more than three trading months) of data; it would seem that this ought to be enough data to allow us to accurately estimate the "true" underlying correlation under an assumption that we have identically distributed data for that period.

Figure 6.4 displays the results of a simulation experiment where we have generated 3000 bivariate samples of iid data from a t distribution with three degrees of freedom and correlation $\rho = 0.5$; this is a heavy-tailed elliptical distribution. The distribution of the values of the standard correlation coefficient (also known as the Pearson correlation coefficient) is not particularly closely concentrated around the true value and produces some very poor estimates for a number of samples. On the other hand, the Kendall's tau transform method produces estimates that are generally much closer to the true value, and thus provides a more efficient way of estimating ρ.

Notes and Comments

A comprehensive reference for spherical and elliptical distributions is Fang, Kotz and Ng (1990); we have based our brief presentation of the theory on this account. Other references for the theory are Kelker (1970), Cambanis, Huang and Simons (1981) and Bingham and Kiesel (2002), the latter in the context of financial modelling. The original reference for Theorem 6.21 is Schoenberg (1938). Frahm (2004) suggests a generalization of the elliptical class to allow asymmetric models while preserving many of the attractive properties of elliptical distributions. For a more historical discussion (going back to Archimedes) and some surprising properties of the uniform distribution on the unit d-sphere, see Letac (2004).

There is a vast literature on alternative estimators of dispersion and correlation matrices, particularly with regard to better robustness properties. Textbooks with relevant sections include Hampel et al. (1986), Marazzi (1993), Wilcox (1997) and Huber and Ronchetti (2009); the last of those books is recommended more generally for applications of robust statistics in econometrics and finance.

We have concentrated on M-estimation of dispersion matrices, since this is related to the maximum likelihood estimation of alternative elliptical models. M-estimators have a relatively long history and are known to have good local robustness properties (insensitivity to small data perturbations); they do, however, have relatively low breakdown points in high dimensions, so their performance can be poor when data are more contaminated. A small selection of papers on M-estimation is Maronna (1976), Devlin, Gnanadesikan and Kettenring (1975, 1981) and Tyler (1983, 1987); see also Frahm (2004), in which an interesting alternative derivation of a Tyler estimator is given. The method based on Kendall's tau was suggested in Lindskog, McNeil and Schmock (2003).

6.4 Dimension-Reduction Techniques

The techniques of dimension reduction, such as factor modelling and principal components, are central to multivariate statistical analysis and are widely used in econometric model building. In the high-dimensional world of financial risk management they are essential tools.

6.4.1 Factor Models

By using a factor model we attempt to explain the randomness in the components of a d-dimensional vector $\boldsymbol{X}$ in terms of a smaller set of *common factors*. If the components of $\boldsymbol{X}$ represent, for example, equity returns, it is clear that a large part of their variation can be explained in terms of the variation of a smaller set of market index returns. Formally, we define a factor model as follows.

Definition 6.31 (linear factor model). The random vector $\boldsymbol{X}$ is said to follow a p-factor model if it can be decomposed as

$$\boldsymbol{X} = \boldsymbol{a} + B\boldsymbol{F} + \boldsymbol{\varepsilon}, \tag{6.50}$$

where

(i) $\boldsymbol{F} = (F_1, \dots, F_p)'$ is a random vector of *common factors* with $p < d$ and a covariance matrix that is positive definite,

(ii) $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_d)'$ is a random vector of *idiosyncratic error terms*, which are uncorrelated and have mean 0,

(iii) $B \in \mathbb{R}^{d \times p}$ is a matrix of constant *factor loadings* and $\boldsymbol{a} \in \mathbb{R}^d$ is a vector of constants, and

(iv) $\operatorname{cov}(\boldsymbol{F}, \boldsymbol{\varepsilon}) = E((\boldsymbol{F} - E(\boldsymbol{F}))\boldsymbol{\varepsilon}') = 0$.

The assumptions that the errors are uncorrelated with each other (ii) and also with the common factors (iv) are important parts of this definition. We do not in general require independence, only uncorrelatedness. However, if the vector $\boldsymbol{X}$ is multivariate normally distributed and follows the factor model in (6.50), then it is possible to find a version of the factor model where $\boldsymbol{F}$ and $\boldsymbol{\varepsilon}$ are Gaussian and the errors can be assumed to be mutually independent and independent of the common factors. We elaborate on this assertion in Example 6.32 below.

It follows from the basic assumptions that factor models imply a special structure for the covariance matrix Σ of $\boldsymbol{X}$. If we denote the covariance matrix of $\boldsymbol{F}$ by Ω and that of $\boldsymbol{\varepsilon}$ by the diagonal matrix Υ, it follows that

$$\Sigma = \operatorname{cov}(\boldsymbol{X}) = B\Omega B' + \Upsilon. \tag{6.51}$$

If the factor model holds, the common factors can always be transformed so that they have mean 0 and are orthogonal. By setting $\boldsymbol{F}^* = \Omega^{-1/2}(\boldsymbol{F} - E(\boldsymbol{F}))$ and $B^* = B\Omega^{1/2}$, we have a representation of the factor model of the form $\boldsymbol{X} = \boldsymbol{\mu} + B^*\boldsymbol{F}^* + \boldsymbol{\varepsilon}$, where $\boldsymbol{\mu} = E(\boldsymbol{X})$, as usual, and $\Sigma = B^*(B^*)' + \Upsilon$.

Conversely, it can be shown that whenever a random vector $\boldsymbol{X}$ has a covariance matrix that satisfies

$$\Sigma = BB' + \Upsilon \tag{6.52}$$

for some $B \in \mathbb{R}^{d \times p}$ with $\operatorname{rank}(B) = p < d$ and diagonal matrix Υ, then $\boldsymbol{X}$ has a factor-model representation for some p-dimensional factor vector $\boldsymbol{F}$ and d-dimensional error vector $\boldsymbol{\varepsilon}$.

Example 6.32 (equicorrelation model). Suppose $\boldsymbol{X}$ is a random vector with standardized margins (zero mean and unit variance) and an *equicorrelation matrix*; in other words, the correlation between each pair of components is equal to $\rho > 0$. This means that the covariance matrix Σ can be written as $\Sigma = \rho J_d + (1 - \rho) I_d$, where J_d is the d-dimensional square matrix of ones and I_d is the identity matrix, so that Σ is obviously of the form (6.52) for the d-vector $B = \sqrt{\rho}\mathbf{1}$.

To find a factor decomposition of $\boldsymbol{X}$, take *any* zero-mean, unit-variance rv Y that is *independent* of $\boldsymbol{X}$ and define a single common factor F and errors $\boldsymbol{\varepsilon}$ by

$$F = \frac{\sqrt{\rho}}{1 + \rho(d-1)} \sum_{j=1}^{d} X_j + \sqrt{\frac{1-\rho}{1+\rho(d-1)}}\, Y, \qquad \varepsilon_j = X_j - \sqrt{\rho} F,$$

where we note that in this construction F also has mean 0 and variance 1. We therefore have the factor decomposition $\boldsymbol{X} = BF + \boldsymbol{\varepsilon}$, and it may be verified by calculation that $\operatorname{cov}(F, \varepsilon_j) = 0$ for all j and $\operatorname{cov}(\varepsilon_j, \varepsilon_k) = 0$ when $j \neq k$, so that the requirements of Definition 6.31 are satisfied. A random vector with an equicorrelation matrix can be thought of as following a factor model with a single common factor.

Since we can take any Y, the factors and errors in this decomposition are non-unique. Consider the case where the vector $\boldsymbol{X}$ is Gaussian; it is most convenient to take Y to also be Gaussian, since in that case the common factor is normally distributed, the error vector is multivariate normally distributed, Y is independent of ε_j, for all j, and ε_j and ε_k are independent for $j \neq k$. Since $\operatorname{var}(\varepsilon_j) = 1 - \rho$, it is most convenient to write the factor model implied by the equicorrelation model as

$$X_j = \sqrt{\rho} F + \sqrt{1-\rho} Z_j, \quad j = 1, \dots, d, \tag{6.53}$$

where $F, Z_1, \dots, Z_d$ are mutually independent standard Gaussian rvs. This model will be used in Section 11.1.5 in the context of modelling homogeneous credit portfolios. For the more general construction on which this example is based, see Mardia, Kent and Bibby (1979, Exercise 9.2.2).

6.4.2 Statistical Estimation Strategies

Now assume that we have data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n \in \mathbb{R}^d$ representing risk-factor changes at times $t = 1, \dots, n$. Each vector observation $\boldsymbol{X}_t$ is assumed to be a realization from a factor model of the form (6.50) so that we have

$$\boldsymbol{X}_t = \boldsymbol{a} + B\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \dots, n, \tag{6.54}$$

for common-factor vectors $\boldsymbol{F}_t = (F_{t,1}, \dots, F_{t,p})'$, error vectors $\boldsymbol{\varepsilon}_t$, a vector of constants $\boldsymbol{a} \in \mathbb{R}^d$, and loading matrix $B \in \mathbb{R}^{d \times p}$. There are occasionally situations where we might wish to model $\boldsymbol{a}$ and B as time dependent, but mostly they are assumed to be fixed over time.

The model (6.54) is clearly an idealization. Data will seldom be perfectly explained by a factor model; the aim is to find an approximating factor model that captures the main sources of variability in the data. Three general types of factor model are commonly used in financial risk applications; they are known as *macroeconomic*, *fundamental* and *statistical* factor models.

Macroeconomic factor models. In these models we assume that appropriate factors $\boldsymbol{F}_t$ are also observable and we collect time-series data $\boldsymbol{F}_1, \dots, \boldsymbol{F}_n \in \mathbb{R}^p$. The name comes from the fact that, in many applications of these models in economics and finance, the observed factors are macroeconomic variables, such as changes in GDP, inflation and interest rates.

A simple example of a macroeconomic model in finance is Sharpe's single-index model, where $F_1, \dots, F_n$ are observations of the return on a market index and $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are individual equity returns that are explained in terms of the market return. Fitting of the model (estimation of B and $\boldsymbol{a}$) is accomplished by time-series regression techniques; it is described in Section 6.4.3.

Fundamental factor models. In contrast to the macroeconomic factor models, here we assume that the loading matrix B is known but that the underlying factors $\boldsymbol{F}_t$ are unobserved. Factor values $\boldsymbol{F}_1, \dots, \boldsymbol{F}_n$ have to be estimated from the data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ using cross-sectional regression at each time point.

The name comes from applications in modelling equity returns where the stocks are classified according to their "fundamentals", such as country, industry sector and size (small cap, large cap, etc.). These are generally categorical variables and it is assumed that there are underlying, unobserved factors associated with each level of the categorical variable, e.g. a factor for each country or each industry sector.

If each risk-factor change $X_{t,i}$ can be identified with a unique set of values for the fundamentals, e.g. a unique country or industry, then the matrix B is a matrix consisting of zeros and ones. If $X_{t,i}$ is attributed to different values of the fundamental variable, then B might contain factor weights summing to 1; for example, 60% of a stock return for a multinational company might be attributed to an unobserved US factor and 40% to an unobserved UK factor. There may also be situations in fundamental factor modelling where time-dependent loading matrices B_t are used.

Statistical factor models. In these models we observe neither the factors $\boldsymbol{F}_t$ nor the loadings B. Instead, we use statistical techniques to estimate both from the data $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$. This can be a very powerful approach to explaining the variability in data, but we note that the factors we obtain, while being explanatory in a statistical sense, may not have any obvious interpretation.

There are two general methods for finding factors. The first method, which is quite common in finance, is to use *principal component analysis* to construct factors; we discuss this technique in detail in Section 6.4.5. The second method, *classical statistical factor analysis*, is less commonly used in finance (see Notes and Comments).

Factor models and systematic risk. In the context of risk management, the goal of all approaches to factor modelling is either to identify or to estimate appropriate factor data $\boldsymbol{F}_1, \dots, \boldsymbol{F}_n$. If this is achieved, we can then concentrate on modelling the distribution or dynamics of the factors, which is a lower-dimensional problem than modelling $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$.

The factors describe the *systematic risk* and are of primary importance. The unobserved errors $\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n$ describe the *idiosyncratic risk* and are of secondary importance. In situations where we have many risk factors, the risk embodied in the errors is partly mitigated by a diversification effect, whereas the risk embodied in the common factors remains. The following simple example gives an idea why this is the case.

Example 6.33. We continue our analysis of the one-factor model in Example 6.32. Suppose that the random vector $\boldsymbol{X}$ in that example represents the return on d different companies so that the rv $Z_{(d)} = (1/d)\sum_{j=1}^{d} X_j$ can be thought of as the portfolio return for an equal investment in each of the companies. We calculate that

$$Z_{(d)} = \frac{1}{d}\mathbf{1}'BF + \frac{1}{d}\mathbf{1}'\boldsymbol{\varepsilon} = \sqrt{\rho}F + \frac{1}{d}\sum_{j=1}^{d}\varepsilon_j.$$

The risk in the first term is not affected by increasing the size of the portfolio d, whereas the risk in the second term can be reduced. Suppose we measure risk by simply calculating variances; we get

$$\operatorname{var}(Z_{(d)}) = \rho + \frac{1-\rho}{d} \to \rho, \quad d \to \infty,$$

showing that the systematic factor is the main contributor to the risk in a large-portfolio situation.

6.4.3 Estimating Macroeconomic Factor Models

Two equivalent approaches may be used to estimate the model parameters in a macroeconomic factor model of the form (6.54). In the first approach we perform d univariate regression analyses, one for each component of the individual return series. In the second approach we estimate all parameters in a single multivariate regression.

Univariate regression. Writing $X_{t,j}$ for the observation at time t of instrument j, we consider the univariate regression model

$$X_{t,j} = a_j + \boldsymbol{b}_j' \boldsymbol{F}_t + \varepsilon_{t,j}, \quad t = 1, \dots, n.$$

This is known as a time-series regression, since the responses $X_{1,j}, \dots, X_{n,j}$ form a univariate time series and the factors $\boldsymbol{F}_1, \dots, \boldsymbol{F}_n$ form a possibly multivariate time series. Without going into technical details we simply remark that the parameters a_j and $\boldsymbol{b}_j$ are estimated using the standard ordinary least-squares (OLS) method found in all textbooks on linear regression. To justify the use of the method and to derive statistical properties of the method it is usually assumed that, conditional on the factors, the errors $\varepsilon_{1,j}, \dots, \varepsilon_{n,j}$ are identically distributed and serially uncorrelated. In other words, they form a white noise process as defined in Chapter 4.

The estimate $\hat{a}_j$ obviously estimates the jth component of $\boldsymbol{a}$, while $\hat{\boldsymbol{b}}_j$ is an estimate of the jth row of the matrix B. By performing a regression for each of the univariate time series $X_{1,j}, \dots, X_{n,j}$ for $j = 1, \dots, d$, we complete the estimation of the parameters $\boldsymbol{a}$ and B.

Multivariate regression. To set the problem up as a multivariate linear-regression problem, we construct a number of large matrices:

$$X = \underbrace{\begin{pmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix}}_{n \times d}, \qquad F = \underbrace{\begin{pmatrix} 1 & \boldsymbol{F}_1' \\ \vdots & \vdots \\ 1 & \boldsymbol{F}_n' \end{pmatrix}}_{n \times (p+1)}, \qquad B_2 = \underbrace{\begin{pmatrix} \boldsymbol{a}' \\ B' \end{pmatrix}}_{(p+1) \times d}, \qquad E = \underbrace{\begin{pmatrix} \boldsymbol{\varepsilon}_1' \\ \vdots \\ \boldsymbol{\varepsilon}_n' \end{pmatrix}}_{n \times d}.$$

Each row of the data X corresponds to a vector observation at a fixed time point t, and each column corresponds to a univariate time series for one of the individual returns. The model (6.54) can then be expressed by the matrix equation

$$X = F B_2 + E, \tag{6.55}$$

where B_2 is the matrix of regression parameters to be estimated.

If we assume that the unobserved error vectors $\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n$ comprising the rows of E are identically distributed and serially uncorrelated, conditional on $\boldsymbol{F}_1, \dots, \boldsymbol{F}_n$, then the equation (6.55) defines a standard multivariate linear regression (see, for example, Mardia, Kent and Bibby (1979) for the standard assumptions). An estimate of B_2 is obtained by multivariate OLS according to the formula

$$\hat{B}_2 = (F'F)^{-1}F'X. \tag{6.56}$$

The factor model is now essentially calibrated, since we have estimates for $\boldsymbol{a}$ and B. The model can now be critically examined with respect to the original conditions of Definition 6.31. Do the error vectors $\boldsymbol{\varepsilon}_t$ come from a distribution with diagonal covariance matrix, and are they uncorrelated with the factors?

To learn something about the errors we can form the model residual matrix $\hat{E} = X - F\hat{B}_2$. Each row of this matrix contains an inferred value of an error vector $\hat{\boldsymbol{\varepsilon}}_t$ at a fixed point in time. Examination of the sample correlation matrix of these inferred error vectors will hopefully show that there is little remaining correlation in the errors (or at least much less than in the original data vectors $\boldsymbol{X}_t$). If this is the case, then the diagonal elements of the sample covariance matrix of the $\hat{\boldsymbol{\varepsilon}}_t$ could be taken as an estimator $\hat{\Upsilon}$ for Υ. It is sometimes of interest to form the covariance matrix implied by the factor model and compare this with the original sample covariance matrix S of the data. The implied covariance matrix is

$$\hat{\Sigma}^{(\mathrm{F})} = \hat{B}\hat{\Omega}\hat{B}' + \hat{\Upsilon}, \quad \text{where } \hat{\Omega} = \frac{1}{n-1}\sum_{t=1}^{n}(\boldsymbol{F}_t - \bar{\boldsymbol{F}})(\boldsymbol{F}_t - \bar{\boldsymbol{F}})'.$$

We would hope that $\hat{\Sigma}^{(\mathrm{F})}$ captures much of the structure of S and that the correlation matrix $R^{(\mathrm{F})} := \wp(\hat{\Sigma}^{(\mathrm{F})})$ captures much of the structure of the sample correlation matrix $R = \wp(S)$.

Example 6.34 (single-index model for Dow Jones 30 returns). As a simple example of the regression approach to fitting factor models we have fitted a single factor model to a set of ten Dow Jones 30 daily stock-return series from 1992 to 1998. Note that these are different returns to those analysed in previous sections of this chapter. They have been chosen to be of two types: technology-related titles such as Hewlett-Packard, Intel, Microsoft and IBM; and food- and consumer-related titles such as Philip Morris, Coca-Cola, Eastman Kodak, McDonald's, Wal-Mart and Disney. The factor chosen is the corresponding return on the Dow Jones 30 index itself.

The estimate of B implied by formula (6.56) is shown in the first line of Table 6.5. The highest values of B correspond to so-called *high-beta* stocks; since a one-factor model implies the relationship $E(X_j) = a_j + B_j E(F)$, these stocks potentially offer high expected returns relative to the market (but are often riskier titles); in this case, the four technology-related stocks have the highest beta values. In the second row, values of r^2, the so-called coefficient of determination, are given for each of the univariate regression models. This number measures the strength of the regression relationship between X_j and F and can be interpreted as the proportion of the variation of the stock return that is explained by variation in the market return;

Table 6.5. The first line gives estimates of B for a multivariate regression model fitted to ten Dow Jones 30 stocks where the observed common factor is the return on the Dow Jones 30 index itself. The second row gives r^2 values for a univariate regression model for each individual time series. The next ten lines of the table give the sample correlation matrix of the data R, while the middle ten lines give the correlation matrix implied by the factor model. The final ten lines show the estimated correlation matrix of the residuals from the regression model, with entries less than 0.1 in absolute value being omitted. See Example 6.34 for full details.

	MO	KO	EK	HWP	INTC	MSFT	IBM	MCD	WMT	DIS
$\hat{B}$	0.87	1.01	0.77	1.12	1.12	1.11	1.07	0.86	1.02	1.03
r^2	0.17	0.33	0.14	0.18	0.17	0.21	0.22	0.23	0.24	0.26
MO	1.00	0.27	0.14	0.17	0.16	0.25	0.18	0.22	0.16	0.22
KO	0.27	1.00	0.17	0.22	0.21	0.25	0.18	0.36	0.33	0.32
EK	0.14	0.17	1.00	0.17	0.17	0.18	0.15	0.14	0.17	0.16
HWP	0.17	0.22	0.17	1.00	0.42	0.38	0.36	0.20	0.22	0.23
INTC	0.16	0.21	0.17	0.42	1.00	0.53	0.36	0.19	0.22	0.21
MSFT	0.25	0.25	0.18	0.38	0.53	1.00	0.33	0.22	0.28	0.26
IBM	0.18	0.18	0.15	0.36	0.36	0.33	1.00	0.20	0.20	0.20
MCD	0.22	0.36	0.14	0.20	0.19	0.22	0.20	1.00	0.26	0.26
WMT	0.16	0.33	0.17	0.22	0.22	0.28	0.20	0.26	1.00	0.28
DIS	0.22	0.32	0.16	0.23	0.21	0.26	0.20	0.26	0.28	1.00
MO	1.00	0.24	0.16	0.18	0.17	0.19	0.20	0.20	0.20	0.21
KO	0.24	1.00	0.22	0.24	0.23	0.26	0.27	0.28	0.28	0.29
EK	0.16	0.22	1.00	0.16	0.15	0.17	0.18	0.18	0.18	0.19
HWP	0.18	0.24	0.16	1.00	0.17	0.19	0.20	0.20	0.21	0.22
INTC	0.17	0.23	0.15	0.17	1.00	0.19	0.19	0.19	0.20	0.21
MSFT	0.19	0.26	0.17	0.19	0.19	1.00	0.22	0.22	0.22	0.23
IBM	0.20	0.27	0.18	0.20	0.19	0.22	1.00	0.23	0.23	0.24
MCD	0.20	0.28	0.18	0.20	0.19	0.22	0.23	1.00	0.23	0.24
WMT	0.20	0.28	0.18	0.21	0.20	0.22	0.23	0.23	1.00	0.25
DIS	0.21	0.29	0.19	0.22	0.21	0.23	0.24	0.24	0.25	1.00
MO	1.00									
KO		1.00					−0.12	0.12		
EK			1.00							
HWP				1.00	0.30	0.24	0.20			
INTC				0.30	1.00	0.43	0.20			
MSFT				0.24	0.43	1.00	0.14			
IBM		−0.12		0.20	0.20	0.14	1.00			
MCD		0.12						1.00		
WMT										
DIS										1.00

the highest r^2 corresponds to Coca-Cola (33%), and in general it seems that about 20% of individual stock-return variation is explained by market-return variation.

The next ten lines of the table give the sample correlation matrix of the data R, while the middle ten lines give the correlation matrix implied by the factor model (corresponding to $\hat{\Sigma}^{(\mathrm{F})}$). The latter matrix picks up much, but not all, of the structure of the former matrix. The final ten lines show the estimated correlation matrix of the residuals from the regression model, but only those elements that exceed 0.1 in absolute value. The residuals are indeed much less correlated than the original data, but a few larger entries indicate imperfections in the factor-model representation of the data, particularly for the technology stocks. The index return for the broader market is clearly an important common factor, but further systematic effects that are not captured by the index appear to be present in these data.

6.4.4 Estimating Fundamental Factor Models

To estimate a fundamental factor model we consider, at each time point t, a cross-sectional regression model of the form

$$\boldsymbol{X}_t = B\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \tag{6.57}$$

where $\boldsymbol{X}_t \in \mathbb{R}^d$ are the risk-factor change data, $B \in \mathbb{R}^{d\times p}$ is a known matrix of factor loadings (which may be time dependent in some applications), $\boldsymbol{F}_t \in \mathbb{R}^p$ are the factors to be estimated, and $\boldsymbol{\varepsilon}_t$ are errors with diagonal covariance matrix Υ. There is no need for an intercept $\boldsymbol{a}$ in the estimation of a fundamental factor model, as this can be absorbed into the factor estimates.

To obtain precision in the estimation of $\boldsymbol{F}_t$, the dimension d of the risk-factor vector needs to be large with respect to the number of factors p to be estimated. Note also that the components of the error vector $\boldsymbol{\varepsilon}_t$ cannot generally be assumed to have equal variance, so (6.57) is a regression problem with so-called heteroscedastic errors.

We recall that, in typical applications in equity return modelling, the factors are frequently identified with country, industry-sector and company-size effects. The rows of the matrix B can consist of zeros and ones, if $X_{t,i}$ is associated with a single country or industry sector, or weights, if $X_{t,i}$ is attributed to more than one country or industry sector. This kind of interpretation for the factors is also quite common in the factor models used for modelling portfolio credit risk, as we discuss in Section 11.5.1.

Unbiased estimators of the factors $\boldsymbol{F}_t$ may be obtained by forming the OLS estimates

$$\hat{\boldsymbol{F}}_t^{\mathrm{OLS}} = (B'B)^{-1}B'\boldsymbol{X}_t,$$

and these are the best linear unbiased estimates in the case where the errors are homoscedastic, so that $\Upsilon = \upsilon^2 I_d$ for some scalar υ. However, in general, the OLS estimates are not efficient and it is possible to obtain linear unbiased estimates with a smaller covariance matrix using the method of generalized least squares (GLS). If Υ were a known matrix, the GLS estimates would be given by

$$\hat{\boldsymbol{F}}_t^{\mathrm{GLS}} = (B'\Upsilon^{-1}B)^{-1}B'\Upsilon^{-1}\boldsymbol{X}_t. \tag{6.58}$$

In practice, we replace Υ in (6.58) with an estimate $\hat{\Upsilon}$ obtained as follows. Under an assumption that the model (6.57) holds at every time point $t = 1, \dots, n$, we first carry out OLS estimation at each time t and form the model residual vectors

$$\hat{\boldsymbol{\varepsilon}}_t = \boldsymbol{X}_t - B\hat{\boldsymbol{F}}_t^{\mathrm{OLS}}, \quad t = 1, \dots, n.$$

We then form the sample covariance matrix of the residuals $\hat{\boldsymbol{\varepsilon}}_1, \dots, \hat{\boldsymbol{\varepsilon}}_n$. This matrix should be approximately diagonal, if the factor model assumption holds. We can set off-diagonal elements equal to zero to form an estimate of $\varUpsilon$.

We give an example of the estimation of a fundamental factor model in the context of modelling the yield curve in Section 9.1.4.

6.4.5 Principal Component Analysis

The aim of principal component analysis (PCA) is to reduce the dimensionality of highly correlated data by finding a small number of uncorrelated linear combinations that account for most of the variance of the original data. PCdimensional reductionA is not itself a model, but rather a data-rotation technique. However, it can be used as a way of constructing factors for use in factor modelling, and this is the main application we consider in this section.

The key mathematical result behind the technique is the *spectral decomposition theorem* of linear algebra, which says that any symmetric matrix $A \in \mathbb{R}^{d\times d}$ can be written as

$$A = \varGamma \varLambda \varGamma', \tag{6.59}$$

where

(i) $\varLambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ is the diagonal matrix of *eigenvalues of* A that, without loss of generality, are ordered so that $\lambda_1 \geqslant \lambda_2 \geqslant \dots \geqslant \lambda_d$, and

(ii) $\varGamma$ is an orthogonal matrix satisfying $\varGamma\varGamma' = \varGamma'\varGamma = I_d$ whose columns are standardized *eigenvectors of* A (i.e. eigenvectors with length 1).

Theoretical principal components. Obviously we can apply this decomposition to any covariance matrix $\varSigma$, and in this case the positive semidefiniteness of $\varSigma$ ensures that $\lambda_j \geqslant 0$ for all j. Suppose the random vector $\boldsymbol{X}$ has mean vector $\boldsymbol{\mu}$ and covariance matrix $\varSigma$ and we make the decomposition $\varSigma = \varGamma\varLambda\varGamma'$ as in (6.59). The principal components transform of $\boldsymbol{X}$ is then defined to be

$$\boldsymbol{Y} = \varGamma'(\boldsymbol{X} - \boldsymbol{\mu}), \tag{6.60}$$

and it can be thought of as a rotation and a recentring of $\boldsymbol{X}$. The jth component of the rotated vector $\boldsymbol{Y}$ is known as the *jth principal component* of $\boldsymbol{X}$ and is given by

$$Y_j = \boldsymbol{\gamma}_j'(\boldsymbol{X} - \boldsymbol{\mu}), \tag{6.61}$$

where $\boldsymbol{\gamma}_j$ is the eigenvector of $\varSigma$ corresponding to the jth ordered eigenvalue; this vector is also known as the jth vector of *loadings*.

Simple calculations show that

$$E(\boldsymbol{Y}) = \boldsymbol{0} \quad \text{and} \quad \mathrm{cov}(\boldsymbol{Y}) = \varGamma'\varSigma\varGamma = \varGamma'\varGamma\varLambda\varGamma'\varGamma = \varLambda,$$

so that the principal components of $\boldsymbol{Y}$ are uncorrelated and have variances $\mathrm{var}(Y_j) = \lambda_j$, $\forall j$. The components are thus ordered by variance, from largest to smallest. Moreover, the first principal component can be shown to be the standardized linear combination of $\boldsymbol{X}$ that has maximal variance among all such combinations; in other words,

$$\operatorname{var}(\boldsymbol{\gamma}_1'\boldsymbol{X}) = \max\{\operatorname{var}(\boldsymbol{a}'\boldsymbol{X})\colon \boldsymbol{a}'\boldsymbol{a} = 1\}.$$

For $j = 2, \dots, d$, the jth principal component can be shown to be the standardized linear combination of $\boldsymbol{X}$ with maximal variance among all such linear combinations that are *orthogonal* to (and hence uncorrelated with) the first $j - 1$ linear combinations. The final dth principal component has minimum variance among standardized linear combinations of $\boldsymbol{X}$.

To measure the ability of the first few principal components to explain the variance of $\boldsymbol{X}$, we observe that

$$\sum_{j=1}^{d} \operatorname{var}(Y_j) = \sum_{j=1}^{d} \lambda_j = \operatorname{trace}(\Sigma) = \sum_{j=1}^{d} \operatorname{var}(X_j).$$

If we interpret $\operatorname{trace}(\Sigma) = \sum_{j=1}^{d} \operatorname{var}(X_j)$ as a measure of the total variance of $\boldsymbol{X}$, then, for $k \leqslant d$, the ratio $\sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{d} \lambda_j$ represents the amount of this variance that is explained by the first k principal components.

Principal components as factors. We note that, by inverting the principal components transform (6.60), we obtain

$$\boldsymbol{X} = \boldsymbol{\mu} + \Gamma \boldsymbol{Y} = \boldsymbol{\mu} + \Gamma_1 \boldsymbol{Y}_1 + \Gamma_2 \boldsymbol{Y}_2,$$

where we have partitioned $\boldsymbol{Y}$ into vectors $\boldsymbol{Y}_1 \in \mathbb{R}^k$ and $\boldsymbol{Y}_2 \in \mathbb{R}^{d-k}$, such that $\boldsymbol{Y}_1$ contains the first k principal components, and we have partitioned Γ into matrices $\Gamma_1 \in \mathbb{R}^{d\times k}$ and $\Gamma_2 \in \mathbb{R}^{d\times(d-k)}$ correspondingly. Let us assume that the first k principal components explain a large part of the total variance and we decide to focus our attention on them and ignore the further principal components in $\boldsymbol{Y}_2$. If we set $\boldsymbol{\varepsilon} = \Gamma_2 \boldsymbol{Y}_2$, we obtain

$$\boldsymbol{X} = \boldsymbol{\mu} + \Gamma_1 \boldsymbol{Y}_1 + \boldsymbol{\varepsilon}, \tag{6.62}$$

which is reminiscent of the basic factor model (6.50) with the vector $\boldsymbol{Y}_1$ playing the role of the factors and the matrix Γ_1 playing the role of the factor loading matrix. Although the components of the error vector $\boldsymbol{\varepsilon}$ will tend to have small variances, the assumptions of the factor model are generally violated in (6.62) since $\boldsymbol{\varepsilon}$ need not have a diagonal covariance matrix and need not be uncorrelated with $\boldsymbol{Y}_1$. Nevertheless, principal components are often interpreted as factors and used to develop approximate factor models. We now describe the estimation process that is followed when data are available.

Sample principal components. Assume that we have a time series of multivariate data observations $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ with identical distribution, unknown mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, with the spectral decomposition $\Sigma = \Gamma \Lambda \Gamma'$ as before.

To construct sample principal components we need to estimate the unknown parameters. We estimate $\boldsymbol{\mu}$ by $\bar{\boldsymbol{X}}$, the sample mean vector, and we estimate Σ by the sample covariance matrix

$$S_x = \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'.$$

We apply the spectral decomposition (6.59) to the symmetric, positive-semidefinite matrix S_x to get

$$S_x = GLG', \tag{6.63}$$

where G is the eigenvector matrix, $L = \operatorname{diag}(l_1, \dots, l_d)$ is the diagonal matrix consisting of ordered eigenvalues, and we switch to roman letters to emphasize that these are now calculated from an empirical covariance matrix. The matrix G provides an estimate of Γ and L provides an estimate of Λ.

By analogy with (6.60) we define vectors of sample principal components

$$\boldsymbol{Y}_t = G'(\boldsymbol{X}_t - \bar{\boldsymbol{X}}), \quad t = 1, \dots, n. \tag{6.64}$$

The jth component of $\boldsymbol{Y}_t$ is known as the jth sample principal component at time t and is given by

$$Y_{t,j} = \boldsymbol{g}_j'(\boldsymbol{X}_t - \bar{\boldsymbol{X}}),$$

where $\boldsymbol{g}_j$ is the jth column of G, that is, the eigenvector of S_x corresponding to the jth largest eigenvalue.

The rotated vectors $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ have the property that their sample covariance matrix is L, as is easily verified:

$$\begin{aligned} S_y &= \frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})(\boldsymbol{Y}_t - \bar{\boldsymbol{Y}})' = \frac{1}{n}\sum_{t=1}^{n}\boldsymbol{Y}_t\boldsymbol{Y}_t' \\ &= \frac{1}{n}\sum_{t=1}^{n}G'(\boldsymbol{X}_t - \bar{\boldsymbol{X}})(\boldsymbol{X}_t - \bar{\boldsymbol{X}})'G = G'S_xG = L. \end{aligned}$$

The rotated vectors therefore show no correlation between components, and the components are ordered by their sample variances, from largest to smallest.

Remark 6.35. In a situation where the different components of the data vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ have very different sample variances (particularly if they are measured on very different scales), it is to be expected that the component (or components) with largest variance will dominate the first loading vector $\boldsymbol{g}_1$ and dominate the first principal component. In these situations the data are often transformed to have identical variances, which effectively means that principal component analysis is applied to the sample correlation matrix R_x. Note also that we could derive sample principal components from a robust estimate of the correlation matrix or a multivariate dispersion matrix.

We can now use the sample eigenvector matrix G and the sample principal components $\boldsymbol{Y}_t$ to calibrate an approximate factor model of the form (6.62). We assume that our data are realizations from the model

$$\boldsymbol{X}_t = \bar{\boldsymbol{X}} + G_1 \boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \quad t = 1, \dots, n, \tag{6.65}$$

where G_1 consists of the first k columns of G and $\boldsymbol{F}_t = (Y_{t,1}, \dots, Y_{t,k})'$, $t = 1, \dots, n$. The choice of k is based on a subjective choice of the number of sample principal components that are required to explain a substantial part of the total sample variance (see Example 6.36).

Equation (6.65) bears a resemblance to the factor model (6.54) except that, in practice, the errors $\boldsymbol{\varepsilon}_t$ do not generally have a diagonal covariance matrix and are not generally uncorrelated with $\boldsymbol{F}_t$. Nevertheless, the method is a popular approach to constructing time series of statistically explanatory factors from multivariate time series of risk-factor changes.

Example 6.36 (PCA-based factor model for Dow Jones 30 returns). We consider the data in Example 6.34 again. Principal component analysis is applied to the sample covariance matrix of the return data and the results are summarized in Figures 6.5 and 6.6. In the former we see a bar plot of the sample variances of the first eight principal components l_j; the cumulative proportion of the total variance explained by the components is given above each bar; the first two components explain almost 50% of the variation. In the latter figure the first two loading vectors $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ are summarized.

The first vector of loadings is positively weighted for all stocks and can be thought of as describing a kind of index portfolio; of course, the weights in the loading vector do not sum to 1, but they can be scaled to do so and this gives a so-called principal-component-mimicking portfolio. The second vector has positive weights for the consumer titles and negative weights for the technology titles; as a portfolio it can be thought of as prescribing a programme of short selling of technology to buy consumer titles. These first two sample principal components loading vectors are used to define factors.

In Table 6.6 the transpose of the matrix $\hat{B}$ (containing the loadings estimates in the factor model) is shown; the rows are merely the first two loading vectors from the principal component analysis. In the third row, values of r^2, the so-called coefficient of determination, are given for each of the univariate regression models, and these indicate that more of the variation in the data is explained by the two PCA-based factors than was explained by the observed factor in Example 6.34; it seems that the model is best able to explain Intel returns.

The next ten lines give the correlation matrix implied by the factor model (corresponding to $\hat{\Sigma}^{(\mathrm{F})}$). Compared with the true sample correlation matrix in Example 6.34 this seems to pick up more of the structure than did the correlation matrix implied by the observed factor model.

The final ten lines show the estimated correlation matrix of the residuals from the regression model, but only those elements that exceed 0.1 in absolute value. The residuals are again less correlated than the original data, but there are quite a number

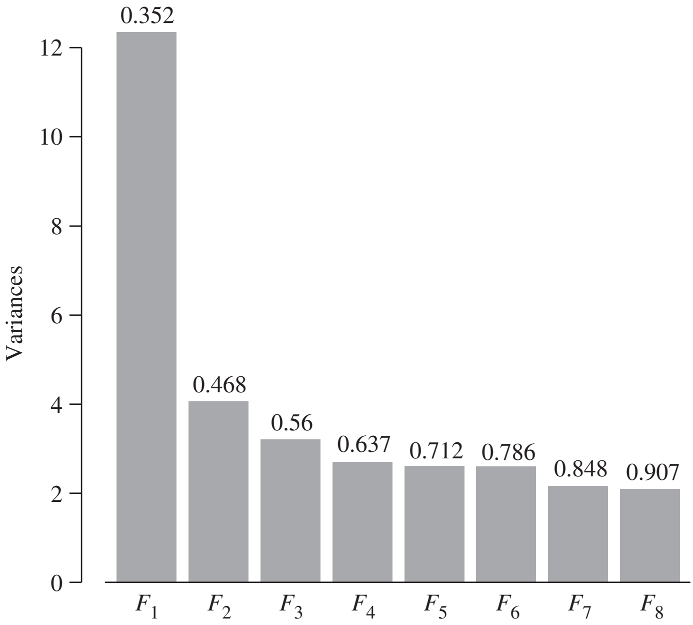

Figure 6.5. Bar plot of the sample variances l_j of the first eight principal components; the cumulative proportion of the total variance explained by the components is given above each bar ($\sum_{j=1}^{k} l_j / \sum_{j=1}^{10} l_j, k = 1, \ldots, 8$).

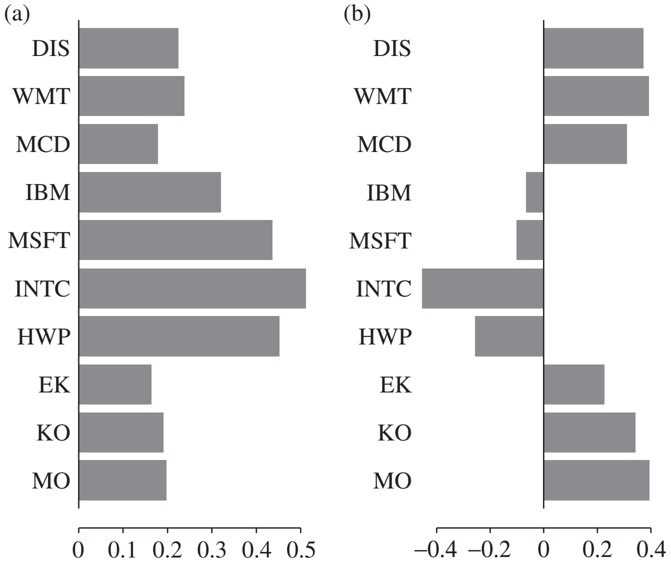

Figure 6.6. Bar plot summarizing the loading vectors $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ defining the first two principal components: (a) factor 1 loadings; (b) factor 2 loadings.

Table 6.6. The first two lines give estimates of the transpose of B for a factor model fitted to ten Dow Jones 30 stocks, where the factors are constructed from the first two sample principal components. The third row gives r^2 values for the univariate regression model for each individual time series. The next ten lines give the correlation matrix implied by the factor model. The final ten lines show the estimated correlation matrix of the residuals from the regression model, with entries less than 0.1 in absolute value omitted. See Example 6.36 for full details.

	MO	KO	EK	HWP	INTC	MSFT	IBM	MCD	WMT	DIS
$\hat{B}'$	0.20	0.19	0.16	0.45	0.51	0.44	0.32	0.18	0.24	0.22
	0.39	0.34	0.23	−0.26	−0.45	−0.10	−0.07	0.31	0.39	0.37
r^2	0.35	0.42	0.18	0.55	0.75	0.56	0.35	0.34	0.42	0.41
MO	1.00	0.39	0.25	0.17	0.13	0.25	0.20	0.35	0.38	0.38
KO	0.39	1.00	0.28	0.21	0.17	0.29	0.23	0.38	0.42	0.42
EK	0.25	0.28	1.00	0.18	0.15	0.22	0.18	0.25	0.28	0.27
HWP	0.17	0.21	0.18	1.00	0.64	0.55	0.43	0.20	0.23	0.23
INTC	0.13	0.17	0.15	0.64	1.00	0.61	0.48	0.16	0.19	0.18
MSFT	0.25	0.29	0.22	0.55	0.61	1.00	0.44	0.27	0.31	0.30
IBM	0.20	0.23	0.18	0.43	0.48	0.44	1.00	0.21	0.25	0.24
MCD	0.35	0.38	0.25	0.20	0.16	0.27	0.21	1.00	0.38	0.37
WMT	0.38	0.42	0.28	0.23	0.19	0.31	0.25	0.38	1.00	0.41
DIS	0.38	0.42	0.27	0.23	0.18	0.30	0.24	0.37	0.41	1.00
MO	1.00	−0.19	−0.15					−0.19	−0.37	−0.26
KO	−0.19	1.00	−0.15		0.11				−0.16	−0.17
EK	−0.15	−0.15	1.00					−0.15	−0.16	−0.16
HWP				1.00	−0.63	−0.37	−0.14			
INTC		0.11		−0.63	1.00	−0.24	−0.31			
MSFT				−0.37	−0.24	1.00	−0.22			
IBM				−0.14	−0.31	−0.22	1.00			
MCD	−0.19		−0.15					1.00	−0.19	−0.19
WMT	−0.37	−0.16	−0.16					−0.19	1.00	−0.23
DIS	−0.26	−0.17	−0.16					−0.19	−0.23	1.00

of larger entries, indicating imperfections in the factor-model representation of the data. In particular, we have introduced a number of larger negative correlations into the residuals; in practice, we seldom expect to find a factor model in which the residuals have a covariance matrix that appears perfectly diagonal.

Notes and Comments

For a more detailed discussion of factor models see the paper by Connor (1995), which provides a comparison of the three types of model, and the book by Campbell, Lo and MacKinlay (1997). An excellent practical introduction to these models with examples in S-Plus is Zivot and Wang (2003). Other accounts of factor models and PCA in finance are found in Alexander (2001) and Tsay (2002).

Much of our discussion of factor models, multivariate regression and principal components is based on Mardia, Kent and Bibby (1979). Statistical approaches to factor models are also treated in Seber (1984) and Johnson and Wichern (2002); these include classical statistical factor analysis, which we have omitted from our account.

7
Copulas and Dependence

In this chapter we use the concept of a copula to look more closely at the issue of modelling a random vector of dependent financial risk factors. In Section 7.1 we define copulas, give a number of examples and establish their basic properties.

Dependence concepts and dependence measures are considered in Section 7.2, beginning with the notion of perfect positive dependence or comonotonicity. This is a very important concept in risk management because it formalizes the idea of undiversifiable risks and therefore has important implications for determining risk-based capital. Dependence measures provide a scalar-valued summary of the strength of dependence between risks and there are many different measures; we consider linear correlation and two further classes of measures—rank correlations and coefficients of tail dependence—that can be directly related to copulas.

Linear correlation is a standard measure for describing the dependence between financial assets but it has a number of limitations, particularly when we leave the multivariate normal and elliptical distributions of Chapter 6 behind. Rank correlations are mainly used to calibrate copulas to data, while tail dependence is an important theoretical concept, since it addresses the phenomenon of joint extreme values in several risk factors, which is one of the major concerns in financial risk management (see also Section 3.2).

In Section 7.3 we look in more detail at the copulas of normal mixture distributions; these are the copulas that are used implicitly when normal mixture distributions are fitted to multivariate risk-factor change data, as in Chapter 6. In Section 7.4 we consider Archimedean copulas, which are widely used as dependence models in low-dimensional applications and which have also found an important niche in portfolio credit risk modelling, as will be seen in Chapters 11 and 12. The chapter ends with a section on fitting copulas to data.

7.1 Copulas

In a sense, every joint distribution function for a random vector of risk factors implicitly contains both a description of the marginal behaviour of individual risk factors and a description of their *dependence structure*; the copula approach provides a way of isolating the description of the dependence structure. We view copulas as an extremely useful concept and see several advantages in introducing and studying them.

First, copulas help in the understanding of dependence at a deeper level. They allow us to see the potential pitfalls of approaches to dependence that focus only on correlation and show us how to define a number of useful alternative dependence measures. Copulas express dependence on a *quantile scale*, which is useful for describing the dependence of extreme outcomes and is natural in a risk-management context, where VaR has led us to think of risk in terms of quantiles of loss distributions.

Moreover, copulas facilitate a *bottom-up approach to multivariate model building*. This is particularly useful in risk management, where we very often have a much better idea about the marginal behaviour of individual risk factors than we do about their dependence structure. An example is furnished by credit risk, where the individual default risk of an obligor, while in itself difficult to estimate, is at least something we can get a better handle on than the dependence among default risks for several obligors. The copula approach allows us to combine our more developed marginal models with a variety of possible dependence models and to investigate the sensitivity of risk to the dependence specification. Since the copulas we present are easily simulated, they lend themselves particularly well to Monte Carlo studies of risk. Of course, while the flexibility of the copula approach allows us, in theory, to build an unlimited number of models with given marginal distributions, we should stress that it is important to have a good understanding of the behaviour of different copulas and their appropriateness for particular kinds of modelling application.

7.1.1 Basic Properties

Definition 7.1 (copula). A d-dimensional copula is a distribution function on $[0, 1]^d$ with standard uniform marginal distributions.

We reserve the notation $C(\boldsymbol{u}) = C(u_1, \dots, u_d)$ for the multivariate dfs that are copulas. Hence C is a mapping of the form $C\colon [0, 1]^d \to [0, 1]$, i.e. a mapping of the unit hypercube into the unit interval. The following three properties must hold.

(1) $C(u_1, \dots, u_d) = 0$ if $u_i = 0$ for any i.

(2) $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i$ for all $i \in \{1, \dots, d\}$, $u_i \in [0, 1]$.

(3) For all $(a_1, \dots, a_d)$, $(b_1, \dots, b_d) \in [0, 1]^d$ with $a_i \leqslant b_i$ we have

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} C(u_{1i_1}, \dots, u_{di_d}) \geqslant 0, \tag{7.1}$$

where $u_{j1} = a_j$ and $u_{j2} = b_j$ for all $j \in \{1, \dots, d\}$.

Note that the second property corresponds to the requirement that marginal distributions are uniform. The so-called rectangle inequality in (7.1) ensures that if the random vector $(U_1, \dots, U_d)'$ has df C, then $P(a_1 \leqslant U_1 \leqslant b_1, \dots, a_d \leqslant U_d \leqslant b_d)$ is non-negative. These three properties characterize a copula; if a function C fulfills them, then it is a copula. Note also that, for $2 \leqslant k < d$, the k-dimensional margins of a d-dimensional copula are themselves copulas.

Some preliminaries. In working with copulas we must be familiar with the operations of *probability* and *quantile transformation*, as well as the properties of generalized inverses, which are summarized in Section A.1.2. The following elementary proposition is found in many probability texts.

Proposition 7.2. *Let F be a distribution function and let $F^{\leftarrow}$ denote its generalized inverse, i.e. the function $F^{\leftarrow}(u) = \inf\{x : F(x) \geqslant u\}$.*

(1) Quantile transform. *If $U \sim U(0,1)$ has a standard uniform distribution, then $P(F^{\leftarrow}(U) \leqslant x) = F(x)$.*

(2) Probability transform. *If X has df F, where F is a continuous univariate df, then $F(X) \sim U(0,1)$.*

Proof. Let $x \in \mathbb{R}$ and $u \in (0,1)$. For the first part use the fact that

$$F(x) \geqslant u \iff F^{\leftarrow}(u) \leqslant x$$

(see Proposition A.3 (iv) in Section A.1.2), from which it follows that

$$P(F^{\leftarrow}(U) \leqslant x) = P(U \leqslant F(x)) = F(x).$$

For the second part we infer that

$$\begin{aligned} P(F(X) \leqslant u) &= P(F^{\leftarrow} \circ F(X) \leqslant F^{\leftarrow}(u)) \\ &= P(X \leqslant F^{\leftarrow}(u)) = F \circ F^{\leftarrow}(u) \\ &= u, \end{aligned}$$

where the first inequality follows from the fact that $F^{\leftarrow}$ is strictly increasing (Proposition A.3 (ii)), the second follows from Proposition A.4, and the final equality is Proposition A.3 (viii). □

Proposition 7.2 (1) is the key to stochastic simulation. If we can generate a uniform variate U and compute the inverse of a df F, then we can sample from that df. Both parts of the proposition taken together imply that we can transform risks with a particular continuous df to have any other continuous distribution. For example, if X has a standard normal distribution, then $\Phi(X)$ is uniform by Proposition 7.2 (2), and, since the quantile function of a standard exponential df G is $G^{\leftarrow}(u) = -\ln(1-u)$, the transformed variable $Y := -\ln(1 - \Phi(X))$ has a standard exponential distribution by Proposition 7.2 (1).

Sklar's Theorem. The importance of copulas in the study of multivariate distribution functions is summarized by the following elegant theorem, which shows, firstly, that all multivariate dfs contain copulas and, secondly, that copulas may be used in conjunction with univariate dfs to construct multivariate dfs.

Theorem 7.3 (Sklar 1959). *Let F be a joint distribution function with margins $F_1, \dots, F_d$. Then there exists a copula $C : [0,1]^d \to [0,1]$ such that, for all $x_1, \dots, x_d$ in $\bar{\mathbb{R}} = [-\infty, \infty]$,*

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)). \tag{7.2}$$

If the margins are continuous, then C is unique; otherwise C is uniquely determined on $\operatorname{Ran} F_1 \times \operatorname{Ran} F_2 \times \cdots \times \operatorname{Ran} F_d$*, where* $\operatorname{Ran} F_i = F_i(\bar{\mathbb{R}})$ *denotes the range of F_i. Conversely, if C is a copula and $F_1, \dots, F_d$ are univariate distribution functions, then the function F defined in (7.2) is a joint distribution function with margins $F_1, \dots, F_d$.*

Proof. We prove the existence and uniqueness of a copula in the case when $F_1, \dots, F_d$ are continuous and the converse statement in its general form. Remark 7.4 explains how the general result may be proved with the more complicated distributional transform, which is given in Appendix A.1.3.

Let $\boldsymbol{X}$ be a random vector with df F and continuous margins $F_1, \dots, F_d$, and, for $i = 1, \dots, d$, set $U_i = F_i(X_i)$. By Proposition 7.2 (2), $U_i \sim U(0, 1)$, and, by Proposition A.4 in the appendix, $F_i^{\leftarrow}(U_i) = X_i$, almost surely. Let C denote the distribution function of $(U_1, \dots, U_d)$, which is a copula by Definition 7.1. For any $x_1, \dots, x_d$ in $\bar{\mathbb{R}} = [-\infty, \infty]$ we infer, using Proposition A.3 (iv), that

$$\begin{aligned} F(x_1, \dots, x_d) &= P(X_1 \leqslant x_1, \dots, X_d \leqslant x_d) \\ &= P(F_1^{\leftarrow}(U_1) \leqslant x_1, \dots, F_d^{\leftarrow}(U_d) \leqslant x_d) \\ &= P(U_1 \leqslant F_1(x_1), \dots, U_d \leqslant F(x_d)) \\ &= C(F_1(x_1), \dots, F_d(x_d)), \end{aligned} \tag{7.3}$$

and thus we obtain the identity (7.2).

If we evaluate (7.2) at the arguments $x_i = F_i^{\leftarrow}(u_i)$, $0 \leqslant u_i \leqslant 1$, $i = 1, \dots, d$, and use Proposition A.3 (viii), we obtain

$$C(u_1, \dots, u_d) = F(F_1^{\leftarrow}(u_1), \dots, F_d^{\leftarrow}(u_d)), \tag{7.4}$$

which gives an explicit representation of C in terms of F and its margins, and thus shows uniqueness.

For the converse statement assume that C is a copula and that $F_1, \dots, F_d$ are arbitrary univariate dfs. We construct a random vector with df (7.2) by taking $\boldsymbol{U}$ to be *any* random vector with df C and setting $\boldsymbol{X} := (F_1^{\leftarrow}(U_1), \dots, F_d^{\leftarrow}(U_d))$. We can then follow exactly the same sequence of equations commencing with (7.3) to establish that the df of $\boldsymbol{X}$ satisfies (7.2). □

Remark 7.4. The general form of Sklar's Theorem can be proved by using the distributional transform in Appendix A.1.3 instead of the probability transform. For a random vector $\boldsymbol{X}$ with arbitrary df F and margins $F_1, \dots, F_d$ we can set $U_i = \tilde{F}_i(X_i, V_i)$, where $\tilde{F}_i$ is the modified distribution function of X_i defined in (A.6) and $V_1, \dots, V_d$ are uniform rvs that are independent of $X_1, \dots, X_d$. Proposition A.6 shows that $U_i \sim U(0, 1)$ and $F_i^{\leftarrow}(U_i) = X_i$, almost surely, so an otherwise-identical proof may be used. The non-uniqueness of the copula is related to the fact that there are different ways of choosing the V_i variables; they need not themselves be independent and could in fact be identical for all i.

Formulas (7.2) and (7.4) are fundamental in dealing with copulas. The former shows how joint distributions F are formed by *coupling together* marginal distributions with copulas C; the latter shows how copulas are *extracted* from multivariate dfs with continuous margins. Moreover, (7.4) shows how copulas express dependence on a quantile scale, since the value $C(u_1, \dots, u_d)$ is the joint probability that X_1 lies below its u_1-quantile, X_2 lies below its u_2-quantile, and so on. Sklar's Theorem also suggests that, in the case of continuous margins, it is natural to define the notion of the copula of a distribution.

Definition 7.5 (copula of $\boldsymbol{F}$). If the random vector $\boldsymbol{X}$ has joint df F with continuous marginal distributions $F_1, \dots, F_d$, then the copula of F (or $\boldsymbol{X}$) is the df C of $(F_1(X_1), \dots, F_d(X_d))$.

Discrete distributions. The copula concept is slightly less natural for multivariate discrete distributions. This is because there is more than one copula that can be used to join the margins to form the joint df, as the following example shows.

Example 7.6 (copulas of bivariate Bernoulli). Let (X_1, X_2) have a bivariate Bernoulli distribution satisfying

$$P(X_1 = 0, X_2 = 0) = \tfrac{1}{8}, \quad P(X_1 = 1, X_2 = 1) = \tfrac{3}{8},$$
$$P(X_1 = 0, X_2 = 1) = \tfrac{2}{8}, \quad P(X_1 = 1, X_2 = 0) = \tfrac{2}{8}.$$

Clearly, $P(X_1 = 0) = P(X_2 = 0) = \frac{3}{8}$ and the marginal distributions F_1 and F_2 of X_1 and X_2 are the same. From Sklar's Theorem we know that

$$P(X_1 \leqslant x_1, X_2 \leqslant x_2) = C(P(X_1 \leqslant x_1), P(X_2 \leqslant x_2))$$

for all x_1, x_2 and some copula C. Since $\operatorname{Ran} F_1 = \operatorname{Ran} F_2 = \{0, \frac{3}{8}, 1\}$, clearly the only constraint on C is that $C(\frac{3}{8}, \frac{3}{8}) = \frac{1}{8}$. Any copula fulfilling this constraint is a copula of (X_1, X_2), and there are infinitely many such copulas.

Invariance. A useful property of the copula of a distribution is its invariance under *strictly increasing* transformations of the marginals. In view of Sklar's Theorem and this invariance property, we interpret the copula of a distribution as a very natural way of representing the dependence structure of that distribution, certainly in the case of continuous margins.

Proposition 7.7. *Let $(X_1, \dots, X_d)$ be a random vector with continuous margins and copula C and let $T_1, \dots, T_d$ be strictly increasing functions. Then $(T_1(X_1), \dots, T_d(X_d))$ also has copula C.*

Proof. We first note that $(T_1(X_1), \dots, T_d(X_d))$ is also a random vector with continuous margins and that it will have the same distribution regardless of whether each T_i is left continuous or right continuous at its (countably many) points of discontinuity. By starting with the expression (7.4) for the unique copula C of $(X_1, \dots, X_d)$, using the fact that $\{X_i \leqslant x\} = \{T_i(X_i) \leqslant T_i(x)\}$ for a strictly increasing transformation

T_i and then applying Proposition A.5 for left-continuous transformations, we obtain

$$\begin{aligned} C(u_1, \dots, u_d) &= P(X_1 \leqslant F_1^{\leftarrow}(u_1), \dots, X_d \leqslant F_d^{\leftarrow}(u_d)) \\ &= P(T_1(X_1) \leqslant T_1 \circ F_1^{\leftarrow}(u_1), \dots, T_d(X_d) \leqslant T_d \circ F_d^{\leftarrow}(u_d)) \\ &= P(T_1(X_1) \leqslant F_{T_1(X_1)}^{\leftarrow}(u_1), \dots, T_d(X_d) \leqslant F_{T_d(X_d)}^{\leftarrow}(u_d)), \end{aligned}$$

which shows that C is also the unique copula of $(T_1(X_1), \dots, T_d(X_d))$. □

Fréchet bounds. We close this section by establishing the important *Fréchet bounds* for copulas, which turn out to have important dependence interpretations that are discussed further in Sections 7.1.2 and 7.2.1.

Theorem 7.8. *For every copula $C(u_1, \dots, u_d)$ we have the bounds*

$$\max\left(\sum_{i=1}^{d} u_i + 1 - d, 0\right) \leqslant C(\boldsymbol{u}) \leqslant \min(u_1, \dots, u_d). \tag{7.5}$$

Proof. The second inequality follows from the fact that, for all i,

$$\bigcap_{1 \leqslant j \leqslant d} \{U_j \leqslant u_j\} \subset \{U_i \leqslant u_i\}.$$

For the first inequality observe that

$$\begin{aligned} C(\boldsymbol{u}) = P\left(\bigcap_{1 \leqslant i \leqslant d} \{U_i \leqslant u_i\}\right) &= 1 - P\left(\bigcup_{1 \leqslant i \leqslant d} \{U_i > u_i\}\right) \\ &\geqslant 1 - \sum_{i=1}^{d} P(U_i > u_i) = 1 - d + \sum_{i=1}^{d} u_i. \end{aligned}$$

□

The lower and upper bounds will be given the notation $W(u_1, \dots, u_d)$ and $M(u_1, \dots, u_d)$, respectively.

Remark 7.9. Although we give Fréchet bounds for a copula, Fréchet bounds may be given for any multivariate df. For a multivariate df F with margins $F_1, \dots, F_d$ we establish by similar reasoning that

$$\max\left(\sum_{i=1}^{d} F_i(x_i) + 1 - d, 0\right) \leqslant F(\boldsymbol{x}) \leqslant \min(F_1(x_1), \dots, F_d(x_d)), \tag{7.6}$$

so we have bounds for F in terms of its own marginal distributions.

7.1.2 Examples of Copulas

We provide a number of examples of copulas in this section and these are subdivided into three categories: *fundamental* copulas represent a number of important special dependence structures; *implicit* copulas are extracted from well-known multivariate distributions using Sklar's Theorem, but do not necessarily possess simple closed-form expressions; *explicit* copulas have simple closed-form expressions and follow mathematical constructions known to yield copulas. Note, however, that implicit and explicit are not mutually exclusive categories, and some copulas may have both implicit and explicit representations, as shown later in Example 7.14.

Fundamental copulas. The *independence copula* is

$$\Pi(u_1, \dots, u_d) = \prod_{i=1}^{d} u_i. \tag{7.7}$$

It is clear from Sklar's Theorem, and equation (7.2) in particular, that rvs with continuous distributions are independent if and only if their dependence structure is given by (7.7).

The *comonotonicity copula* is the Fréchet upper bound copula from (7.5):

$$M(u_1, \dots, u_d) = \min(u_1, \dots, u_d). \tag{7.8}$$

Observe that this special copula is the joint df of the random vector $(U, \dots, U)$, where $U \sim U(0, 1)$. Suppose that the rvs $X_1, \dots, X_d$ have continuous dfs and are *perfectly positively dependent* in the sense that they are almost surely strictly increasing functions of each other, so that $X_i = T_i(X_1)$ almost surely for $i = 2, \dots, d$. By Proposition 7.7, the copula of $(X_1, \dots, X_d)$ and the copula of $(X_1, \dots, X_1)$ are the same. But the copula of $(X_1, \dots, X_1)$ is just the df of $(U, \dots, U)$, where $U = F_1(X_1)$, i.e. the copula (7.8).

The *countermonotonicity copula* is the two-dimensional Fréchet lower bound copula from (7.5) given by

$$W(u_1, u_2) = \max(u_1 + u_2 - 1, 0). \tag{7.9}$$

This copula is the joint df of the random vector $(U, 1 - U)$, where $U \sim U(0, 1)$. If X_1 and X_2 have continuous dfs and are *perfectly negatively dependent* in the sense that X_2 is almost surely a strictly decreasing function of X_1, then (7.9) is their copula.

We discuss both perfect positive and perfect negative dependence in more detail in Section 7.2.1, where we see that an extension of the countermonotonicity concept to dimensions higher than two is not possible.

Perspective pictures and contour plots for the three fundamental copulas are given in Figure 7.1. The Fréchet bounds (7.5) imply that all bivariate copulas lie between the surfaces in (a) and (c).

Implicit copulas. If $\boldsymbol{Y} \sim N_d(\boldsymbol{\mu}, \Sigma)$ is a multivariate normal random vector, then its copula is a so-called *Gauss copula* (or Gaussian copula). Since the operation of standardizing the margins amounts to applying a series of strictly increasing transformations, Proposition 7.7 implies that the copula of $\boldsymbol{Y}$ is exactly the same as the copula of $\boldsymbol{X} \sim N_d(\boldsymbol{0}, P)$, where $P = \wp(\Sigma)$ is the correlation matrix of $\boldsymbol{Y}$. By Definition 7.5 this copula is given by

$$\begin{aligned} C_P^{\mathrm{Ga}}(\boldsymbol{u}) &= P(\Phi(X_1) \leqslant u_1, \dots, \Phi(X_d) \leqslant u_d) \\ &= \boldsymbol{\Phi}_P(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)), \end{aligned} \tag{7.10}$$

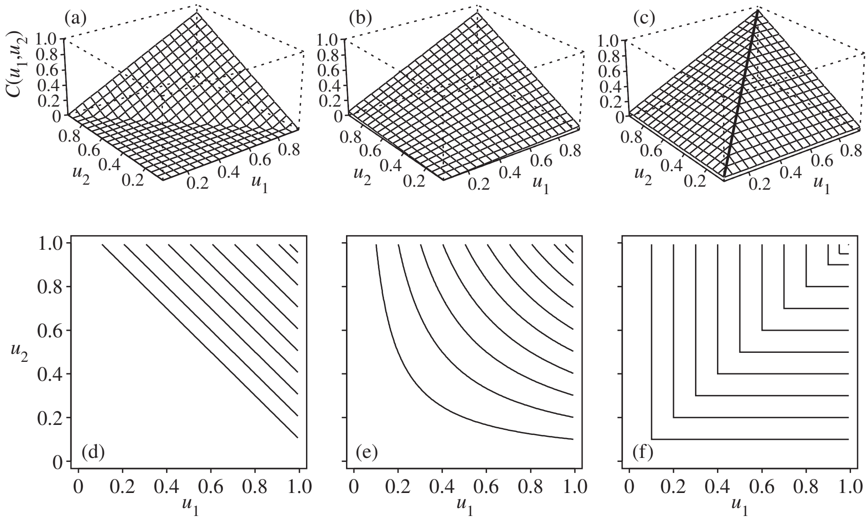

Figure 7.1. (a)–(c) Perspective plots and (d)–(f) contour plots of the three fundamental copulas: (a), (d) countermonotonicity, (b), (e) independence and (c), (f) comonotonicity. Note that these are plots of distribution functions.

where Φ denotes the standard univariate normal df and $\boldsymbol{\Phi}_P$ denotes the joint df of $\boldsymbol{X}$. The notation C_P^{Ga} emphasizes that the copula is parametrized by the $\frac{1}{2}d(d-1)$ parameters of the correlation matrix; in two dimensions we write C_ρ^{Ga}, where $\rho = \rho(X_1, X_2)$.

The Gauss copula does not have a simple closed form, but can be expressed as an integral over the density of $\boldsymbol{X}$; in two dimensions for $|\rho| < 1$ we have, using (7.10), that

$$C_\rho^{\text{Ga}}(u_1, u_2) = \int_{-\infty}^{\Phi^{-1}(u_1)} \int_{-\infty}^{\Phi^{-1}(u_2)} \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp\left(\frac{-(s_1^2 - 2\rho s_1 s_2 + s_2^2)}{2(1-\rho^2)}\right) \mathrm{d}s_1\, \mathrm{d}s_2.$$

Note that both the independence and comonotonicity copulas are special cases of the Gauss copula. If $P = I_d$, we obtain the independence copula (7.7); if $P = J_d$, the $d \times d$ matrix consisting entirely of ones, then we obtain the comonotonicity copula (7.8). Also, for $d = 2$ and $\rho = -1$ the Gauss copula is equal to the countermonotonicity copula (7.9). Thus in two dimensions the Gauss copula can be thought of as a dependence structure that interpolates between perfect positive and negative dependence, where the parameter ρ represents the strength of dependence.

Perspective plots and contour lines of the bivariate Gauss copula with $\rho = 0.7$ are shown in Figure 7.2 (a),(c); these may be compared with the contour lines of the independence and perfect dependence copulas in Figure 7.1. Note that these pictures show contour lines of distribution functions and not densities; a picture of the Gauss copula density is given in Figure 7.5.

In the same way that we can extract a copula from the multivariate normal distribution, we can extract an implicit copula from any other distribution with continuous

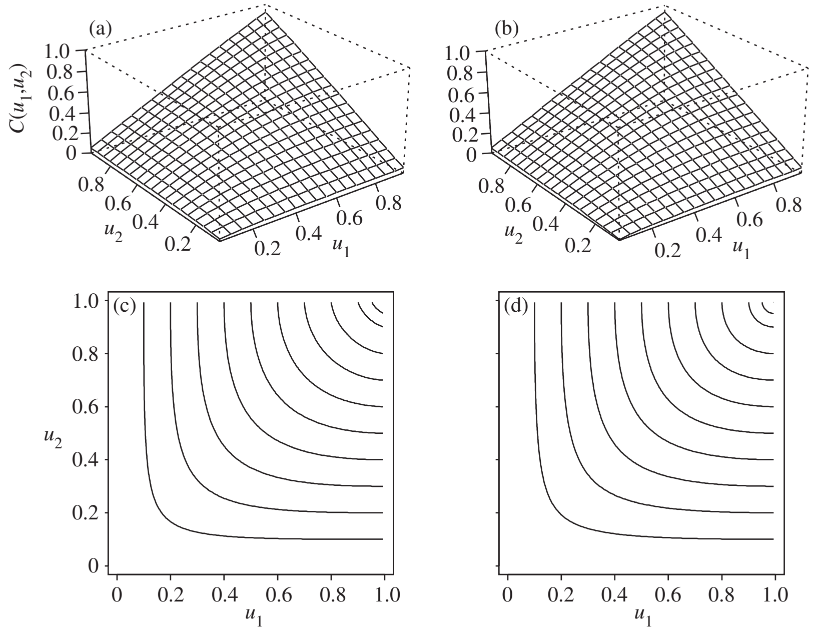

Figure 7.2. (a), (b) Perspective plots and (c), (d) contour plots of the Gaussian and Gumbel copulas, with parameters $\rho = 0.7$ and $\theta = 2$, respectively. Note that these are plots of distribution functions; a picture of the Gauss copula density is given in Figure 7.5.

marginal dfs. For example, the d-dimensional *t copula* takes the form

$$C^t_{\nu,P}(\boldsymbol{u}) = \boldsymbol{t}_{\nu,P}(t_\nu^{-1}(u_1), \dots, t_\nu^{-1}(u_d)), \qquad (7.11)$$

where t_ν is the df of a standard univariate t distribution with ν degrees of freedom, $\boldsymbol{t}_{\nu,P}$ is the joint df of the vector $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{0}, P)$, and P is a correlation matrix. As in the case of the Gauss copula, if $P = J_d$ then we obtain the comonotonicity copula (7.8). However, in contrast to the Gauss copula, if $P = I_d$ we do not obtain the independence copula (assuming $\nu < \infty$) since uncorrelated multivariate t-distributed rvs are not independent (see Lemma 6.5).

Explicit copulas. While the Gauss and t copulas are copulas implied by well-known multivariate dfs and do not themselves have simple closed forms, we can write down a number of copulas that do have simple closed forms. An example is the bivariate *Gumbel copula*:

$$C^{\mathrm{Gu}}_\theta(u_1, u_2) = \exp(-((-\ln u_1)^\theta + (-\ln u_2)^\theta)^{1/\theta}), \quad 1 \leqslant \theta < \infty. \qquad (7.12)$$

If $\theta = 1$ we obtain the independence copula as a special case, and the limit of C^{Gu}_θ as $\theta \to \infty$ is the two-dimensional comonotonicity copula. Thus the Gumbel copula interpolates between independence and perfect dependence and the parameter θ represents the strength of dependence. Perspective plot and contour lines for the Gumbel copula with parameter $\theta = 2$ are shown in Figure 7.2 (b),(d). They appear to be very similar to the picture for the Gauss copula, but Example 7.13 will show that the Gaussian and Gumbel dependence structures are quite different.

A further example is the bivariate *Clayton copula*:

$$C_\theta^{\mathrm{Cl}}(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}, \quad 0 \leqslant \theta < \infty. \tag{7.13}$$

The case $\theta = 0$ should be interpreted as the limit of (7.13) as $\theta \to 0$, which is the independence copula; as $\theta \to \infty$ we approach the two-dimensional comonotonicity copula.

The Gumbel and Clayton copulas belong to the *Archimedean* copula family and we provide more discussion of this family in Sections 7.4 and 15.2.

7.1.3 Meta Distributions

The converse statement of Sklar's Theorem provides a very powerful technique for constructing multivariate distributions with arbitrary margins and copulas; we know that if we start with a copula C and margins $F_1, \dots, F_d$, then $F(\boldsymbol{x}) := C(F_1(x_1), \dots, F_d(x_d))$ defines a multivariate df with margins $F_1, \dots, F_d$.

Consider, for example, building a distribution with the Gauss copula C_P^{Ga} but arbitrary margins; such a model is sometimes called a *meta-Gaussian* distribution. We extend the meta terminology to other distributions, so, for example, a *meta-t_ν* distribution has the copula $C_{\nu,P}^t$ and arbitrary margins, and a *meta-Clayton* distribution has the Clayton copula and arbitrary margins.

7.1.4 Simulation of Copulas and Meta Distributions

It should be apparent from the way the implicit copulas in Section 7.1.2 were extracted from well-known distributions that it is particularly easy to sample from these copulas, provided we can sample from the distribution from which they are extracted. The steps are summarized in the following algorithm.

Algorithm 7.10 (simulation of implicit copulas).

(1) Generate $\boldsymbol{X} \sim F$, where F is a df with continuous margins $F_1, \dots, F_d$.

(2) Return $\boldsymbol{U} = (F_1(X_1), \dots, F_d(X_d))'$. The random vector $\boldsymbol{U}$ has df C, where C is the unique copula of F.

Particular examples are given in the following algorithms.

Algorithm 7.11 (simulation of Gauss copula).

(1) Generate $\boldsymbol{Z} \sim N_d(\boldsymbol{0}, P)$ using Algorithm 6.2.

(2) Return $\boldsymbol{U} = (\Phi(Z_1), \dots, \Phi(Z_d))'$, where Φ is the standard normal df. The random vector $\boldsymbol{U}$ has df C_P^{Ga}.

Algorithm 7.12 (simulation of *t* copula).

(1) Generate $\boldsymbol{X} \sim t_d(\nu, \boldsymbol{0}, P)$ using Algorithm 6.10.

(2) Return $\boldsymbol{U} = (t_\nu(X_1), \dots, t_\nu(X_d))'$, where t_ν denotes the df of a standard univariate t distribution. The random vector $\boldsymbol{U}$ has df $C_{\nu,P}^t$.

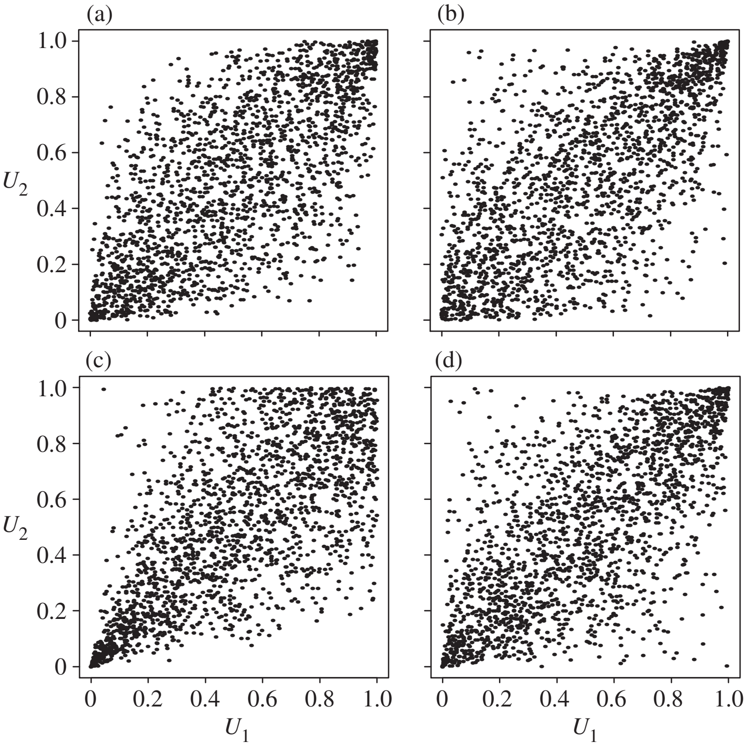

Figure 7.3. Two thousand simulated points from the (a) Gaussian, (b) Gumbel, (c) Clayton and (d) t copulas. See Example 7.13 for parameter choices and interpretation.

The Clayton and Gumbel copulas present slightly more challenging simulation problems and we give algorithms in Section 7.4 after looking at the structure of these copulas in more detail. These algorithms will, however, be used in Example 7.13 below.

Assume that the problem of generating realizations $\boldsymbol{U}$ from a particular copula has been solved. The converse of Sklar's Theorem shows us how we can sample from meta distributions that combine this copula with an arbitrary choice of marginal distribution. If $\boldsymbol{U}$ has df C, then we use quantile transformation to obtain $\boldsymbol{X} := (F_1^{\leftarrow}(U_1), \dots, F_d^{\leftarrow}(U_d))'$, which is a random vector with margins $F_1, \dots, F_d$ and multivariate df $C(F_1(x_1), \dots, F_d(x_d))$. This technique is extremely useful in Monte Carlo studies of risk and will be discussed further in the context of Example 7.58.

Example 7.13 (various copulas compared). In Figure 7.3 we show 2000 simulated points from four copulas: the Gauss copula (7.10) with parameter $\rho = 0.7$; the Gumbel copula (7.12) with parameter $\theta = 2$; the Clayton copula (7.13) with parameter $\theta = 2.2$; and the t copula (7.11) with parameters $\nu = 4$ and $\rho = 0.71$.

In Figure 7.4 we transform these points componentwise using the quantile function of the standard normal distribution to get realizations from four different meta distributions with standard normal margins. The Gaussian picture shows data generated from a standard bivariate normal distribution with correlation 70%. The

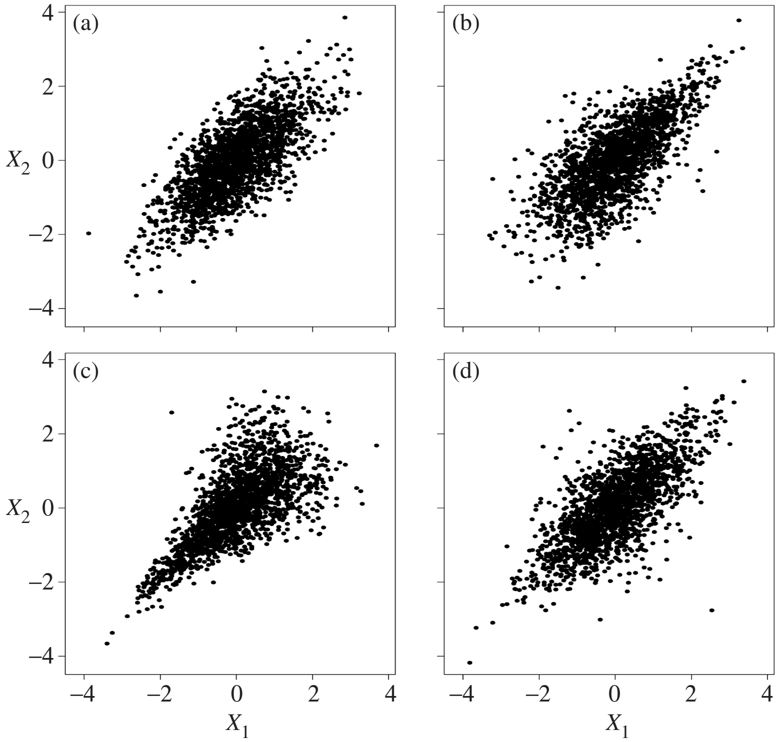

Figure 7.4. Two thousand simulated points from four distributions with standard normal margins, constructed using the copula data from Figure 7.3 ((a) Gaussian, (b) Gumbel, (c) Clayton and (d) t). The Gaussian picture shows points from a standard bivariate normal with correlation 70%; other pictures show distributions with non-Gauss copulas constructed to have a linear correlation of roughly 70%. See Example 7.13 for parameter choices and interpretation.

other pictures show data generated from unusual distributions that have been created using the converse of Sklar's Theorem; the parameters of the copulas have been chosen so that all of these distributions have a linear correlation that is roughly 70%.

Considering the Gumbel picture, these are bivariate data with a meta-Gumbel distribution with df $C_\theta^{\mathrm{Gu}}(\Phi(x_1), \Phi(x_2))$, where $\theta = 2$. The Gumbel copula causes this distribution to have *upper tail dependence*, a concept defined formally in Section 7.2.4. Roughly speaking, there is much more of a tendency for X_2 to be extreme when X_1 is extreme, and vice versa, a phenomenon that would obviously be worrying when X_1 and X_2 are interpreted as potential financial losses. The Clayton copula turns out to have *lower tail dependence*, and the t copula to have both lower and upper tail dependence; in contrast, the Gauss copula does not have tail dependence and this can also be glimpsed in Figure 7.2. In the upper-right-hand corner the contours of the Gauss copula are more like those of the independence copula of Figure 7.1 than the perfect dependence copula.

Note that the qualitative differences between the distributions are explained by the copula alone; we can construct similar pictures where the marginal distributions are exponential or Student t, or any other univariate distribution.

7.1.5 Further Properties of Copulas

Survival copulas. A version of Sklar's identity (7.2) also applies to multivariate survival functions of distributions. Let $\boldsymbol{X}$ be a random vector with multivariate survival function $\bar{F}$, marginal dfs $F_1, \dots, F_d$ and marginal survival functions $\bar{F}_1, \dots, \bar{F}_d$, i.e. $\bar{F}_i = 1 - F_i$. We have the identity

$$\bar{F}(x_1, \dots, x_d) = \hat{C}(\bar{F}_1(x_1), \dots, \bar{F}_d(x_d)) \tag{7.14}$$

for a copula $\hat{C}$, which is known as a survival copula. In the case where $F_1, \dots, F_d$ are continuous this identity is easily established by noting that

$$\begin{aligned}\bar{F}(x_1, \dots, x_d) &= P(X_1 > x_1, \dots, X_d > x_d)\\ &= P(1 - F_1(X_1) \leqslant \bar{F}_1(x_1), \dots, 1 - F_d(X_d) \leqslant \bar{F}_d(x_d)),\end{aligned}$$

so (7.14) follows by writing $\hat{C}$ for the distribution function of $\mathbf{1} - \boldsymbol{U}$, where $\boldsymbol{U} := (F_1(X_1), \dots, F_d(X_d))'$ and $\mathbf{1}$ is the vector of ones in $\mathbb{R}^d$. In general, the term *survival copula of a copula* C will be used to denote the df of $\mathbf{1} - \boldsymbol{U}$ when $\boldsymbol{U}$ has df C.

In the case where $F_1, \dots, F_d$ are continuous and strictly increasing we can give a representation for $\hat{C}$ in (7.14) by setting $x_i = \bar{F}_i^{-1}(u_i)$ for $i = 1, \dots, d$ to obtain

$$\hat{C}(u_1, \dots, u_d) = \bar{F}(\bar{F}_1^{-1}(u_1), \dots, \bar{F}_d^{-1}(u_d)). \tag{7.15}$$

The next example illustrates the derivation of a survival copula from a bivariate survival function using (7.15).

Example 7.14 (survival copula of a bivariate Pareto distribution). A well-known generalization of the important univariate Pareto distribution is the bivariate Pareto distribution with survivor function given by

$$\bar{F}(x_1, x_2) = \left(\frac{x_1 + \kappa_1}{\kappa_1} + \frac{x_2 + \kappa_2}{\kappa_2} - 1\right)^{-\alpha}, \quad x_1, x_2 \geqslant 0,\ \alpha, \kappa_1, \kappa_2 > 0.$$

It is easily confirmed that the marginal survivor functions are given by $\bar{F}_i(x) = (\kappa_i/(\kappa_i + x))^{\alpha}$, $i = 1, 2$, and we then infer using (7.15) that the survival copula is given by $\hat{C}(u_1, u_2) = (u_1^{-1/\alpha} + u_2^{-1/\alpha} - 1)^{-\alpha}$. Comparison with (7.13) reveals that this is the Clayton copula with parameter $\theta = 1/\alpha$.

The useful concept of *radial symmetry* can be expressed in terms of copulas and survival copulas.

Definition 7.15 (radial symmetry). A random vector $\boldsymbol{X}$ (or its df) is radially symmetric about a point $\boldsymbol{a}$ if $\boldsymbol{X} - \boldsymbol{a} \overset{\mathrm{d}}{=} \boldsymbol{a} - \boldsymbol{X}$.

An elliptical random vector $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ is obviously radially symmetric about $\boldsymbol{\mu}$. If $\boldsymbol{U}$ has df C, where C is a copula, then the only possible centre of symmetry is $(0.5, \dots, 0.5)$, so C is radially symmetric if

$$(U_1 - 0.5, \dots, U_d - 0.5) \overset{\mathrm{d}}{=} (0.5 - U_1, \dots, 0.5 - U_d) \iff \boldsymbol{U} \overset{\mathrm{d}}{=} \mathbf{1} - \boldsymbol{U}.$$

Thus if a copula C is radially symmetric and $\hat{C}$ is its survival copula, we have $\hat{C} = C$. It is easily seen that the copulas of elliptical distributions are radially symmetric but the Gumbel and Clayton copulas are not.

Survival copulas should not be confused with the *survival functions* of copulas, which are not themselves copulas. Since copulas are simply multivariate dfs, they have survival or tail functions, which we denote by $\bar{C}$. If $\boldsymbol{U}$ has df C and the survival copula of C is $\hat{C}$, then

$$\begin{aligned}\bar{C}(u_1, \dots, u_d) &= P(U_1 > u_1, \dots, U_d > u_d)\\ &= P(1 - U_1 \leqslant 1 - u_1, \dots, 1 - U_d \leqslant 1 - u_d)\\ &= \hat{C}(1 - u_1, \dots, 1 - u_d).\end{aligned}$$

A useful relationship between a copula and its survival copula in the bivariate case is that

$$\hat{C}(1 - u_1, 1 - u_2) = 1 - u_1 - u_2 + C(u_1, u_2). \tag{7.16}$$

Conditional distributions of copulas. It is often of interest to look at conditional distributions of copulas. We concentrate on two dimensions and suppose that (U_1, U_2) has df C. Since a copula is an increasing continuous function in each argument,

$$\begin{aligned}C_{U_2 \mid U_1}(u_2 \mid u_1) = P(U_2 \leqslant u_2 \mid U_1 = u_1) &= \lim_{\delta \to 0} \frac{C(u_1 + \delta, u_2) - C(u_1, u_2)}{\delta}\\ &= \frac{\partial}{\partial u_1} C(u_1, u_2),\end{aligned} \tag{7.17}$$

where this partial derivative exists almost everywhere (see Nelsen (2006) for precise details). The conditional distribution is a distribution on the interval [0, 1] that is only a uniform distribution in the case where C is the independence copula. A risk-management interpretation of the conditional distribution is the following. Suppose continuous risks (X_1, X_2) have the (unique) copula C. Then $1 - C_{U_2 \mid U_1}(q \mid p)$ is the probability that X_2 exceeds its qth quantile given that X_1 attains its pth quantile.

Copula densities. Copulas do not always have joint densities; the comonotonicity and countermonotonicity copulas are examples of copulas that are not absolutely continuous. However, the parametric copulas that we have met so far do have densities given by

$$c(u_1, \dots, u_d) = \frac{\partial C(u_1, \dots, u_d)}{\partial u_1 \cdots \partial u_d}, \tag{7.18}$$

and we are sometimes required to calculate them, e.g. if we wish to fit copulas to data by maximum likelihood.

It is useful to note that, for the implicit copula of an absolutely continuous joint df F with strictly increasing, continuous marginal dfs $F_1, \dots, F_d$, we may differentiate $C(u_1, \dots, u_d) = F(F_1^{\leftarrow}(u_1), \dots, F_d^{\leftarrow}(u_d))$ to see that the copula density is given by

$$c(u_1, \dots, u_d) = \frac{f(F_1^{-1}(u_1), \dots, F_d^{-1}(u_d))}{f_1(F_1^{-1}(u_1)) \cdots f_d(F_d^{-1}(u_d))}, \tag{7.19}$$

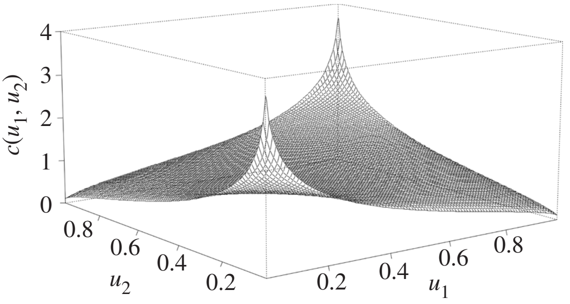

Figure 7.5. Perspective plot of the density of the bivariate Gauss copula with parameter $\rho = 0.3$.

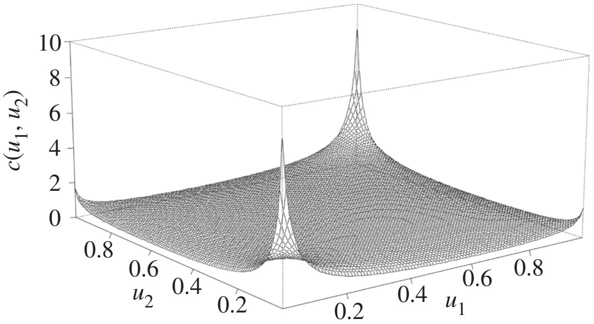

Figure 7.6. Perspective plot of the density of the bivariate t copula with parameters $\nu = 4$ and $\rho = 0.3$.

where f is the joint density of F, $f_1, \dots, f_d$ are the marginal densities, and $F_1^{-1}, \dots, F_d^{-1}$ are the ordinary inverses of the marginal dfs.

Using this technique we can calculate the densities of the Gaussian and t copulas as shown in Figures 7.5 and 7.6, respectively. Observe that the t copula assigns much more probability mass to the corners of the unit square; this may be explained by the tail dependence of the t copula, as discussed in Section 7.2.4.

Exchangeability.

Definition 7.16 (exchangeability). A random vector $\boldsymbol{X}$ is exchangeable if

$$(X_1, \dots, X_d) \overset{\mathrm{d}}{=} (X_{\Pi(1)}, \dots, X_{\Pi(d)})$$

for any permutation $(\Pi(1), \dots, \Pi(d))$ of $(1, \dots, d)$.

We will refer to a copula as an *exchangeable* copula if it is the df of an exchangeable random vector of uniform variates $\boldsymbol{U}$. Clearly, for such a copula we must have

$$C(u_1, \dots, u_d) = C(u_{\Pi(1)}, \dots, u_{\Pi(d)}) \tag{7.20}$$

for all possible permutations of the arguments of C. Such copulas will prove useful in modelling the default dependence for homogeneous groups of companies in the context of credit risk.

Examples of exchangeable copulas include both the Gumbel and Clayton copulas as well as the Gaussian and t copulas, C_P^{Ga} and $C_{\nu,P}^t$, in the case where P is an *equicorrelation matrix*, i.e. a matrix of the form $P = \rho J_d + (1-\rho)I_d$, where J_d is the square matrix consisting entirely of ones and $\rho \geqslant -1/(d-1)$.

It follows from (7.20) and (7.17) that if the df of the vector (U_1, U_2) is an exchangeable bivariate copula, then

$$P(U_2 \leqslant u_2 \mid U_1 = u_1) = P(U_1 \leqslant u_2 \mid U_2 = u_1), \tag{7.21}$$

which implies quite strong symmetry. If a random vector (X_1, X_2) has such a copula, then the probability that X_2 exceeds its u_2-quantile given that X_1 attains its u_1-quantile is exactly the same as the probability that X_1 exceeds its u_2-quantile given that X_2 attains its u_1-quantile. Not all bivariate copulas must satisfy (7.21). For an example of a non-exchangeable bivariate copula see Section 15.2.2 and Figure 15.4.

Notes and Comments

Sklar's Theorem is first found in Sklar (1959); see also Schweizer and Sklar (1983) or Nelsen (2006) for a proof of the result. The elegant proof using the distributional transform, as mentioned in Remark 7.4, is due to Rüschendorf (2009). A systematic development of the theory of copulas, particularly bivariate ones, with many examples is found in Nelsen (2006). Pitfalls related to discontinuity of marginal distributions are presented in Marshall (1996), and a primer on copulas for discrete count data is given in Genest and Nešlehová (2007). For extensive lists of parametric copula families see Hutchinson and Lai (1990), Joe (1997) and Nelsen (2006).

A reference on copula methods in finance is Cherubini, Luciano and Vecchiato (2004). Embrechts (2009) contains some references to the discussion concerning the pros and cons of copula modelling in insurance and finance.

7.2 Dependence Concepts and Measures

In this section we first provide formal definitions of the concepts of perfect positive and negative dependence (comonotonicity and countermonotonicity) and we present some of the properties of perfectly dependent random vectors.

We then focus on three kinds of dependence measure: the usual Pearson linear correlation, rank correlation, and the coefficients of tail dependence. All of these dependence measures yield a *scalar measurement* of the "strength of the dependence" for a pair of rvs (X_1, X_2), although the nature and properties of the measure are different in each case.

The rank correlations and tail-dependence coefficients are *copula-based* dependence measures. In contrast to ordinary correlation, these measures are functions of the copula only and can thus be used in the parametrization of copulas, as discussed in Section 7.5.

7.2.1 Perfect Dependence

We define the concepts of perfect positive and perfect negative dependence using the fundamental copulas of Section 7.1.2. Alternative names for these concepts are comonotonicity and countermonotonicity, respectively.

Comonotonicity. This concept may be defined in a number of equivalent ways. We give a copula-based definition and then derive alternative representations that show that comonotonic random variables can be thought of as undiversifiable random variables.

Definition 7.17 (comonotonicity). The rvs $X_1, \dots, X_d$ are said to be comonotonic if they admit as copula the Fréchet upper bound $M(u_1, \dots, u_d) = \min(u_1, \dots, u_d)$.

The following result shows that comonotonic rvs are monotonically increasing functions of a single rv. In other words, there is a single source of risk and the comonotonic variables move deterministically in lockstep with that risk.

Proposition 7.18. *$X_1, \dots, X_d$ are comonotonic if and only if*

$$(X_1, \dots, X_d) \overset{\mathrm{d}}{=} (v_1(Z), \dots, v_d(Z)) \tag{7.22}$$

for some rv Z and increasing functions $v_1, \dots, v_d$.

Proof. Assume that $X_1, \dots, X_d$ are comonotonic according to Definition 7.2.1. Let U be any uniform rv and write $F, F_1, \dots, F_d$ for the joint df and marginal dfs of $X_1, \dots, X_d$, respectively. From (7.2) we have

$$\begin{aligned} F(x_1, \dots, x_d) &= \min(F_1(x_1), \dots, F_d(x_d)) \\ &= P(U \leqslant \min(F_1(x_1), \dots, F_d(x_d))) \\ &= P(U \leqslant F_1(x_1), \dots, U \leqslant F_d(x_d)) \\ &= P(F_1^{\leftarrow}(U) \leqslant x_1, \dots, F_d^{\leftarrow}(U) \leqslant x_d) \end{aligned}$$

for any $U \sim U(0, 1)$, where we use Proposition A.3 (iv) in the last equality. It follows that

$$(X_1, \dots, X_d) \overset{\mathrm{d}}{=} (F_1^{\leftarrow}(U), \dots, F_d^{\leftarrow}(U)), \tag{7.23}$$

which is of the form (7.22). Conversely, if (7.22) holds, then

$$F(x_1, \dots, x_d) = P(v_1(Z) \leqslant x_1, \dots, v_d(Z) \leqslant x_d) = P(Z \in A_1, \dots, Z \in A_d),$$

where each A_i is an interval of the form $(-\infty, k_i]$ or $(-\infty, k_i)$, so one interval A_i is a subset of all other intervals. Therefore,

$$\begin{aligned} F(x_1, \dots, x_d) &= \min(P(Z \in A_1), \dots, P(Z \in A_d)) \\ &= \min(F_1(x_1), \dots, F_d(x_d)), \end{aligned}$$

which proves comonotonicity. □

In the case of rvs with continuous marginal distributions we have a simpler and stronger result.

Corollary 7.19. *Let $X_1, \dots, X_d$ be rvs with continuous dfs. They are comonotonic if and only if for every pair (i, j) we have $X_j = T_{ji}(X_i)$ almost surely for some increasing transformation T_{ji}.*

Proof. The result follows from the proof of Proposition 7.18 by noting that the rv U may be taken to be $F_i(X_i)$ for any i. Without loss of generality set $d = 2$ and $i = 1$ and use (7.23) and Proposition A.4 to obtain

$$(X_1, X_2) \stackrel{\mathrm{d}}{=} (F_1^{\leftarrow} \circ F_1(X_1), F_2^{\leftarrow} \circ F_1(X_1)) \stackrel{\mathrm{d}}{=} (X_1, F_2^{\leftarrow} \circ F_1(X_1)).$$

□

Comonotone additivity of quantiles. A very important result for comonotonic rvs is the additivity of the quantile function as shown in the following proposition. In addition to the VaR risk measure, the property of so-called comonotone additivity will be shown to apply to a class of risk measures known as distortion risk measures in Section 8.2.1; this class includes expected shortfall.

Proposition 7.20. *Let $0 < \alpha < 1$ and $X_1, \dots, X_d$ be comonotonic rvs with dfs $F_1, \dots, F_d$. Then*

$$F^{\leftarrow}_{X_1+\cdots+X_d}(\alpha) = F_1^{\leftarrow}(\alpha) + \cdots + F_d^{\leftarrow}(\alpha). \tag{7.24}$$

Proof. For ease of notation take $d = 2$. From Proposition 7.18 we have that $(X_1, X_2) \stackrel{\mathrm{d}}{=} (F_1^{\leftarrow}(U), F_2^{\leftarrow}(U))$ for some $U \sim U(0, 1)$. It follows that

$$F^{\leftarrow}_{X_1+X_2}(\alpha) = F^{\leftarrow}_{T(U)}(\alpha),$$

where T is the increasing left-continuous function given by $T(x) = F_1^{\leftarrow}(x) + F_2^{\leftarrow}(x)$. The result follows by applying Proposition A.5 to get

$$F^{\leftarrow}_{T(U)}(\alpha) = T(F_U^{\leftarrow}(\alpha)) = T(\alpha) = F_1^{\leftarrow}(\alpha) + F_2^{\leftarrow}(\alpha).$$

□

Countermonotonicity. In an analogous manner to the way we have defined comonotonicity, we define countermonotonicity as a copula concept, albeit restricted to the case $d = 2$.

Definition 7.21 (countermonotonicity). The rvs X_1 and X_2 are countermonotonic if they have as copula the Fréchet lower bound $W(u_1, u_2) = \max(u_1 + u_2 - 1, 0)$.

Proposition 7.22. *X_1 and X_2 are countermonotonic if and only if*

$$(X_1, X_2) \stackrel{\mathrm{d}}{=} (v_1(Z), v_2(Z))$$

for some rv Z with v_1 increasing and v_2 decreasing, or vice versa.

Proof. The proof is similar to that of Proposition 7.18 and is given in Embrechts, McNeil and Straumann (2002). □

Remark 7.23. In the case where X_1 and X_2 are continuous we have the simpler result that countermonotonicity is equivalent to $X_2 = T(X_1)$ almost surely for some decreasing function T.

The concept of countermonotonicity does not generalize to higher dimensions. The Fréchet lower bound $W(u_1, \dots, u_d)$ is not itself a copula for $d > 2$ since it is not a proper distribution function and does not satisfy (7.1), as the following example taken from Nelsen (2006, Exercise 2.36) shows.

Example 7.24 (the Fréchet lower bound is not a copula for $d > 2$). Consider the d-cube $[\frac{1}{2}, 1]^d \subset [0, 1]^d$. If the Fréchet lower bound for copulas were a df on $[0, 1]^d$, then (7.1) would imply that the probability mass $P(d)$ of this cube would be given by

$$\begin{aligned} P(d) &= \max(1 + \cdots + 1 - d + 1, 0) - d \max(\tfrac{1}{2} + 1 + \cdots + 1 - d + 1, 0) \\ &\quad + \binom{d}{2} \max(\tfrac{1}{2} + \tfrac{1}{2} + \cdots + 1 - d + 1, 0) - \cdots \\ &\quad + \max(\tfrac{1}{2} + \cdots + \tfrac{1}{2} - d + 1, 0) \\ &= 1 - \tfrac{1}{2}d. \end{aligned}$$

Hence the Fréchet lower bound cannot be a copula for $d > 2$.

Some additional insight into the impossibility of countermonotonicity for dimensions higher than two is given by the following simple example.

Example 7.25. Let X_1 be a positive-valued rv and take $X_2 = 1/X_1$ and $X_3 = \mathrm{e}^{-X_1}$. Clearly, (X_1, X_2) and (X_1, X_3) are countermonotonic random vectors. However, (X_2, X_3) is comonotonic and the copula of the vector (X_1, X_2, X_3) is the df of the vector $(U, 1 - U, 1 - U)$, which may be calculated to be

$$C(u_1, u_2, u_3) = \max(\min(u_2, u_3) + u_1 - 1, 0).$$

7.2.2 Linear Correlation

Correlation plays a central role in financial theory, but it is important to realize that the concept is only really a natural one in the context of multivariate normal or, more generally, elliptical models. As we have seen, elliptical distributions are fully described by a mean vector, a covariance matrix and a characteristic generator function. Since means and variances are features of marginal distributions, the copulas of elliptical distributions can be thought of as depending only on the correlation matrix and characteristic generator; the correlation matrix thus has a natural parametric role in these models, which it does not have in more general multivariate models. Our discussion of correlation will focus on the shortcomings of correlation and the subtle pitfalls that the naive user of correlation may encounter when moving away from elliptical models. The concept of copulas will help us to illustrate these pitfalls.

The correlation $\rho(X_1, X_2)$ between rvs X_1 and X_2 was defined in (6.3). It is a measure of *linear* dependence and takes values in $[-1, 1]$. If X_1 and X_2 are independent, then $\rho(X_1, X_2) = 0$, but it is important to be clear that the converse is false:

the uncorrelatedness of X_1 and X_2 does not in general imply their independence. Examples are provided by the class of uncorrelated normal mixture distributions (see Lemma 6.5) and the class of spherical distributions (with the single exception of the multivariate normal). For an even simpler example, we can take $X_1 = Z \sim N(0, 1)$ and $X_2 = Z^2$; these are clearly dependent rvs but have zero correlation.

If $|\rho(X_1, X_2)| = 1$, then this is equivalent to saying that X_2 and X_1 are *perfectly linearly dependent*, meaning that $X_2 = \alpha + \beta X_1$ almost surely for some $\alpha \in \mathbb{R}$ and $\beta \neq 0$, with $\beta > 0$ for positive linear dependence and $\beta < 0$ for negative linear dependence. Moreover, for $\beta_1, \beta_2 > 0$,

$$\rho(\alpha_1 + \beta_1 X_1, \alpha_2 + \beta_2 X_2) = \rho(X_1, X_2),$$

so correlation is invariant under strictly increasing *linear* transformations. However, correlation is *not* invariant under nonlinear strictly increasing transformations $T : \mathbb{R} \to \mathbb{R}$. For two real-valued rvs we have, in general, $\rho(T(X_1), T(X_2)) \neq \rho(X_1, X_2)$.

Another obvious, but important, remark is that correlation is only defined when the variances of X_1 and X_2 are finite. This restriction to finite-variance models is not ideal for a dependence measure and can cause problems when we work with heavy-tailed distributions. For example, actuaries who model losses in different business lines with infinite-variance distributions may not describe the dependence of their risks using correlation. We will encounter similar examples in Section 13.1.4 on operational risk.

Correlation fallacies. We now discuss further pitfalls in the use of correlation, which we present in the form of fallacies. We believe these fallacies are worth highlighting because they illustrate the dangers of attempting to construct multivariate risk models starting from marginal distributions and estimates of the correlations between risks. The statements we make are true if we restrict our attention to elliptically distributed risk factors, but they are false in general. For background to these fallacies, alternative examples and a discussion of the relevance to multivariate Monte Carlo simulation, see Embrechts, McNeil and Straumann (2002).

Fallacy 1. The marginal distributions and pairwise correlations of a random vector determine its joint distribution.

It should already be clear to readers of this chapter that this is not true. Figure 7.4 shows the key to constructing counterexamples. Suppose the rvs X_1 and X_2 have continuous marginal distributions F_1 and F_2 and joint df $C(F_1(x_1), F_2(x_2))$ for some copula C, and suppose their linear correlation is $\rho(X_1, X_2) = \rho$. It will generally be possible to find an alternative copula $C_2 \neq C$ and to construct a random vector (Y_1, Y_2) with df $C_2(F_1(x_1), F_2(x_2))$ such that $\rho(Y_1, Y_2) = \rho$. The following example illustrates this idea in a case where $\rho = 0$.

Example 7.26. Consider two rvs representing profits and losses on two portfolios. Suppose we are given the information that both risks have standard normal distributions and that their correlation is 0. We construct two random vectors that are consistent with this information.

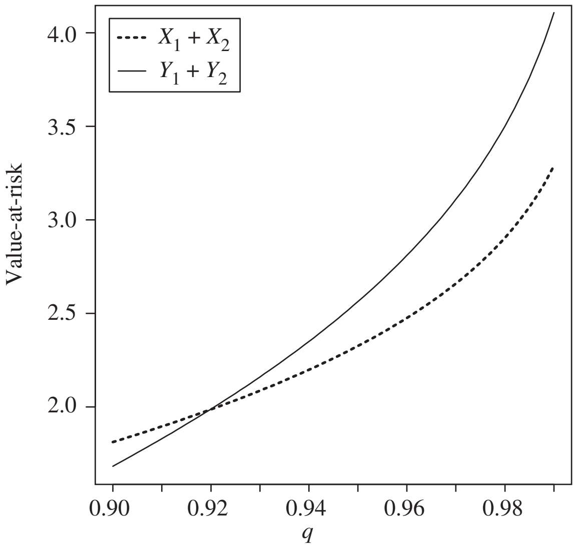

Figure 7.7. VaR for the risks X_1+X_2 and Y_1+Y_2 as described in Example 7.26. Both these pairs have standard normal margins and a correlation of zero; X_1 and X_2 are independent, whereas Y_1 and Y_2 are dependent.

Model 1 is the standard bivariate normal $\boldsymbol{X} \sim N_2(\boldsymbol{0}, I_2)$. Model 2 is constructed by taking V to be an independent discrete rv such that $P(V = 1) = P(V = -1) = 0.5$ and setting $(Y_1, Y_2) = (X_1, VX_1)$ with X_1 as in Model 1. This model obviously also has normal margins and correlation zero; its copula is given by

$$C(u_1, u_2) = 0.5 \max(u_1 + u_2 - 1, 0) + 0.5 \min(u_1, u_2),$$

which is a mixture of the two-dimensional comonotonicity and countermonotonicity copulas. This could be roughly interpreted as representing two equiprobable states of the world: in one state financial outcomes in the two portfolios are comonotonic and we are certain to make money in both or lose money in both; in the other state they are countermonotonic and we will make money in one and lose money in the other.

We can calculate analytically the distribution of the total losses $X_1 + X_2$ and $Y_1 + Y_2$; the latter sum does not itself have a univariate normal distribution. For $k \geqslant 0$ we get that

$$P(X_1 + X_2 > k) = \bar{\Phi}(k/\sqrt{2}), \qquad P(Y_1 + Y_2 > k) = \tfrac{1}{2}\bar{\Phi}(\tfrac{1}{2}k),$$

from which it follows that, for $\alpha > 0.75$,

$$F^{\leftarrow}_{X_1+X_2}(\alpha) = \sqrt{2}\Phi^{-1}(\alpha), \qquad F^{\leftarrow}_{Y_1+Y_2}(\alpha) = 2\Phi^{-1}(2\alpha - 1).$$

In Figure 7.7 we see that the quantile of $Y_1 + Y_2$ dominates that of $X_1 + X_2$ for probability levels above 93%. This example also illustrates that the VaR of a sum of risks is clearly not determined by marginal distributions and pairwise correlations.

The correlation of two risks does not only depend on their copula—if it did, then Proposition 7.7 would imply that correlation would be invariant under strictly increasing transformations, which is not the case. Correlation is also inextricably linked to the marginal distributions of the risks, and this imposes certain constraints on the values that correlation can take. This is the subject of the second fallacy.

Fallacy 2. For given univariate distributions F_1 and F_2 and any correlation value ρ in $[-1, 1]$ it is always possible to construct a joint distribution F with margins F_1 and F_2 and correlation ρ.

Again, this statement is true if F_1 and F_2 are the margins of an elliptical distribution, but is in general false. The so-called *attainable* correlations can form a strict subset of the interval $[-1, 1]$, as is shown in the next theorem. In the proof of the theorem we require the formula of Höffding, which is given in the next lemma.

Lemma 7.27. *If (X_1, X_2) has joint df F and marginal dfs F_1 and F_2, then the covariance of X_1 and X_2, when finite, is given by*

$$\operatorname{cov}(X_1, X_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (F(x_1, x_2) - F_1(x_1)F_2(x_2))\,\mathrm{d}x_1\,\mathrm{d}x_2. \tag{7.25}$$

Proof. Let (X_1, X_2) have df F and let $(\tilde{X}_1, \tilde{X}_2)$ be an *independent copy* (i.e. a second pair with df F independent of (X_1, X_2)). We have

$$2\operatorname{cov}(X_1, X_2) = E((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2)).$$

We now use a useful identity that says that, for any $a \in \mathbb{R}$ and $b \in \mathbb{R}$, we can always write $(a - b) = \int_{-\infty}^{\infty}(I_{\{b\leqslant x\}} - I_{\{a\leqslant x\}})\,\mathrm{d}x$, and we apply this to the random pairs $(X_1 - \tilde{X}_1)$ and $(X_2 - \tilde{X}_2)$. We obtain

$$\begin{aligned}
&2\operatorname{cov}(X_1, X_2)\\
&\quad= E\biggl(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(I_{\{\tilde{X}_1\leqslant x_1\}} - I_{\{X_1\leqslant x_1\}})(I_{\{\tilde{X}_2\leqslant x_2\}} - I_{\{X_2\leqslant x_2\}})\,\mathrm{d}x_1\,\mathrm{d}x_2\biggr)\\
&\quad= 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(P(X_1 \leqslant x_1, X_2 \leqslant x_2) - P(X_1 \leqslant x_1)P(X_2 \leqslant x_2))\,\mathrm{d}x_1\,\mathrm{d}x_2.
\end{aligned}$$

□

Theorem 7.28 (attainable correlations). *Let (X_1, X_2) be a random vector with finite-variance marginal dfs F_1 and F_2 and an unspecified joint df; assume also that* $\operatorname{var}(X_1) > 0$ *and* $\operatorname{var}(X_2) > 0$. *The following statements hold.*

(1) *The attainable correlations form a closed interval $[\rho_{\min}, \rho_{\max}]$ with $\rho_{\min} < 0 < \rho_{\max}$.*

(2) *The minimum correlation $\rho = \rho_{\min}$ is attained if and only if X_1 and X_2 are countermonotonic. The maximum correlation $\rho = \rho_{\max}$ is attained if and only if X_1 and X_2 are comonotonic.*

(3) *$\rho_{\min} = -1$ if and only if X_1 and $-X_2$ are of the same type (see Section A.1.1), and $\rho_{\max} = 1$ if and only if X_1 and X_2 are of the same type.*

Proof. We begin with (2) and use the identity (7.25). We also recall the two-dimensional Fréchet bounds for a general df in (7.6):

$$\max(F_1(x_1) + F_2(x_2) - 1, 0) \leqslant F(x_1, x_2) \leqslant \min(F_1(x_1), F_2(x_2)).$$

Clearly, when F_1 and F_2 are fixed, the integrand in (7.25) is maximized pointwise when X_1 and X_2 have the Fréchet upper bound copula $C(u_1, u_2) = \min(u_1, u_2)$, i.e. when they are comonotonic. Similarly, the integrand is minimized when X_1 and X_2 are countermonotonic.

To complete the proof of (1), note that clearly $\rho_{\max} \geqslant 0$. However, $\rho_{\max} = 0$ can be ruled out since this would imply that $\min(F_1(x_1), F_2(x_2)) = F_1(x_1)F_2(x_2)$ for all x_1, x_2. This can only occur if F_1 or F_2 is a degenerate distribution consisting of point mass at a single point, but this is excluded by the assumption that variances are non-zero. By a similar argument we have that $\rho_{\min} < 0$. If $W(F_1, F_2)$ and $M(F_1, F_2)$ denote the Fréchet lower and upper bounds, respectively, then the mixture $\lambda W(F_1, F_2) + (1-\lambda)M(F_1, F_2)$, $0 \leqslant \lambda \leqslant 1$, has correlation $\lambda\rho_{\min} + (1-\lambda)\rho_{\max}$. Thus for any $\rho \in [\rho_{\min}, \rho_{\max}]$ we can set $\lambda = (\rho_{\max} - \rho)/(\rho_{\max} - \rho_{\min})$ to construct a joint df that attains the correlation value ρ.

Part (3) is clear since $\rho_{\min} = -1$ or $\rho_{\max} = 1$ if and only if there is an almost sure linear relationship between X_1 and X_2. □

Example 7.29 (attainable correlations for lognormal rvs). An example where the maximal and minimal correlations can be easily calculated occurs when $\ln X_1 \sim N(0, 1)$ and $\ln X_2 \sim N(0, \sigma^2)$. For $\sigma \neq 1$ the lognormally distributed rvs X_1 and X_2 are not of the same type (although $\ln X_1$ and $\ln X_2$ are) so that, by part (3) of Theorem 7.28, we have $\rho_{\max} < 1$. The rvs X_1 and $-X_2$ are also not of the same type, so $\rho_{\min} > -1$.

To calculate the actual boundaries of the attainable interval let $Z \sim N(0, 1)$ and observe that if X_1 and X_2 are comonotonic, then $(X_1, X_2) \overset{\mathrm{d}}{=} (\mathrm{e}^Z, \mathrm{e}^{\sigma Z})$. Clearly, $\rho_{\max} = \rho(\mathrm{e}^Z, \mathrm{e}^{\sigma Z})$ and, by a similar argument, $\rho_{\min} = \rho(\mathrm{e}^Z, \mathrm{e}^{-\sigma Z})$. The analytical calculation now follows easily and yields

$$\rho_{\min} = \frac{\mathrm{e}^{-\sigma} - 1}{\sqrt{(\mathrm{e} - 1)(\mathrm{e}^{\sigma^2} - 1)}}, \qquad \rho_{\max} = \frac{\mathrm{e}^{\sigma} - 1}{\sqrt{(\mathrm{e} - 1)(\mathrm{e}^{\sigma^2} - 1)}}.$$

See Figure 7.8 for an illustration of the attainable correlation interval for different values of σ and note how the boundaries of the interval both tend rapidly to zero as σ is increased. This shows, for example, that we can have situations where comonotonic rvs have very small correlation values. Since comonotonicity is the strongest form of positive dependence, this provides a correction to the widely held view that small correlations always imply weak dependence.

Fallacy 3. For rvs $X_1 \sim F_1$ and $X_2 \sim F_2$ and for given α, the quantile of the sum $F^{\leftarrow}_{X_1+X_2}(\alpha)$ is maximized when the joint distribution F has maximal correlation.

While once again this is true if (X_1, X_2) are jointly elliptical, the statement is not true in general and any example of the superadditivity of the quantile function (or VaR risk measure) yields a counterexample.

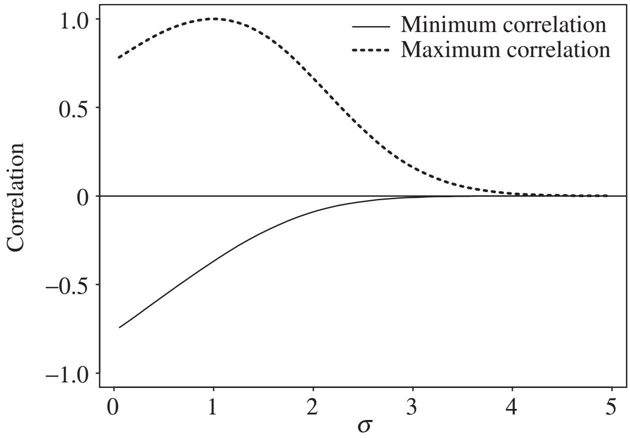

Figure 7.8. Maximum and minimum attainable correlations for lognormal rvs X_1 and X_2, where $\ln X_1$ is standard normal and $\ln X_2$ is normal with mean 0 and variance σ^2.

In a superadditive VaR example, we have, for some value of α, that

$$F^{\leftarrow}_{X_1+X_2}(\alpha) > F^{\leftarrow}_{X_1}(\alpha) + F^{\leftarrow}_{X_2}(\alpha), \tag{7.26}$$

but, by Proposition 7.20, the right-hand side of this inequality is equal to $F^{\leftarrow}_{Y_1+Y_2}(\alpha)$ for a pair of comonotonic rvs (Y_1, Y_2) with $Y_1 \overset{\mathrm{d}}{=} X_1$ and $Y_2 \overset{\mathrm{d}}{=} X_2$. Moreover, by part (2) of Theorem 7.28, (Y_1, Y_2) will attain the maximal correlation $\rho_{\max}$ and $\rho_{\max} > \rho(X_1, X_2)$. A simple example where this occurs is as follows.

Example 7.30. Let $X_1 \sim \mathrm{Exp}(1)$ and $X_2 \sim \mathrm{Exp}(1)$ be two independent standard exponential rvs. Let $Y_1 = Y_2 = X_1$ and take $\alpha = 0.7$. Since $X_1 + X_2 \sim \mathrm{Ga}(2, 1)$ (see Appendix A.2.4) it is easily checked that

$$F^{\leftarrow}_{X_1+X_2}(\alpha) > F^{\leftarrow}_{X_1}(\alpha) + F^{\leftarrow}_{X_2}(\alpha) = F^{\leftarrow}_{Y_1+Y_2}(\alpha)$$

but $\rho(X_1, X_2) = 0$ and $\rho(Y_1, Y_2) = 1$. This example is also discussed in Section 8.3.3.

In Section 8.4.4 we will look at the problem of discovering how "bad" the quantile of the sum of the two risks in (7.26) can be when the marginal distributions are known.

A common message can be extracted from the fallacies of this section: namely that the concept of correlation is meaningless unless applied in the context of a well-defined joint model. Any interpretation of correlation values in the absence of such a model should be avoided.

7.2.3 Rank Correlation

Rank correlations are simple scalar measures of dependence that depend only on the copula of a bivariate distribution and not on the marginal distributions, unlike linear correlation, which depends on both. The standard empirical estimators of rank correlation may be calculated by looking at the *ranks* of the data alone, hence the name. In other words, we only need to know the ordering of the sample for each variable of interest and not the actual numerical values.

The main practical reason for looking at rank correlations is that they can be used to calibrate copulas to empirical data. At a theoretical level, being direct functionals of the copula, rank correlations have more appealing properties than linear correlations, as is discussed below. There are two main varieties of rank correlation, Kendall's and Spearman's, and both can be understood as a measure of concordance for bivariate random vectors. Two points in $\mathbb{R}^2$, denoted by (x_1, x_2) and $(\tilde{x}_1, \tilde{x}_2)$, are said to be *concordant* if $(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2) > 0$ and to be *discordant* if $(x_1 - \tilde{x}_1)(x_2 - \tilde{x}_2) < 0$.

Kendall's tau. Consider a random vector (X_1, X_2) and an independent copy $(\tilde{X}_1, \tilde{X}_2)$ (i.e. a second vector with the same distribution, but independent of the first). If X_2 tends to increase with X_1, then we expect the probability of concordance to be high relative to the probability of discordance; if X_2 tends to decrease with increasing X_1, then we expect the opposite. This motivates Kendall's rank correlation, which is simply the probability of concordance minus the probability of discordance for these pairs:

$$\rho_\tau(X_1, X_2) = P((X_1-\tilde{X}_1)(X_2-\tilde{X}_2) > 0)-P((X_1-\tilde{X}_1)(X_2-\tilde{X}_2) < 0). \quad (7.27)$$

It is easily seen that there is a more compact way of writing this as an expectation, which also leads to an obvious estimator in Section 7.5.1.

Definition 7.31. For rvs X_1 and X_2 Kendall's tau is given by

$$\rho_\tau(X_1, X_2) = E(\operatorname{sign}((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2))),$$

where $(\tilde{X}_1, \tilde{X}_2)$ is an independent copy of (X_1, X_2) and $\operatorname{sign}(x) = I_{\{x>0\}} - I_{\{x<0\}}$.

In higher dimensions the Kendall's tau matrix of a random vector $\boldsymbol{X}$ may be written as $\rho_\tau(\boldsymbol{X}) = \operatorname{cov}(\boldsymbol{Y})$, where $\boldsymbol{Y} = \operatorname{sign}(\boldsymbol{X} - \tilde{\boldsymbol{X}})$ and $\tilde{\boldsymbol{X}}$ is an independent copy of $\boldsymbol{X}$; note that $\boldsymbol{Y}$ is obtained by the componentwise application of the sign function, so that $\boldsymbol{Y} = (Y_1, \dots, Y_d)'$, where $Y_i = \operatorname{sign}(X_i - \tilde{X}_i)$ for $i = 1, \dots, d$. Since $\rho_\tau(\boldsymbol{X})$ can be expressed as the covariance matrix of $\boldsymbol{Y}$, it is obviously positive semidefinite.

We now show that, for random variables with continuous dfs, Kendall's tau depends only on the unique copula C of (X_1, X_2) and we give an explicit formula for computing ρ_τ from C.

Proposition 7.32. *Suppose X_1 and X_2 have continuous marginal distributions and unique copula C. Then*

$$\rho_\tau(X_1, X_2) = 4 \int_0^1 \int_0^1 C(u_1, u_2)\, \mathrm{d}C(u_1, u_2) - 1. \quad (7.28)$$

Proof. Starting from (7.27) and writing F_1 and F_2 for the marginals dfs, we have

$$\begin{aligned} \rho_\tau(X_1, X_2) &= 2P((X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0) - 1 \\ &= 4P(X_1 < \tilde{X}_1, X_2 < \tilde{X}_2) - 1 \qquad (7.29) \\ &= 4P(F_1(X_1) < F_1(\tilde{X}_1), F_2(X_2) < F_2(\tilde{X}_2)) - 1, \end{aligned}$$

where the second equality follows from the interchangeability of the pairs (X_1, X_2) and $(\tilde{X}_1, \tilde{X}_2)$ and the third equality from the continuity of F_1 and F_2. Introducing the uniform random variables $U_i = F_i(X_i)$ and $\tilde{U}_i = F_i(\tilde{X}_i)$, $i = 1, 2$, and noting that the df of the pairs (U_1, U_2) and $(\tilde{U}_1, \tilde{U}_2)$ is C, we obtain

$$\begin{aligned}\rho_\tau(X_1, X_2) &= 4E(P(U_1 < \tilde{U}_1, U_2 < \tilde{U}_2 \mid \tilde{U}_1, \tilde{U}_2)) - 1 \\ &= 4\int_0^1 \int_0^1 P(U_1 < u_1, U_2 < u_2)\, \mathrm{d}C(u_1, u_2) - 1,\end{aligned}$$

from which (7.28) follows. □

Spearman's rho. This measure can also be defined in terms of the concordance and discordance of two bivariate random vectors, but this time we consider the independent pairs (X_1, X_2) and $(\tilde{X}_1, \bar{X}_2)$ and assume that they have identical marginal distributions but that the second pair is a pair of *independent* random variables.

Definition 7.33. For rvs X_1 and X_2 Spearman's rho is given by

$$\rho_S(X_1, X_2) = 3(P((X_1 - \tilde{X}_1)(X_2 - \bar{X}_2) > 0) - P((X_1 - \tilde{X}_1)(X_2 - \bar{X}_2) < 0)), \tag{7.30}$$

where $\tilde{X}_1$ and $\bar{X}_2$ are random variables satisfying $\tilde{X}_1 \overset{\mathrm{d}}{=} X_1$ and $\bar{X}_2 \overset{\mathrm{d}}{=} X_2$ and where (X_1, X_2), $\tilde{X}_1$ and $\bar{X}_2$ are all independent.

It is not immediately apparent that this definition gives a sensible correlation measure, i.e. a number in the interval $[-1, 1]$. The following proposition gives an alternative representation, which makes this clear for continuous random variables.

Proposition 7.34. *If X_1 and X_2 have continuous marginal distributions F_1 and F_2, then $\rho_S(X_1, X_2) = \rho(F_1(X_1), F_2(X_2))$.*

Proof. If the random vectors (X_1, X_2) and $(\tilde{X}_1, \bar{X}_2)$ have continuous marginal distributions, we may write

$$\begin{aligned}\rho_S(X_1, X_2) &= 6P((X_1 - \tilde{X}_1)(X_2 - \bar{X}_2) > 0) - 3 \qquad (7.31)\\ &= 6P((F_1(X_1) - F_1(\tilde{X}_1))(F_2(X_2) - F_2(\bar{X}_2)) > 0) - 3 \\ &= 6P((U_1 - \tilde{U}_1)(U_2 - \bar{U}_2) > 0) - 3,\end{aligned}$$

where $(U_1, U_2) := (F_1(X_1), F_2(X_2))$, $\tilde{U}_1 := F_1(\tilde{X}_1)$ and $\bar{U}_2 = F_2(\bar{X}_2)$ and U_1, U_2, $\tilde{U}_1$ and $\bar{U}_2$ all have standard uniform distributions. Conditioning on U_1 and U_2 we obtain

$$\begin{aligned}\rho_S(X_1, X_2) &= 6E(P((U_1 - \tilde{U}_1)(U_2 - \bar{U}_2) > 0 \mid U_1, U_2)) - 3 \\ &= 6E(P(\tilde{U}_1 < U_1, \bar{U}_2 < U_2 \mid U_1, U_2) \\ &\qquad + P(\tilde{U}_1 > U_1, \bar{U}_2 > U_2 \mid U_1, U_2)) - 3 \\ &= 6E(U_1U_2 + (1 - U_1)(1 - U_2)) - 3 \\ &= 12E(U_1U_2) - 6E(U_1) - 6E(U_2) + 3 \\ &= 12\operatorname{cov}(U_1, U_2), \qquad (7.32)\end{aligned}$$

where we have used the fact that $E(U_1) = E(U_2) = \frac{1}{2}$. The conclusion that $\rho_S(X_1, X_2) = \rho(F_1(X_1), F_2(X_2))$ follows from noting that $\operatorname{var}(U_1) = \operatorname{var}(U_2) = \frac{1}{12}$. □

In other words, Spearman's rho, for continuous random variables, is simply the linear correlation of their unique copula. The Spearman's rho matrix for the general multivariate random vector $\boldsymbol{X}$ with continuous margins is given by $\rho_S(\boldsymbol{X}) = \rho(F_1(X_1), \dots, F_d(X_d))$ and must be positive semidefinite. In the bivariate case, the formula for Spearman's rho in terms of the copula C of (X_1, X_2) follows from a simple application of Höffding's formula (7.25) to formula (7.32).

Corollary 7.35. *Suppose X_1 and X_2 have continuous marginal distributions and unique copula C. Then*

$$\rho_S(X_1, X_2) = 12 \int_0^1 \int_0^1 (C(u_1, u_2) - u_1 u_2)\, \mathrm{d}u_1\, \mathrm{d}u_2. \tag{7.33}$$

Properties of rank correlation. Kendall's tau and Spearman's rho have many properties in common. They are both symmetric dependence measures taking values in the interval $[-1, 1]$. They give the value zero for independent rvs, although a rank correlation of 0 does not necessarily imply independence. It can be shown that they take the value 1 if and only if X_1 and X_2 are comonotonic (see Embrechts, McNeil and Straumann 2002) and the value -1 if and only if they are countermonotonic (which contrasts with the behaviour of linear correlation observed in Theorem 7.28). They are invariant under strictly increasing transformations of X_1 and X_2.

To what extent do the fallacies of linear correlation identified in Section 7.2.2 carry over to rank correlation? Clearly, Fallacy 1 remains relevant: marginal distributions and pairwise rank correlations do not fully determine the joint distribution of a vector of risks. Fallacy 3 also still applies when we switch from linear to rank correlations; although two comonotonic risks will have the maximum rank correlation value of one, this does not imply that the quantile of their sum is maximized over the class of all joint models with the same marginal distributions.

However, Fallacy 2 no longer applies when we consider rank correlations: for any choice of continuous marginal distributions it is possible to specify a bivariate distribution that has any desired rank correlation value in $[-1, 1]$. One way of doing this is to take a convex combination of the form

$$F(x_1, x_2) = \lambda W(F_1(x_1), F_2(x_2)) + (1 - \lambda) M(F_1(x_1), F_2(x_2)),$$

where W and M are the countermonotonicity and comonotonicity copulas, respectively. A random pair (X_1, X_2) with this df has rank correlation

$$\rho_\tau(X_1, X_2) = \rho_S(X_1, X_2) = 1 - 2\lambda,$$

which yields any desired value in $[-1, 1]$ for an appropriate choice of λ in $[0, 1]$. But this is only one of many possible constructions; a model with the Gauss copula of the form $F(x_1, x_2) = C_\rho^{\mathrm{Ga}}(F_1(x_1), F_2(x_2))$ can also be parametrized by an appropriate choice of $\rho \in [-1, 1]$ to have any rank correlation in $[-1, 1]$. In Section 7.3.2 we will calculate Spearman's rho and Kendall's tau values for the Gauss copula and other copulas of normal variance mixture distributions.

7.2.4 Coefficients of Tail Dependence

Like the rank correlations, the coefficients of tail dependence are measures of pairwise dependence that depend only on the copula of a pair of rvs X_1 and X_2 with continuous marginal dfs. The motivation for looking at these coefficients is that they provide measures of *extremal dependence* or, in other words, measures of the strength of dependence in the tails of a bivariate distribution. The coefficients we describe are defined in terms of limiting conditional probabilities of *quantile exceedances*. We note that there are a number of other definitions of tail-dependence measures in the literature (see Notes and Comments).

In the case of upper tail dependence we look at the probability that X_2 exceeds its q-quantile, given that X_1 exceeds its q-quantile, and then consider the limit as q goes to one. Obviously the roles of X_1 and X_2 are interchangeable. Formally we have the following.

Definition 7.36. Let X_1 and X_2 be rvs with dfs F_1 and F_2. The coefficient of upper tail dependence of X_1 and X_2 is

$$\lambda_{\mathrm{u}} := \lambda_{\mathrm{u}}(X_1, X_2) = \lim_{q \to 1^-} P(X_2 > F_2^{\leftarrow}(q) \mid X_1 > F_1^{\leftarrow}(q)),$$

provided a limit $\lambda_{\mathrm{u}} \in [0, 1]$ exists. If $\lambda_{\mathrm{u}} \in (0, 1]$, then X_1 and X_2 are said to show upper tail dependence or extremal dependence in the upper tail; if $\lambda_{\mathrm{u}} = 0$, they are *asymptotically independent* in the upper tail. Analogously, the coefficient of lower tail dependence is

$$\lambda_{\mathrm{l}} := \lambda_{\mathrm{l}}(X_1, X_2) = \lim_{q \to 0^+} P(X_2 \leqslant F_2^{\leftarrow}(q) \mid X_1 \leqslant F_1^{\leftarrow}(q)),$$

provided a limit $\lambda_{\mathrm{l}} \in [0, 1]$ exists.

If F_1 and F_2 are continuous dfs, then we get simple expressions for λ_{l} and λ_{u} in terms of the unique copula C of the bivariate distribution. Using elementary conditional probability and (7.4) we have

$$\begin{aligned}\lambda_{\mathrm{l}} &= \lim_{q \to 0^+} \frac{P(X_2 \leqslant F_2^{\leftarrow}(q), X_1 \leqslant F_1^{\leftarrow}(q))}{P(X_1 \leqslant F_1^{\leftarrow}(q))} \\ &= \lim_{q \to 0^+} \frac{C(q, q)}{q}. \end{aligned} \tag{7.34}$$

For upper tail dependence we use (7.14) to obtain

$$\lambda_{\mathrm{u}} = \lim_{q \to 1^-} \frac{\hat{C}(1-q, 1-q)}{1-q} = \lim_{q \to 0^+} \frac{\hat{C}(q, q)}{q}, \tag{7.35}$$

where $\hat{C}$ is the survival copula of C (see (7.16)). For radially symmetric copulas we must have $\lambda_{\mathrm{l}} = \lambda_{\mathrm{u}}$, since $C = \hat{C}$ for such copulas.

Calculation of these coefficients is straightforward if the copula in question has a simple closed form, as is the case for the Gumbel copula in (7.12) and the Clayton copula in (7.13). In Section 7.3.1 we will use a slightly more involved method to calculate tail-dependence coefficients for copulas of normal variance mixture distributions, such as the Gaussian and t copulas.

Example 7.37 (Gumbel and Clayton copulas). Writing $\hat{C}_\theta^{\mathrm{Gu}}$ for the Gumbel survival copula we first use (7.16) to infer that

$$\lambda_{\mathrm{u}} = \lim_{q \to 1^-} \frac{\hat{C}_\theta^{\mathrm{Gu}}(1-q, 1-q)}{1-q} = 2 - \lim_{q \to 1^-} \frac{C_\theta^{\mathrm{Gu}}(q, q) - 1}{q-1}.$$

We now use L'Hôpital's rule and the fact that $C_\theta^{\mathrm{Gu}}(u, u) = u^{2^{1/\theta}}$ to infer that

$$\lambda_{\mathrm{u}} = 2 - \lim_{q \to 1^-} \frac{\mathrm{d}C_\theta^{\mathrm{Gu}}(q, q)}{\mathrm{d}q} = 2 - 2^{1/\theta}.$$

Provided that $\theta > 1$, the Gumbel copula has upper tail dependence. The strength of this tail dependence tends to 1 as $\theta \to \infty$, which is to be expected since the Gumbel copula tends to the comonotonicity copula as $\theta \to \infty$. Using a similar technique the coefficient of lower tail dependence for the Clayton copula may be shown to be $\lambda_{\mathrm{l}} = 2^{-1/\theta}$ for $\theta > 0$.

The consequences of the lower tail dependence of the Clayton copula and the upper tail dependence of the Gumbel copula can be seen in Figures 7.3 and 7.4, where there is obviously an increased tendency for these copulas to generate joint extreme values in the respective corners. In Section 7.3.1 we will see that the Gauss copula is asymptotically independent in both tails, while the *t* copula has both upper and lower tail dependence of the same magnitude (due to its radial symmetry).

Notes and Comments

The concept of comonotonicity or perfect positive dependence is discussed by many authors, including Schmeidler (1986) and Yaari (1987). See also Wang and Dhaene (1998), whose proof we use in Proposition 7.18, and the entry in the *Encyclopedia of Actuarial Science* by Vyncke (2004).

The discussion of correlation fallacies is based on Embrechts, McNeil and Straumann (2002), which contains a number of other examples illustrating these pitfalls. Throughout this book we make numerous references to this paper, which also played a major role in popularizing the copula concept mainly, but not solely, in finance, insurance and economics (see, for example, Genest, Gendron and Bourdeau-Brien 2009). The ETH-RiskLab preprint of this paper was available as early as 1998, with a published abridged version appearing as Embrechts, McNeil and Straumann (1999).

For Höffding's formula and its use in proving the bounds on attainable correlations see Höffding (1940), Fréchet (1957) and Shea (1983). Useful references for rank correlations are Kruskal (1958) and Joag-Dev (1984). The relationship between rank correlation and copulas is discussed in Schweizer and Wolff (1981) and Nelsen (2006). The definition of tail dependence that we use stems from Joe (1993, 1997). There are a number of alternative definitions of tail-dependence measures, as discussed, for example, in Coles, Heffernan and Tawn (1999).

Important books that treat dependence concepts and emphasize links to copulas include Joe (1997), Denuit et al. (2005) and Rüschendorf (2013).

7.3 Normal Mixture Copulas

A unique copula is contained in every multivariate distribution with continuous marginal distributions, and a useful class of parametric copulas are those contained in the multivariate normal mixture distributions of Section 6.2. We view these copulas as particularly important in market-risk applications; indeed, in most cases, these copulas are used implicitly, without the user necessarily recognizing the fact. Whenever normal mixture distributions are fitted to multivariate return data or used as innovation distributions in multivariate time-series models, normal mixture copulas are used. They are also found in a number of credit risk models, as we discuss in Section 12.2.

In this section we first focus on normal variance mixture copulas; in Section 7.3.1 we examine their tail-dependence properties; and in Section 7.3.2 we calculate rank correlation coefficients, which are useful for calibrating these copulas to data. Then, in Sections 7.3.3 and 7.3.4, we look at more exotic examples of copulas arising from multivariate normal mixture constructions.

7.3.1 Tail Dependence

Coefficients of tail dependence. Consider a pair of uniform rvs (U_1, U_2) whose distribution $C(u_1, u_2)$ is a normal variance mixture copula. Due to the radial symmetry of C (see Section 7.1.5), it suffices to consider the formula for the lower tail-dependence coefficient in (7.34) to calculate the coefficient of tail dependence λ of C. By applying L'Hôpital's rule and using (7.17) we obtain

$$\lambda = \lim_{q\to 0^+} \frac{\mathrm{d}C(q,q)}{\mathrm{d}q} = \lim_{q\to 0^+} P(U_2 \leqslant q \mid U_1 = q) + \lim_{q\to 0^+} P(U_1 \leqslant q \mid U_2 = q).$$

Since C is *exchangeable*, we have from (7.21) that

$$\lambda = 2 \lim_{q\to 0^+} P(U_2 \leqslant q \mid U_1 = q). \tag{7.36}$$

We now show the interesting contrast between the Gaussian and t copulas that we alluded to in Example 7.13, namely that the t copula has tail dependence whereas the Gauss copula is asymptotically independent in the tail.

Example 7.38 (asymptotic independence of the Gauss copula). To evaluate the tail-dependence coefficient for the Gauss copula C_ρ^{Ga}, let $(X_1, X_2) := (\Phi^{-1}(U_1), \Phi^{-1}(U_2))$, so that (X_1, X_2) has a bivariate normal distribution with standard margins and correlation ρ. It follows from (7.36) that

$$\begin{aligned}\lambda &= 2 \lim_{q\to 0^+} P(\Phi^{-1}(U_2) \leqslant \Phi^{-1}(q) \mid \Phi^{-1}(U_1) = \Phi^{-1}(q)) \\ &= 2 \lim_{x\to -\infty} P(X_2 \leqslant x \mid X_1 = x).\end{aligned}$$

Using the fact that $X_2 \mid X_1 = x \sim N(\rho x, 1-\rho^2)$, it can be calculated that

$$\lambda = 2 \lim_{x\to -\infty} \Phi(x\sqrt{1-\rho}/\sqrt{1+\rho}) = 0,$$

Table 7.1. Values of λ, the coefficient of upper and lower tail dependence, for the t copula $C^t_{\nu,\rho}$ for various values of ν, the degrees of freedom, and ρ, the correlation. The last row represents the Gauss copula.

	ρ				
ν	-0.5	0	0.5	0.9	1
2	0.06	0.18	0.39	0.72	1
4	0.01	0.08	0.25	0.63	1
10	0.00	0.01	0.08	0.46	1
∞	0	0	0	0	1

provided $\rho < 1$. Hence, the Gauss copula is asymptotically independent in both tails. Regardless of how high a correlation we choose, if we go far enough into the tail, extreme events appear to occur independently in each margin.

Example 7.39 (asymptotic dependence of the *t* copula). To evaluate the tail-dependence coefficient for the t copula $C^t_{\nu,\rho}$, let $(X_1, X_2) := (t_\nu^{-1}(U_1), t_\nu^{-1}(U_2))$, where t_ν denotes the df of a univariate t distribution with ν degrees of freedom. Thus $(X_1, X_2) \sim t_2(\nu, \mathbf{0}, P)$, where P is a correlation matrix with off-diagonal element ρ. By calculating the conditional density from the joint and marginal densities of a bivariate t distribution, it may be verified that, conditional on $X_1 = x$,

$$\left(\frac{\nu+1}{\nu+x^2}\right)^{1/2}\frac{X_2-\rho x}{\sqrt{1-\rho^2}} \sim t_{\nu+1}. \tag{7.37}$$

Using an argument similar to Example 7.38 we find that

$$\lambda = 2t_{\nu+1}\left(-\sqrt{\frac{(\nu+1)(1-\rho)}{1+\rho}}\right). \tag{7.38}$$

Provided that $\rho > -1$, the copula of the bivariate t distribution is asymptotically dependent in both the upper and lower tails.

In Table 7.1 we tabulate the coefficient of tail dependence for various values of ν and ρ. For fixed ρ the strength of the tail dependence increases as ν decreases, and for fixed ν tail dependence increases as ρ increases. Even for zero or negative correlation values there is some tail dependence. This is not too surprising and can be grasped intuitively by recalling from Section 6.2.1 that the t distribution is a normal mixture distribution with a mixing variable W whose distribution is inverse gamma (which is a heavy-tailed distribution): if $|X_1|$ is large, there is a good chance that this is because W is large, increasing the probability of $|X_2|$ being large.

We could use the same method used in the previous examples to calculate tail-dependence coefficients for other copulas of normal variance mixtures. In doing so we would find that most examples, such as copulas of symmetric hyperbolic or NIG distributions, fell into the same category as the Gauss copula and were asymptotically independent in the tails. The essential determinant of whether the copula of a normal variance mixture has tail dependence or not is the tail of the distribution of the mixing variable W in Definition 6.4. If W has a distribution with a power tail, then we get tail dependence, otherwise we get asymptotic independence. This is a consequence of a general result for elliptical distributions given in Section 16.1.3.

Joint quantile exceedance probabilities. Coefficients of tail dependence are of course asymptotic quantities, and in the remainder of this section we look at joint exceedances of *finite high quantiles* for the Gauss and t copulas in order to learn more about the practical consequences of the differences between the extremal behaviours of these two models.

As motivation we consider Figure 7.9, where 5000 simulated points from four different distributions are displayed. The distributions in (a) and (b) are meta-Gaussian distributions (see Section 7.1.3); they share the same copula C_ρ^{Ga}. The distributions in (c) and (d) are meta-t distributions; they share the same copula $C_{\nu,\rho}^t$. The values of ν and ρ in all parts are 4 and 0.5, respectively. The distributions in (a) and (c) share the same margins, namely standard normal margins. The distributions in (b) and (d) both have Student t margins with four degrees of freedom. The distributions in (a) and (d) are, of course, elliptical, being a standard bivariate normal and a bivariate t distribution with four degrees of freedom; they both have linear correlation $\rho = 0.5$. The other distributions are not elliptical and do not necessarily have linear correlation 50%, since altering the margins alters the linear correlation. All four distributions have identical Kendall's tau values (see Proposition 7.43). The meta-Gaussian distributions have the same Spearman's rho value, as do the meta-t distributions, although the two values are not identical (see Section 7.2.3).

The vertical and horizontal lines mark the true theoretical 0.005 and 0.995 quantiles for all distributions. Note that for the meta-t distributions the number of points that lie below both 0.005 quantiles or exceed both 0.995 quantiles is clearly greater than for the meta-Gaussian distributions, and this can be explained by the tail dependence of the t copula. The true theoretical ratio by which the number of these joint exceedances in the meta-t models should exceed the number in the meta-Gaussian models is 2.79, as may be read from Table 7.2, whose interpretation we now discuss.

In Table 7.2 we have calculated values of $C_\rho^{\mathrm{Ga}}(u,u)/C_{\nu,\rho}^t(u,u)$ for various ρ and ν and $u = 0.05, 0.01, 0.005, 0.001$. The rows marked Gauss contain values of $C_\rho^{\mathrm{Ga}}(u,u)$, which is the probability that two rvs with this copula are below their u-quantiles; we term this event a joint quantile exceedance (thinking of exceedance in the downwards direction). It is obviously identical to the probability that both rvs are larger than their $(1-u)$-quantiles. The remaining rows give the values of the ratio and thus express the amount by which the joint quantile exceedance probabilities must be inflated when we move from models with a Gauss copula to models with a t copula.

In Table 7.3 we extend Table 7.2 to higher dimensions. We now focus only on joint exceedances of the 1% (or 99%) quantile(s). We tabulate values of the ratio $C_P^{\mathrm{Ga}}(u,\dots,u)/C_{\nu,P}^t(u,\dots,u)$, where P is an equicorrelation matrix with all correlations equal to ρ. It is noticeable that not only do these values grow as the correlation parameter or number of degrees of freedom falls, but they also grow with the

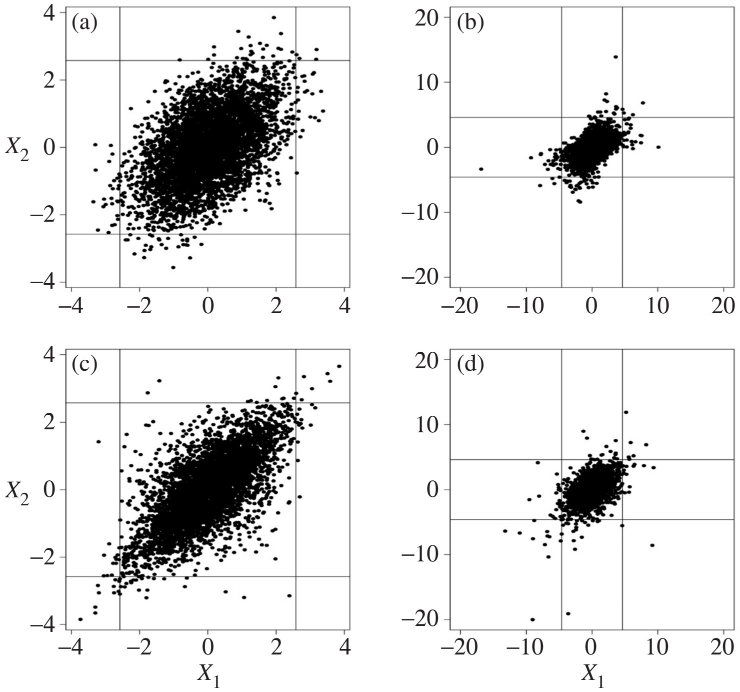

Figure 7.9. Five thousand simulated points from four distributions. (a) Standard bivariate normal with correlation parameter $\rho = 0.5$. (b) Meta-Gaussian distribution with copula C_ρ^{Ga} and Student t margins with four degrees of freedom. (c) Meta-t distribution with copula $C_{4,\rho}^t$ and standard normal margins. (d) Standard bivariate t distribution with four degrees of freedom and correlation parameter $\rho = 0.5$. Horizontal and vertical lines mark the 0.005 and 0.995 quantiles. See Section 7.3.1 for a commentary.

Table 7.2. Joint quantile exceedance probabilities for bivariate Gauss and t copulas with correlation parameter values of 0.5 and 0.7. For Gauss copulas the probability of joint quantile exceedance is given; for the t copulas the factors by which the Gaussian probability must be multiplied are given.

			Quantile			
ρ	Copula	ν	0.05	0.01	0.005	0.001
0.5	Gauss		1.21×10^{-2}	1.29×10^{-3}	4.96×10^{-4}	5.42×10^{-5}
0.5	t	8	1.20	1.65	1.94	3.01
0.5	t	4	1.39	2.22	2.79	4.86
0.5	t	3	1.50	2.55	3.26	5.83
0.7	Gauss		1.95×10^{-2}	2.67×10^{-3}	1.14×10^{-3}	1.60×10^{-4}
0.7	t	8	1.11	1.33	1.46	1.86
0.7	t	4	1.21	1.60	1.82	2.52
0.7	t	3	1.27	1.74	2.01	2.83

Table 7.3. Joint 1% quantile exceedance probabilities for multivariate Gaussian and t equicorrelation copulas with correlation parameter values of 0.5 and 0.7. For Gauss copulas the probability of joint quantile exceedance is given; for the t copulas the factors by which the Gaussian probability must be multiplied are given.

			Dimension d			
ρ	Copula	ν	2	3	4	5
0.5	Gauss		1.29×10^{-3}	3.66×10^{-4}	1.49×10^{-4}	7.48×10^{-5}
0.5	t	8	1.65	2.36	3.09	3.82
0.5	t	4	2.22	3.82	5.66	7.68
0.5	t	3	2.55	4.72	7.35	10.34
0.7	Gauss		2.67×10^{-3}	1.28×10^{-3}	7.77×10^{-4}	5.35×10^{-4}
0.7	t	8	1.33	1.58	1.78	1.95
0.7	t	4	1.60	2.10	2.53	2.91
0.7	t	3	1.74	2.39	2.97	3.45

dimension of the copula. The next example gives an interpretation of one of these numbers.

Example 7.40 (joint quantile exceedances: an interpretation). Consider daily returns on five stocks. Suppose we are unsure about the best multivariate elliptical model for these returns, but we believe that the correlation between any two returns on the same day is 50%. If returns follow a multivariate Gaussian distribution, then the probability that on any day all returns are below the 1% quantiles of their respective distributions is 7.48×10^{-5}. In the long run, such an event will happen once every 13 369 trading days on average: that is, roughly once every 51.4 years (assuming 260 trading days in a year). On the other hand, if returns follow a multivariate t distribution with four degrees of freedom, then such an event will happen 7.68 times more often: that is, roughly once every 6.7 years. In the life of a risk manager, 50-year events and 7-year events have a very different significance.

7.3.2 Rank Correlations

To calculate rank correlations for normal variance mixture copulas we use the following preliminary result for elliptical distributions.

Proposition 7.41. *Let $\boldsymbol{X} \sim E_2(\boldsymbol{0}, \Sigma, \psi)$ and $\rho = \wp(\Sigma)_{12}$, where $\wp$ denotes the correlation operator in (6.5). Assume $P(\boldsymbol{X} = \boldsymbol{0}) = 0$. Then*

$$P(X_1 > 0, X_2 > 0) = \tfrac{1}{4} + \frac{\arcsin \rho}{2\pi}.$$

Proof. First we make a standardization of the variables and observe that if $\boldsymbol{Y} \sim E_2(0, P, \psi)$ and $P = \wp(\Sigma)$, then $P(X_1 > 0, X_2 > 0) = P(Y_1 > 0, Y_2 > 0)$. Now introduce a pair of spherical variates $\boldsymbol{Z} \sim S_2(\psi)$; it follows that

$$\begin{aligned}(Y_1, Y_2) &\overset{\mathrm{d}}{=} (Z_1, \rho Z_1 + \sqrt{1-\rho^2} Z_2)\\ &\overset{\mathrm{d}}{=} R(\cos\Theta, \rho\cos\Theta + \sqrt{1-\rho^2}\sin\Theta),\end{aligned}$$

where R is a positive radial rv and Θ is an independent, uniformly distributed angle on $[-\pi, \pi)$ (see Section 6.3.1 and Theorem 6.21). Let $\phi = \arcsin \rho$ and observe that $\sin \phi = \rho$ and $\cos \phi = \sqrt{1-\rho^2}$. Since $P(R=0) = P(\boldsymbol{X} = \boldsymbol{0}) = 0$ we conclude that

$$\begin{aligned} P(X_1 > 0, X_2 > 0) &= P(\cos \Theta > 0, \sin \phi \cos \Theta + \cos \phi \sin \Theta > 0) \\ &= P(\cos \Theta > 0, \sin(\Theta + \phi) > 0). \end{aligned}$$

The angle Θ must jointly satisfy $\Theta \in (-\frac{1}{2}\pi, \frac{1}{2}\pi)$ and $\Theta + \phi \in (0, \pi)$, and it is easily seen that for any value of ϕ this has probability $(\frac{1}{2}\pi + \phi)/(2\pi)$, which gives the result. □

Theorem 7.42 (rank correlations for the Gauss copula). *Let $\boldsymbol{X}$ have a bivariate meta-Gaussian distribution with copula C_ρ^{Ga} and continuous margins. Then the rank correlations are*

$$\rho_\tau(X_1, X_2) = \frac{2}{\pi} \arcsin \rho, \tag{7.39}$$

$$\rho_{\mathrm{S}}(X_1, X_2) = \frac{6}{\pi} \arcsin \tfrac{1}{2}\rho. \tag{7.40}$$

Proof. Since rank correlation is a copula property, we can of course simply assume that $\boldsymbol{X} \sim N_2(\boldsymbol{0}, P)$, where P is a correlation matrix with off-diagonal element ρ; the calculations are then easy. For Kendall's tau, formula (7.29) implies

$$\rho_\tau(X_1, X_2) = 4P(Y_1 > 0, Y_2 > 0) - 1,$$

where $\boldsymbol{Y} = \tilde{\boldsymbol{X}} - \boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ is an independent copy of $\boldsymbol{X}$. Since, by the convolution property of the multivariate normal distribution in Section 6.1.3, $\boldsymbol{Y} \sim N_2(\boldsymbol{0}, 2P)$, we have that $\rho(Y_1, Y_2) = \rho$ and formula (7.39) follows from Proposition 7.41.

For Spearman's rho, let $\boldsymbol{Z} = (Z_1, Z_2)'$ be a vector consisting of two independent standard normal random variables and observe that formula (7.31) implies

$$\begin{aligned} \rho_{\mathrm{S}}(X_1, X_2) &= 3(2P((X_1 - Z_1)(X_2 - Z_2) > 0) - 1) \\ &= 3(4P(X_1 - Z_1 > 0, X_2 - Z_2 > 0) - 1) \\ &= 3(4P(Y_1 > 0, Y_2 > 0) - 1), \end{aligned}$$

where $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{Z}$. Since $\boldsymbol{Y} \sim N_2(\boldsymbol{0}, (P + I_2))$, the formula (7.40) follows from Proposition 7.41 and the fact that $\rho(Y_1, Y_2) = \rho/2$. □

These relationships between the rank correlations and ρ are illustrated in Figure 7.10. Note that the right-hand side of (7.40) may be approximated by the value ρ itself. This approximation turns out to be very accurate, as shown in the figure; the error bounds are $|6 \arcsin(\rho/2)/\pi - \rho| \leqslant (\pi - 3)|\rho|/\pi \leqslant 0.0181$.

The relationship between Kendall's tau and the correlation parameter of the Gauss copula C_ρ^{Ga} expressed by (7.39) holds more generally for the copulas of all elliptical distributions that exclude point mass at their centre, including, for example, the t copula $C_{\nu,\rho}^t$. This is a consequence of the following result, which was already used to derive an alternative correlation estimator for bivariate distributions in Section 6.3.4.

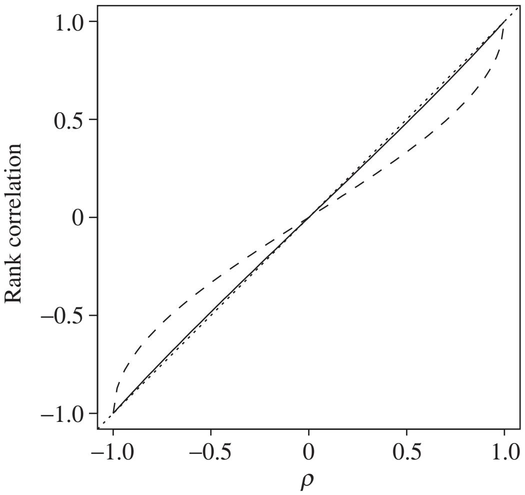

Figure 7.10. The solid line shows the relationship between Spearman's rho and the correlation parameter ρ of the Gauss copula C_ρ^{Ga} for meta-Gaussian rvs with continuous dfs; this is very close to the line $y = x$, which is just visible as a dotted line. The dashed line shows the relationship between Kendall's tau and ρ; this relationship holds for the copulas of other normal variance mixture distributions with correlation parameter ρ, such as the t copula $C^t_{\nu,\rho}$.

Proposition 7.43. *Let $\boldsymbol{X} \sim E_2(\boldsymbol{0}, P, \psi)$ for a correlation matrix P with off-diagonal element ρ, and assume that $P(\boldsymbol{X} = \boldsymbol{0}) = 0$. Then the relationship $\rho_\tau(X_1, X_2) = (2/\pi)\arcsin\rho$ holds.*

Proof. The result relies on the convolution property of elliptical distributions in (6.47). Setting $\boldsymbol{Y} = \tilde{\boldsymbol{X}} - \boldsymbol{X}$, where $\tilde{\boldsymbol{X}}$ is an independent copy of $\boldsymbol{X}$, we note that $\boldsymbol{Y} \sim E_2(\boldsymbol{0}, P, \tilde{\psi})$ for some characteristic generator $\tilde{\psi}$. We need to evaluate $\rho_\tau(X_1, X_2) = 4P(Y_1 > 0, Y_2 > 0) - 1$ as in the proof of Theorem 7.42, but Proposition 7.41 shows that $P(Y_1 > 0, Y_2 > 0)$ takes the same value whenever $\boldsymbol{Y}$ is elliptical. □

The relationship (7.40) between Spearman's rho and the correlation parameter of the Gauss copula does not hold for the copulas of all elliptical distributions. We can, however, derive a formula for the copulas of normal variance mixture distributions based on the following result.

Proposition 7.44. *Let $\boldsymbol{X} \sim M_2(\boldsymbol{0}, P, \hat{H})$ be distributed according to a normal variance mixture distribution for a correlation matrix P with off-diagonal element ρ, and assume that $P(\boldsymbol{X} = \boldsymbol{0}) = 0$. Then*

$$\rho_S(X_1, X_2) = \frac{6}{\pi} E\left(\arcsin\left(\rho \frac{W}{\sqrt{(W + \tilde{W})(W + \bar{W})}}\right)\right), \tag{7.41}$$

where W, $\tilde{W}$ and $\bar{W}$ are independent random variables with df H such that the Laplace–Stieltjes transform of H is $\hat{H}$.

Proof. Assume that $\boldsymbol{X} = \sqrt{W}\boldsymbol{Z}$, where $\boldsymbol{Z} \sim N_2(\boldsymbol{0}, P)$. Let $\tilde{Z}$ and $\bar{Z}$ be standard normal variables and assume that $\boldsymbol{Z}$, $\tilde{Z}$, $\bar{Z}$, W, $\tilde{W}$ and $\bar{W}$ are all independent. Write $\tilde{X} := \sqrt{\tilde{W}}\tilde{Z}$, $\bar{X} := \sqrt{\bar{W}}\bar{Z}$ and

$$\begin{aligned} Y_1 &:= X_1 - \tilde{X} = \sqrt{W}Z_1 - \sqrt{\tilde{W}}\tilde{Z}, \\ Y_2 &:= X_2 - \bar{X} = \sqrt{W}Z_2 - \sqrt{\bar{W}}\bar{Z}. \end{aligned}$$

The result is proved by applying a similar approach to Theorem 7.42 and conditioning on the variables W, $\tilde{W}$ and $\bar{W}$. We note that the conditional distribution of $\boldsymbol{Y} = (Y_1, Y_2)'$ satisfies

$$\boldsymbol{Y} \mid W, \tilde{W}, \bar{W} \sim N_2\left(\boldsymbol{0}, \begin{pmatrix} W + \tilde{W} & W\rho \\ W\rho & W + \bar{W} \end{pmatrix}\right).$$

Using formula (7.31) we calculate

$$\begin{aligned} \rho_S(X_1, X_2) &= 6P((X_1 - \tilde{X}_1)(X_2 - \bar{X}_2) > 0) - 3 \\ &= 3(2E(P(Y_1Y_2 > 0 \mid W, \tilde{W}, \bar{W})) - 1) \\ &= 3E(4P(Y_1 > 0, Y_2 > 0 \mid W, \tilde{W}, \bar{W}) - 1), \end{aligned}$$

and (7.41) is obtained by applying Proposition 7.41 and using the fact that, given W, $\tilde{W}$ and $\bar{W}$,

$$\rho(Y_1, Y_2) = \rho \frac{W}{\sqrt{(W + \tilde{W})(W + \bar{W})}}.$$

□

The formula (7.41) reduces to the formula (7.40) in the case where $W = \tilde{W} = \bar{W} = k$ for some positive constant k. In general, Spearman's rho for the copulas of normal variance mixtures can be calculated accurately by approximating the integral in (7.41) using Monte Carlo. For example, to calculate Spearman's rho for the t copula $C^t_{\nu,\rho}$ we would generate a set of inverse gamma variates $\{W_j, \tilde{W}_j, \bar{W}_j,\ j = 1, \dots, m\}$ for m large, such that each variable in the set had an independent $\mathrm{Ig}(\frac{1}{2}\nu, \frac{1}{2}\nu)$ distribution. We would then use

$$\rho_S(X_1, X_2) \approx \frac{6}{m\pi} \sum_{j=1}^{m} \arcsin\left(\rho \frac{W_j}{\sqrt{(W_j + \tilde{W}_j)(W_j + \bar{W}_j)}}\right). \tag{7.42}$$

7.3.3 Skewed Normal Mixture Copulas

A skewed normal mixture copula is the copula of any normal mixture distribution that is not elliptically symmetric. An example is provided by the *skewed t copula*, which is the copula of the distribution whose density is given in (6.31).

A random vector $\boldsymbol{X}$ with a skewed t distribution and ν degrees of freedom is denoted by $\boldsymbol{X} \sim \mathrm{GH}_d(-\frac{1}{2}\nu, \nu, 0, \boldsymbol{\mu}, \Sigma, \boldsymbol{\gamma})$ in the notation of Section 6.2.3. Its marginal distributions satisfy $X_i \sim \mathrm{GH}_1(-\frac{1}{2}\nu, \nu, 0, \mu_i, \Sigma_{ii}, \gamma_i)$ (from Proposition 6.13) and its copula depends on ν, $P = \wp(\Sigma)$ and $\boldsymbol{\gamma}$ and will be denoted by $C^t_{\nu,P,\boldsymbol{\gamma}}$ or, in the bivariate case, $C^t_{\nu,\rho,\gamma_1,\gamma_2}$. Random sampling from the skewed t copula follows the general approach of Algorithm 7.10.

Algorithm 7.45 (simulation of the skewed t copula).

(1) Generate $\boldsymbol{X} \sim \mathrm{GH}_d(-\frac{1}{2}\nu, \nu, 0, \mathbf{0}, P, \boldsymbol{\gamma})$ using Algorithm 6.10.

(2) Return $\boldsymbol{U} = (F_1(X_1), \dots, F_d(X_d))'$, where F_i is the distribution function of a $\mathrm{GH}_1(-\frac{1}{2}\nu, \nu, 0, 0, 1, \gamma_i)$ distribution. The random vector $\boldsymbol{U}$ has df $C^t_{\nu,P,\boldsymbol{\gamma}}$.

Note that the evaluation of F_i requires the numerical integration of the density of a skewed univariate t distribution.

To appreciate the flexibility of the skewed t copula it suffices to consider the bivariate case for different values of the skewness parameters γ_1 and γ_2. In Figure 7.11 we have plotted simulated points from nine different examples of this copula. Part (e) corresponds to the case when $\gamma_1 = \gamma_2 = 0$ and is thus the ordinary t copula. All other pictures show copulas that are non-radially symmetric (see Section 7.1.5), as is obvious by rotating each picture 180° about the point $(\frac{1}{2}, \frac{1}{2})$; (c), (e) and (g) show exchangeable copulas satisfying (7.20), while the remaining six are non-exchangeable.

Obviously, the main advantage of the skewed t copula over the ordinary t copula is that its asymmetry allows us to have different levels of tail dependence in "opposite corners" of the distribution. In the context of market risk it is often claimed that joint negative returns on stocks show more tail dependence than joint positive returns.

7.3.4 Grouped Normal Mixture Copulas

Technically speaking, a grouped normal mixture copula is not itself the copula of a normal mixture distribution, but rather a way of attaching together a set of normal mixture copulas. We will illustrate the idea by considering the *grouped t copula*. Here, the basic idea is to construct a copula for a random vector $\boldsymbol{X}$ such that certain subvectors of $\boldsymbol{X}$ have t copulas but quite different levels of tail dependence.

We create a distribution using a generalization of the variance-mixing construction $\boldsymbol{X} = \sqrt{W}\boldsymbol{Z}$ in (6.18). Rather than multiplying all components of a correlated Gaussian vector $\boldsymbol{Z}$ with the root of a single inverse-gamma-distributed variate W, as in Example 6.7, we instead multiply different subgroups with different variates W_j, where $W_j \sim \mathrm{Ig}(\frac{1}{2}\nu_j, \frac{1}{2}\nu_j)$ and the W_j are themselves comonotonic (see Section 7.2.1). We therefore create subgroups whose dependence properties are described by t copulas with different ν_j parameters. The groups may even consist of a single member for each ν_j parameter, an idea that has been developed by Luo and Shevchenko (2010) under the name of the generalized t copula.

Like the t copula, the skewed t copula and anything based on a mixture of multivariate normals, a grouped t copula is easy to simulate and thus to use in Monte Carlo risk studies—this has been a major motivation for its development. We formally define the grouped t copula by explaining in more detail how to generate a random vector $\boldsymbol{U}$ with that distribution.

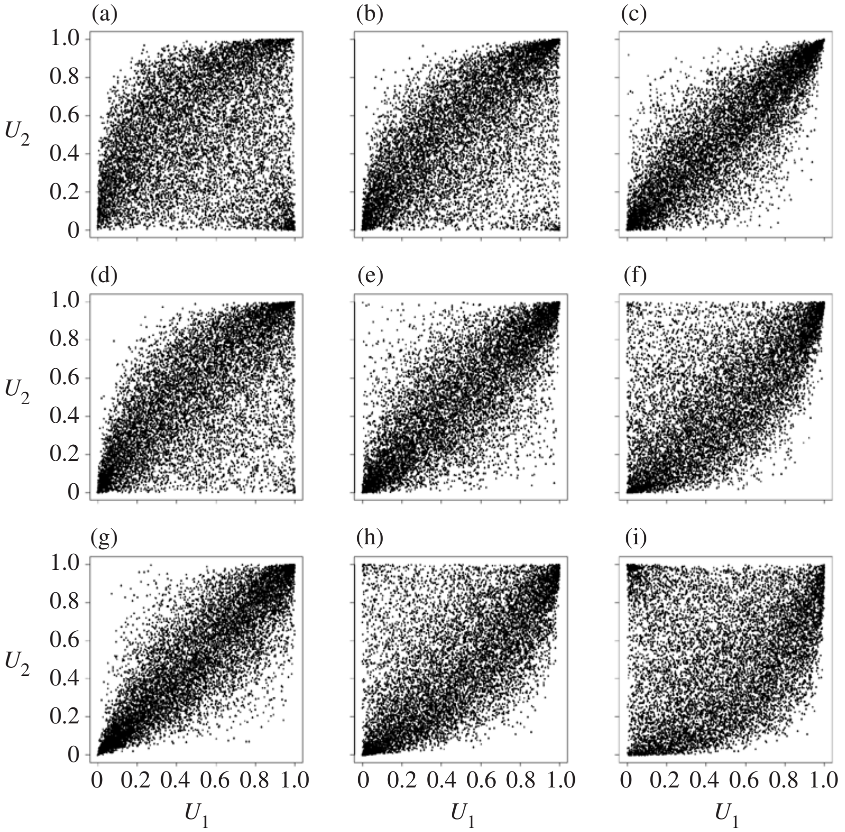

Figure 7.11. Ten thousand simulated points from bivariate skewed t copula $C^t_{\nu,\rho,\gamma_1,\gamma_2}$ for $\nu = 5$, $\rho = 0.8$ and various values of the parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)'$: (a) $\boldsymbol{\gamma} = (0.8, -0.8)'$; (b) $\boldsymbol{\gamma} = (0.8, 0)'$; (c) $\boldsymbol{\gamma} = (0.8, 0.8)'$; (d) $\boldsymbol{\gamma} = (0, -0.8)'$; (e) $\boldsymbol{\gamma} = (0, 0)'$; (f) $\boldsymbol{\gamma} = (0, 0.8)'$; (g) $\boldsymbol{\gamma} = (-0.8, -0.8)'$; (h) $\boldsymbol{\gamma} = (-0.8, 0)'$; and (i) $\boldsymbol{\gamma} = (-0.8, 0.8)'$.

Algorithm 7.46 (simulation of the grouped *t* copula).

(1) Generate independently $\boldsymbol{Z} \sim N_d(\boldsymbol{0}, P)$ and $U \sim U(0, 1)$.

(2) Partition $\{1, \dots, d\}$ into m subsets of sizes $s_1, \dots, s_m$, and for $k = 1, \dots, m$ let ν_k be the degrees-of-freedom parameter associated with group k.

(3) Set $W_k = G^{-1}_{\nu_k}(U)$, where G_ν is the df of the univariate $\text{Ig}(\frac{1}{2}\nu, \frac{1}{2}\nu)$ distribution, so that $W_1, \dots, W_m$ are comonotonic and inverse-gamma-distributed variates.

(4) Construct vectors $\boldsymbol{X}$ and $\boldsymbol{U}$ by

$$\boldsymbol{X} = (\sqrt{W_1}Z_1, \dots, \sqrt{W_1}Z_{s_1}, \sqrt{W_2}Z_{s_1+1}, \dots, \sqrt{W_2}Z_{s_1+s_2}, \dots, \sqrt{W_m}Z_d)',$$
$$\boldsymbol{U} = (t_{\nu_1}(X_1), \dots, t_{\nu_1}(X_{s_1}), t_{\nu_2}(X_{s_1+1}), \dots, t_{\nu_2}(X_{s_1+s_2}), \dots, t_{\nu_m}(X_d))'.$$

The former has a grouped t distribution and the latter is distributed according to a grouped t copula.

If we have an a priori idea of the desired group structure, we can calibrate the grouped t copula to data using a method based on Kendall's tau rank correlations. The use of this method for the ordinary t copula is described later in Section 7.5.1 and Example 7.56.

Notes and Comments

The coefficient of tail dependence for the t copula was first derived in Embrechts, McNeil and Straumann (2002). A more general result for the copulas of elliptical distributions is given in Hult and Lindskog (2002) and will be discussed in Section 16.1.3. The formula for Kendall's tau for elliptical distributions can be found in Lindskog, McNeil and Schmock (2003) and Fang and Fang (2002).

Proposition 7.44 is due to Andrew D. Smith (personal correspondence), who has also derived the attractive alterative formula

$$\rho_{\mathrm{S}}(X_1, X_2) = (6/\pi)E(\arcsin(\rho\sin(\Theta/2))),$$

where Θ is the (random) angle in a triangle with side lengths $(W^{-1} + \tilde{W}^{-1})$, $(W^{-1} + \bar{W}^{-1})$ and $(\tilde{W}^{-1} + \bar{W}^{-1})$.

The skewed t copula was introduced in Demarta and McNeil (2005), which also describes the grouped t copula. The grouped t copula and a method for its calibration were first proposed in Daul et al. (2003). The special case of the grouped t copula with one member in each group has been investigated by Luo and Shevchenko (2010, 2012), who refer to this as a generalized t copula.

7.4 Archimedean Copulas

The Gumbel copula (7.12) and the Clayton copula (7.13) belong to the family of so-called Archimedean copulas, which has been very extensively studied. This family has proved useful for modelling portfolio credit risk, as will be seen in Example 11.4. In this section we look at the simple structure of these copulas and establish some of the properties that we will need.

7.4.1 Bivariate Archimedean Copulas

As well as the Gumbel and Clayton copulas, two further examples we consider are the *Frank copula*,

$$C_\theta^{\mathrm{Fr}}(u_1, u_2) = -\frac{1}{\theta}\ln\left(1 + \frac{(\mathrm{e}^{-\theta u_1} - 1)(\mathrm{e}^{-\theta u_2} - 1)}{\mathrm{e}^{-\theta} - 1}\right), \quad \theta \in \mathbb{R},$$

and a two-parameter copula that we refer to as a *generalized Clayton copula*,

$$C_{\theta,\delta}^{\mathrm{GC}}(u_1, u_2) = (((u_1^{-\theta} - 1)^\delta + (u_2^{-\theta} - 1)^\delta)^{1/\delta} + 1)^{-1/\theta}, \quad \theta \geqslant 0,\ \delta \geqslant 1.$$

It may be verified that, provided the parameter θ lies in the ranges we have specified in the copula definitions, all four examples that we have met have the form

$$C(u_1, u_2) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2)), \tag{7.43}$$

Table 7.4. Table summarizing the generators, permissible parameter values and limiting special cases for four selected Archimedean copulas. The case $\theta = 0$ should be taken to mean the limit $\lim_{\theta \to 0} \psi_\theta(t)$. For the Clayton and Frank copulas this limit is e^{-t}, which is the generator of the independence copula.

Copula	Generator $\psi(t)$	Parameter range	$\psi \in \Psi_\infty$	Lower	Upper
C_θ^{Gu}	$\exp(-t^{1/\theta})$	$\theta \geqslant 1$	Yes	Π	M
C_θ^{Cl}	$\max((1+\theta t)^{-1/\theta}, 0)$	$\theta \geqslant -1$	$\theta \geqslant 0$	W	M
C_θ^{Fr}	$-\dfrac{1}{\theta}\ln(1+(e^{-\theta}-1)e^{-t})$	$\theta \in \mathbb{R}$	$\theta \geqslant 0$	W	M
$C_{\theta,\delta}^{\mathrm{GC}}$	$(1+\theta t^{1/\delta})^{-1/\theta}$	$\theta \geqslant 0, \delta \geqslant 1$	Yes	N/A	N/A

where $\psi\colon [0,\infty) \to [0,1]$ is a decreasing and continuous function that satisfies the conditions $\psi(0) = 1$ and $\lim_{t\to\infty} \psi(t) = 0$. The function ψ is known as the *generator* of the copula. For example, for the Gumbel copula $\psi(t) = \exp(-t^{1/\theta})$ for $\theta \geqslant 1$, and for the other copulas the generators are given in Table 7.4.

When we introduced the Clayton copula in (7.13) we insisted that its parameter should be non-negative. In the table we also define a Clayton copula for $-1 \leqslant \theta < 0$. To accommodate this case, the generator is written $\psi(t) = \max((1+\theta t)^{-1/\theta}, 0)$. We observe that $\psi(t)$ is strictly decreasing on $[0, -1/\theta)$ but $\psi(t) = 0$ on $[-1/\theta, \infty)$. To define the generator inverse uniquely at zero we set $\psi^{-1}(0) = \inf\{t\colon \psi(t) = 0\} = -1/\theta$.

In Table 7.4 we also give the lower and upper limits of the families as the parameter θ goes to the boundaries of the parameter space. Both the Frank and Clayton copulas are known as *comprehensive* copulas, since they interpolate between a lower limit of countermonotonicity and an upper limit of comonotonicity. For a more extensive table of Archimedean copulas see Nelsen (2006).

The following important theorem clarifies the conditions under which a function ψ is the generator of a bivariate Archimedean copula and allows us to define an Archimedean copula generator formally.

Theorem 7.47 (bivariate Archimedean copula). *Let* $\psi\colon [0,\infty) \to [0,1]$ *be a decreasing, continuous function that satisfies the conditions* $\psi(0) = 1$ *and* $\lim_{t\to\infty} \psi(t) = 0$. *Then*

$$C(u_1, u_2) = \psi(\psi^{-1}(u_1) + \psi^{-1}(u_2)) \tag{7.44}$$

is a copula if and only if ψ *is convex.*

Proof. See Nelsen (2006, Theorem 4.1.4). □

Definition 7.48 (Archimedean copula generator). A decreasing, continuous, convex function $\psi\colon [0,\infty) \to [0,1]$ satisfying $\psi(0) = 1$ and $\lim_{t\to\infty} \psi(t) = 0$ is known as an Archimedean copula generator.

Table 7.5. Kendall's rank correlations and coefficients of tail dependence for the copulas of Table 7.4. $D_1(\theta)$ is the Debye function $D_1(\theta) = \theta^{-1}\int_0^\theta t/(e^t - 1)\,dt$.

Copula	ρ_τ	λ_u	λ_l
C_θ^{Gu}	$1 - 1/\theta$	$2 - 2^{1/\theta}$	0
C_θ^{Cl}	$\theta/(\theta + 2)$	0	$\begin{cases} 2^{-1/\theta}, & \theta > 0, \\ 0, & \theta \leqslant 0, \end{cases}$
C_θ^{Fr}	$1 - 4\theta^{-1}(1 - D_1(\theta))$	0	0
$C_{\theta,\delta}^{\mathrm{GC}}$	$\dfrac{(2+\theta)\delta - 2}{(2+\theta)\delta}$	$2 - 2^{1/\delta}$	$2^{-1/(\theta\delta)}$

Note that this definition automatically implies that ψ is strictly decreasing at all values t for which $\psi(t) > 0$, but there may be a flat piece if ψ attains the value zero. This is the only point where there is ambiguity about the inverse ψ^{-1}, and we set $\psi^{-1}(0) = \inf\{t : \psi(t) = 0\}$.

Kendall's rank correlations can be calculated for Archimedean copulas directly from the generator inverse using Proposition 7.49 below. The formula obtained can be used to calibrate Archimedean copulas to empirical data using the sample version of Kendall's tau, as we discuss in Section 7.5.

Proposition 7.49. *Let X_1 and X_2 be continuous rvs with unique Archimedean copula C generated by ψ. Then*

$$\rho_\tau(X_1, X_2) = 1 + 4\int_0^1 \frac{\psi^{-1}(t)}{d\psi^{-1}(t)/dt}\,dt. \tag{7.45}$$

Proof. See Nelsen (2006, Corollary 5.1.4). □

For the closed-form copulas of the Archimedean class, coefficients of tail dependence are easily calculated using methods of the kind used in Example 7.37. Values for Kendall's tau and the coefficients of tail dependence for the copulas of Table 7.4 are given in Table 7.5. It is interesting to note that the generalized Clayton copula $C_{\theta,\delta}^{\mathrm{GC}}$ combines, in a sense, both Gumbel's family and Clayton's family for positive parameter values, and thus succeeds in having tail dependence in both tails.

7.4.2 Multivariate Archimedean Copulas

It seems natural to attempt to construct a higher-dimensional Archimedean copula according to

$$C(u_1, \dots, u_d) = \psi(\psi^{-1}(u_1) + \cdots + \psi^{-1}(u_d)), \tag{7.46}$$

where ψ is an Archimedean generator function as in Definition 7.48. However, this construction may fail to define a proper distribution function for arbitrary dimension d. An example where this occurs is obtained if we take the generator $\psi(t) = 1 - t$, which is the Clayton generator for $\theta = -1$. In this case we obtain the Fréchet lower bound for copulas, which is not itself a copula for $d > 2$.

In order to guarantee that we will obtain a proper copula in any dimension we have to impose the property of *complete monotonicity* on ψ. A decreasing function $f(t)$ is completely monotonic on an interval $[a, b]$ if it satisfies

$$(-1)^k \frac{\mathrm{d}^k}{\mathrm{d}t^k} f(t) \geqslant 0, \quad k \in \mathbb{N},\ t \in (a, b). \tag{7.47}$$

Theorem 7.50. *If $\psi\colon [0, \infty) \to [0, 1]$ is an Archimedean copula generator, then the construction (7.46) gives a copula in any dimension d if and only if ψ is completely monotonic.*

Proof. See Kimberling (1974). □

If an Archimedean copula generator is completely monotonic, we write $\psi \in \Psi_\infty$. A column in Table 7.4 shows the cases where the generators are completely monotonic. For example, the Clayton generator is completely monotonic when $\theta \geqslant 0$ and the d-dimensional Clayton copula takes the form

$$C_\theta^{\mathrm{Cl}}(\boldsymbol{u}) = (u_1^{-\theta} + \cdots + u_d^{-\theta} - d + 1)^{-1/\theta}, \quad \theta \geqslant 0, \tag{7.48}$$

where the limiting case $\theta = 0$ should be interpreted as the d-dimensional independence copula.

The class of completely monotonic Archimedean copula generators is equivalent to the class of Laplace–Stieltjes transforms of dfs G on $[0, \infty)$ such that $G(0) = 0$. Let X be an rv with such a df G. We recall that the Laplace–Stieltjes transform of G (or X) is given by

$$\hat{G}(t) = \int_0^\infty \mathrm{e}^{-tx}\,\mathrm{d}G(x) = E(\mathrm{e}^{-tX}), \quad t \geqslant 0. \tag{7.49}$$

It is not difficult to verify that $\hat{G}\colon [0, \infty) \to [0, 1]$ is a continuous, strictly decreasing function with the property of complete monotonicity (7.47). Moreover, $\hat{G}(0) = 1$ and the exclusion of distributions with point mass at zero ensures $\lim_{t\to\infty} \hat{G}(t) = 0$. $\hat{G}$ therefore provides a candidate for an Archimedean generator that will generate a copula in any dimension.

This insight has a number of practical implications. On the one hand, we can create a rich variety of Archimedean copulas by considering Laplace–Stieltjes transforms of different distributions on $[0, \infty)$. On the other hand, we can derive a generic method of sampling from Archimedean copulas based on the following result.

Proposition 7.51. *Let G be a df on $[0, \infty)$ satisfying $G(0) = 0$ with Laplace–Stieltjes transform $\hat{G}$ as in (7.49). Let V be an rv with df G and let $Y_1, \dots, Y_d$ be a sequence of independent, standard exponential rvs that are also independent of V. Then the following hold.*

(i) *The survival copula of the random vector $\boldsymbol{X} := (Y_1/V, \dots, Y_d/V)$ is an Archimedean copula C with generator $\psi = \hat{G}$.*

(ii) *The random vector $\boldsymbol{U} := (\psi(X_1), \dots, \psi(X_d))'$ is distributed according to C.*

(iii) *The components of $\boldsymbol{U}$ are conditionally independent given V with conditional df $P(U_i \leqslant u \mid V = v) = \exp(-v\psi^{-1}(u))$.*

Proof. For part (i) we calculate that, for $\boldsymbol{x} \in \mathbb{R}^d_+$,

$$\begin{aligned} P(X_1 > x_1, \dots, X_d > x_d) &= \int_0^\infty \prod_{i=1}^d \mathrm{e}^{-x_i v}\, \mathrm{d}G(v) \\ &= \int_0^\infty \mathrm{e}^{-v(x_1+\cdots+x_d)}\, \mathrm{d}G(v) \\ &= \hat{G}(x_1 + \cdots + x_d). \end{aligned} \tag{7.50}$$

Since the marginal survival functions are given by $P(X_i > x) = \hat{G}(x)$ and $\hat{G}$ is continuous and strictly decreasing, the result follows from writing

$$P(X_1 > x_1, \dots, X_d > x_d) = \hat{G}(\hat{G}^{-1}(P(X_1 > x_1)) + \cdots + \hat{G}^{-1}(P(X_d > x_d))).$$

Part (ii) follows easily from (7.50) since

$$\begin{aligned} P(\boldsymbol{U} \leqslant \boldsymbol{u}) &= P(U_1 < u_1, \dots, U_d < u_d) \\ &= P(X_1 > \psi^{-1}(u_1), \dots, X_d > \psi^{-1}(u_d)). \end{aligned}$$

The conditional independence is obvious in part (iii) and we calculate that

$$\begin{aligned} P(U_i \leqslant u \mid V = v) &= P(X_i > \psi^{-1}(u) \mid V = v) \\ &= P(Y_i > v\psi^{-1}(u)) \\ &= \mathrm{e}^{-v\psi^{-1}(u)}. \end{aligned}$$

□

Because of the importance of such copulas, particularly in the field of credit risk, we will call these copulas LT-Archimedean (LT stands for "Laplace transform") and make the following definition.

Definition 7.52 (LT-Archimedean copula). An LT-Archimedean copula is a copula of the form (7.46), where ψ is the Laplace–Stieltjes transform of a df G on $[0, \infty)$ satisfying $G(0) = 0$.

The sampling algorithm is based on parts (i) and (ii) of Proposition 7.51. We give explicit instructions for the Clayton, Gumbel and Frank copulas.

Algorithm 7.53 (simulation of LT-Archimedean copulas).

(1) Generate a variate V with df G such that $\hat{G}$, the Laplace–Stieltjes transform of G, is the generator ψ of the required copula.

(2) Generate independent uniform variates $Z_1, \dots, Z_d$ and set $Y_i = -\ln(Z_i)$ for $i = 1, \dots, d$ so that $Y_1, \dots, Y_d$ are standard exponential.

(3) Return $\boldsymbol{U} = (\psi(Y_1/V), \dots, \psi(Y_d/V))'$.

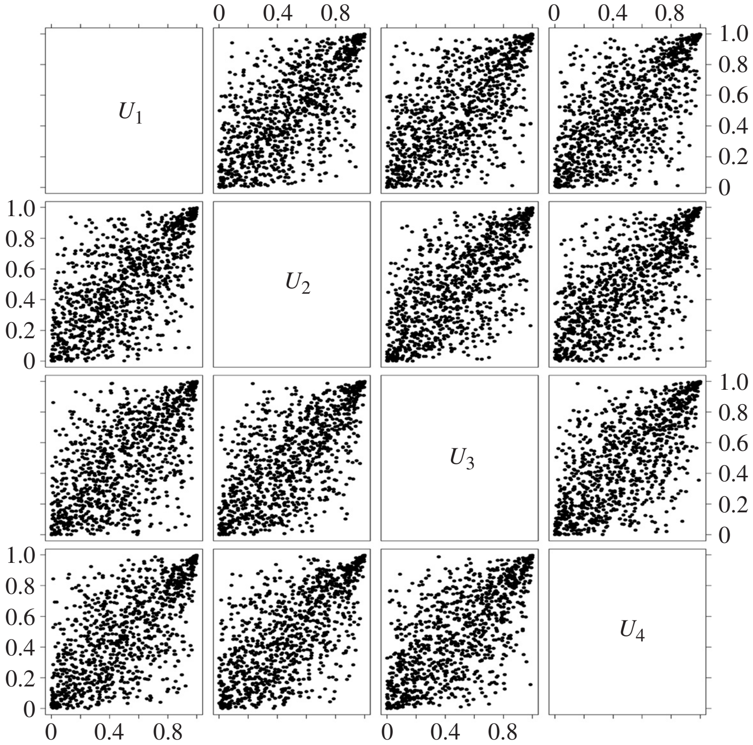

Figure 7.12. Pairwise scatterplots of 1000 simulated points from a four-dimensional exchangeable Gumbel copula with $\theta = 2$. Data are simulated using Algorithm 7.53.

(a) For the special case of the Clayton copula we generate a gamma variate $V \sim \text{Ga}(1/\theta, 1)$ with $\theta > 0$ (see Section A.2.4). The df of V has Laplace transform $\hat{G}(t) = (1 + t)^{-1/\theta}$. This differs slightly from the Clayton generator in Table 7.4 but we note that $\hat{G}(\theta t)$ and $\hat{G}(t)$ generate the same Archimedean copula.

(b) For the special case of the Gumbel copula we generate a positive stable variate $V \sim \text{St}(1/\theta, 1, \gamma, 0)$, where $\gamma = (\cos(\pi/(2\theta)))^{\theta}$ and $\theta > 1$ (see Section A.2.9 for more details and a reference to a simulation algorithm). This df has Laplace transform $\hat{G}(t) = \exp(-t^{1/\theta})$ as desired.

(c) For the special case of the Frank copula we generate a discrete rv V with probability mass function $p(k) = P(V = k) = (1 - \mathrm{e}^{-\theta})^k/(k\theta)$ for $k = 1, 2, \dots$ and $\theta > 0$. This can be achieved by standard simulation methods for discrete distributions (see Ripley 1987, p. 71).

See Figure 7.12 for an example of data simulated from a four-dimensional Gumbel copula using this algorithm. Note the upper tail dependence in each bivariate margin of this copula.

While Archimedean copulas with completely monotonic generators (Laplace–Stieltjes transforms) can be used in any dimension, if we are interested in

Archimedean copulas in a given dimension d, we can relax the requirement of complete monotonicity and substitute the weaker requirement of *d-monotonicity*. See Section 15.2.1 for more details.

A copula obtained from construction (7.46) is obviously an *exchangeable* copula conforming to (7.20). While exchangeable bivariate Archimedean copulas are widely used in modelling applications, their exchangeable multivariate extensions represent a very specialized form of dependence structure and have more limited applications. An exception to this is in the area of credit risk, although even here more general models with group structures are also needed. It is certainly natural to enquire whether there are extensions to the Archimedean class that are not rigidly exchangeable. We present some non-exchangeable Archimedean copulas in Section 15.2.2.

Notes and Comments

The name *Archimedean* relates to an algebraic property of the copulas that resembles the Archimedean axiom for real numbers (see Nelsen 2006, p. 122). The Clayton copula was introduced in Clayton (1978), although it has also been called the Cook and Johnson copula (see Genest and MacKay 1986) and the Pareto copula (see Hutchinson and Lai 1990). For the Frank copula see Frank (1979); this copula has radial symmetry and is the only such Archimedean copula. A useful reference, particularly for bivariate Archimedean copulas, is Nelsen (2006).

Theorem 7.47 is a result of Schweizer and Sklar (1983) (see also Alsina, Frank and Schweizer 2006). The formula for Kendall's tau in the Archimedean family is due to Genest and MacKay (1986). The link between completely monotonic functions and generators which give Archimedean copulas of the form (7.46) is found in Kimberling (1974). See also Feller (1971) for more on the concept of complete monotonicity. For more on the important connection between Archimedean generators and Laplace transforms, see Joe (1997).

Proposition 7.51 and Algorithm 7.53 are essentially due to Marshall and Olkin (1988). See Frees and Valdez (1997), Schönbucher (2005) and Frey and McNeil (2003) for further discussion of this technique. A text on simulation techniques for copula families is Mai and Scherer (2012).

Other copula families we have not considered include the Marshall–Olkin copulas (Marshall and Olkin 1967a,b) and the extremal copulas in Tiit (1996). There is also a large literature on pair copulas and vine copulas; fundamental references include Bedford and Cooke (2001), Kurowicka and Cooke (2006), Aas et al. (2009) and Czado (2010).

7.5 Fitting Copulas to Data

We assume that we have data vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ with identical distribution function F, describing financial losses or financial risk-factor returns; we write $\boldsymbol{X}_t = (X_{t,1}, \dots, X_{t,d})'$ for an individual data vector and $\boldsymbol{X} = (X_1, \dots, X_d)'$ for a generic random vector with df F. We assume further that this df F has continuous margins $F_1, \dots, F_d$ and thus, by Sklar's Theorem, a unique representation $F(x) = C(F_1(x_1), \dots, F_d(x_d))$.

It is often very difficult, particularly in higher dimensions and in situations where we are dealing with skewed loss distributions or heterogeneous risk factors, to find a good multivariate model that describes both marginal behaviour and dependence structure effectively. For multivariate risk-factor return data of a similar kind, such as stock returns or exchange-rate returns, we have discussed useful overall models such as the GH family of Section 6.2.3, but even in these situations there can be value in separating the marginal-modelling and dependence-modelling issues and looking at each in more detail. The copula approach to multivariate models facilitates this approach and allows us to consider, for example, the issue of whether tail dependence appears to be present in our data.

This section is thus devoted to the problem of estimating the parameters $\boldsymbol{\theta}$ of a parametric copula $C_{\boldsymbol{\theta}}$. The main method we consider is maximum likelihood in Section 7.5.3. First we outline a simpler method-of-moments procedure using sample rank correlation estimates. This method has the advantage that marginal distributions do not need to be estimated, and consequently inference about the copula is in a sense "margin free".

7.5.1 Method-of-Moments Using Rank Correlation

Depending on which particular copula we want to fit, it may be easier to use empirical estimates of either Spearman's or Kendall's rank correlation to infer an estimate for the copula parameter. We begin by discussing the standard estimators of both of these rank correlations.

Proposition 7.34 suggests that we could estimate $\rho_S(X_i, X_j)$ by calculating the usual correlation coefficient for the *pseudo-observations*: $\{(F_{i,n}(X_{t,i}), F_{j,n}(X_{t,j})) : t = 1, \dots, n\}$, where $F_{i,n}$ denotes the standard empirical df for the ith margin. In fact, we estimate $\rho_S(X_i, X_j)$ by calculating the correlation coefficient of the *ranks* of the data, a quantity known as the Spearman's rank correlation coefficient. This coincides with the correlation coefficient of the pseudo-observations when there are no tied observations (that is, observations with $X_{t,i} = X_{s,i}$ or $X_{t,j} = X_{s,j}$ for some $t \neq s$).

The rank of $X_{t,i}$, written $\operatorname{rank}(X_{t,i})$, is simply the position of $X_{t,i}$ in the sample $X_{1,i}, \dots, X_{n,i}$ when the observations are ordered from smallest to largest. If there are tied observations, we assign them a rank equal to the average rank that the observations would have if the ties were randomly broken; for example, the sample of four observations $\{2, 3, 2, 1\}$ would have ranks $\{2.5, 4, 2.5, 1\}$.

In the case of no ties the Spearman's rank correlation coefficient is given by the formula

$$r_{ij}^{S} = \frac{12}{n(n^2-1)} \sum_{t=1}^{n} (\operatorname{rank}(X_{t,i}) - \tfrac{1}{2}(n+1))(\operatorname{rank}(X_{t,j}) - \tfrac{1}{2}(n+1)). \qquad (7.51)$$

We denote by $R^{\mathrm{S}} = (r^{\mathrm{S}}_{ij})$ the matrix of pairwise Spearman's rank correlation coefficients; since this is the sample correlation matrix of the vectors of ranks, it is clearly a positive-semidefinite matrix.

The standard estimator of Kendall's tau $\rho_\tau(X_i, X_j)$ is Kendall's rank correlation coefficient:

$$r^\tau_{ij} = \binom{n}{2}^{-1} \sum_{1 \leqslant t < s \leqslant n} \operatorname{sign}((X_{t,i} - X_{s,i})(X_{t,j} - X_{s,j})). \tag{7.52}$$

This is clearly the empirical analogue of the theoretical Kendall's tau in Definition 7.31. Note that the actual evaluation of this estimator for large n is time-consuming (in comparison with Spearman's rho) because every pair of observations must be considered. Again we can collect pairwise Kendall's rank correlation coefficients in a matrix $R^\tau = (r^\tau_{ij})$; by observing that this matrix may be written as

$$R^\tau = \binom{n}{2}^{-1} \sum_{1 \leqslant t < s \leqslant n} \operatorname{sign}(\boldsymbol{X}_t - \boldsymbol{X}_s) \operatorname{sign}(\boldsymbol{X}_t - \boldsymbol{X}_s)',$$

it is again apparent that this gives a positive-semidefinite matrix.

In a series of examples we show how these sample rank correlations can be used to calibrate (or partially calibrate) various copulas. Obviously we assume that there are a priori grounds for considering the chosen copula to be an appropriate model, such as symmetry or the lack of it and the presence or absence of tail dependence. The general method will always be similar: we look for a theoretical relationship between one of the rank correlations and the parameters of the copula and substitute empirical values of the rank correlation into this relationship to get estimates of some or all of the copula parameters.

Example 7.54 (bivariate Archimedean copulas with a single parameter). Suppose our assumed model is of the form $F(x_1, x_2) = C_\theta(F_1(x_1), F_2(x_2))$, where θ is a single parameter to be estimated. For many such copulas a simple functional relationship exists between either Kendall's tau and θ or Spearman's rho and θ. For specific examples consider the Gumbel, Clayton and Frank copulas of Section 7.4; in these cases we have simple relationships of the form $\rho_\tau(X_1, X_2) = f(\theta)$, as shown in Table 7.5. This suggests that we can calibrate these copulas by first calculating a sample value r^τ for Kendall's tau and then solving the equation $r^\tau = f(\hat{\theta})$ for $\hat{\theta}$, assuming that $\hat{\theta}$ is a valid value in the parameter space of the copula. For example, Gumbel's copula is calibrated by taking $\hat{\theta} = (1 - r^\tau)^{-1}$, provided that $r^\tau \geqslant 0$. Clayton's copula interpolates between perfect negative and perfect positive dependence and can be calibrated to any sample Kendall's tau value in $(-1, 1)$. For the calibration of higher-dimensional Archimedean copulas using rank correlations, see Hofert, Mächler and McNeil (2012).

Example 7.55 (calibrating Gauss copulas using Spearman's rho). Suppose we assume a meta-Gaussian model for $\boldsymbol{X}$ with copula C^{Ga}_P, and we wish to estimate the correlation matrix P. It follows from Theorem 7.42 that

$$\rho_{\mathrm{S}}(X_i, X_j) = (6/\pi) \arcsin \tfrac{1}{2}\rho_{ij} \approx \rho_{ij},$$

where the final approximation is very accurate (see Figure 7.10). This suggests we estimate P by the matrix of pairwise Spearman's rank correlation coefficients R^{S}.

The method of Example 7.55 could be used to estimate P in a t copula model $C^t_{\nu,P}(F_1(x_1), \dots, F_d(x_d))$, although the calibration would not be as accurate as in the Gaussian case. Empirical investigations of the relationship (7.41) based on the Monte Carlo approximation (7.42) suggest that the error $|\rho_{\mathrm{S}}(X_i, X_j) - \rho_{ij}|$, while still modest, is larger than in the Gaussian case and increases with decreasing degrees of freedom ν. Instead we propose a method based on Kendall's tau in the next example, which is based on Proposition 7.43 and could be applied to all elliptical copulas.

Example 7.56 (calibrating *t* copulas using Kendall's tau). Suppose we assume a meta-t model for $\boldsymbol{X}$ with copula $C^t_{\nu,P}$ and we wish to estimate the correlation matrix P. It follows from Proposition 7.43 that

$$\rho_\tau(X_i, X_j) = (2/\pi) \arcsin \rho_{ij},$$

so that a possible estimator of P is the matrix R^* with components given by $r^*_{ij} = \sin(\frac{1}{2}\pi r^\tau_{ij})$. However, there is no guarantee that this componentwise transformation of the matrix of Kendall's rank correlation coefficients will remain positive definite (although in our experience it very often does). In this case R^* can be transformed by the eigenvalue method given in Algorithm 7.57 to obtain a positive-definite matrix that is close to R^*. The remaining parameter ν of the copula could then be estimated by maximum likelihood, as discussed in Section 7.5.3.

Algorithm 7.57 (eigenvalue method). Let R^* be a so-called *pseudo*-correlation matrix, i.e. a symmetric matrix of pairwise correlation estimates with unit diagonal entries and off-diagonal entries in $[-1, 1]$ that is not positive semidefinite.

(1) Calculate the spectral decomposition $R^* = GLG'$ as in (6.59), where L is the matrix of eigenvalues and G is an orthogonal matrix whose columns are eigenvectors of R^*.

(2) Replace all negative eigenvalues in L by small values $\delta > 0$ to obtain $\tilde{L}$.

(3) Calculate $Q = G\tilde{L}G'$, which will be symmetric and positive definite but not a correlation matrix, since its diagonal elements will not necessarily equal one.

(4) Return the correlation matrix $R = \wp(Q)$, where $\wp$ denotes the correlation matrix operator defined in (6.5).

In Examples 7.55 and 7.56 we saw that it is relatively easy to calibrate the Gauss copula and the correlation parameter matrix P of the t copula to sample rank correlations. This technique is particularly useful when we have limited multivariate data and formal estimation of a full multivariate model is unrealistic. Consider the following hypothetical example.

Example 7.58 (fictitious risk integration situation). Suppose a company is divided into a number of business units that function semi-autonomously. The company management would like to calculate an enterprise-wide P&L distribution for a one-month period. They have historical data on monthly results for each of the business units for the last two years only, i.e. twenty-four observations. However, each business unit believes that through detailed knowledge of their own business going back over a longer period they can specify their own P&L fairly accurately. Rather than attempting to fit a multivariate distribution to twenty-four observations, the risk-management team decides to combine the individual marginal models provided by each of the business units using a matrix of rank correlations estimated from the twenty-four data points.

In this situation we can build multivariate models by combining the known marginal distributions using any copula that can be calibrated to the estimated rank correlations. The Gaussian and t copulas lend themselves to this purpose and can be used to build meta-Gaussian and meta-t models that are consistent with the available information.

Typically, these models could then be used in a Monte Carlo risk analysis; we have seen in Section 7.1.4 that meta-Gaussian and meta-t models are particularly easy to simulate. Because the approach is obviously prone to model risk (24 observations provide very meagre multivariate data) it should be seen as a form of sensitivity analysis performed using detailed marginal information and only vague dependence information; we might choose to compare a meta-Gaussian model with no tail dependence and a meta-t model with, say, three degrees of freedom and very strong tail dependence. In Section 8.4.4 we will have more to say on this problem of risk integration under dependence uncertainty.

7.5.2 *Forming a Pseudo-sample from the Copula*

We now turn to the estimation of parametric copulas by maximum likelihood (ML). In practical situations we are seldom interested in the copula alone, but also require estimates of the margins to form a full multivariate model; even when the copula is of central interest, as it is for us in this chapter, we are forced to estimate margins in order to estimate the copula, since copula data are almost never observed directly.

While we may attempt to estimate margins and copula in one single optimization, splitting the modelling into two steps can yield more insight and allow a more detailed analysis of the different model components. In this section we describe briefly some general approaches to the first step of estimating margins and constructing a *pseudo-sample* of observations from the copula. In the following section we describe how the copula parameters are estimated by ML from the pseudo-sample.

Let $\hat{F}_1, \dots, \hat{F}_d$ denote estimates of the marginal dfs (possible methods are discussed below). The pseudo-sample from the copula consists of the vectors $\hat{\boldsymbol{U}}_1, \dots, \hat{\boldsymbol{U}}_n$, where

$$\hat{\boldsymbol{U}}_t = (\hat{U}_{t,1}, \dots, \hat{U}_{t,d})' = (\hat{F}_1(X_{t,1}), \dots, \hat{F}_d(X_{t,d}))'. \tag{7.53}$$

Observe that, even if the original data vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ are iid, the pseudo-sample data are generally dependent, because the marginal estimates $\hat{F}_i$ will in most cases be constructed from all of the original data vectors through the univariate samples $X_{1,i}, \dots, X_{n,i}$. Possible methods for obtaining the marginal estimate $\hat{F}_i$ include the following.

(1) Parametric estimation. We choose an appropriate parametric model for the data in question and fit it by ML: for financial risk-factor return data we might consider the GH distribution, or one of its special cases such as Student t or normal inverse Gaussian (NIG); for insurance or operational loss data we might consider a standard actuarial loss distribution such as Pareto or lognormal.

(2) Non-parametric estimation with variant of empirical df. We could estimate F_j using

$$F^*_{i,n}(x) = \frac{1}{n+1}\sum_{t=1}^{n} I_{\{X_{t,i} \leqslant x\}}, \tag{7.54}$$

which differs from the usual empirical df by the use of the denominator $n+1$ rather than n. This guarantees that the pseudo-copula data in (7.53) lie strictly in the interior of the unit cube; to implement ML we must be able to evaluate the copula density at each $\hat{\boldsymbol{U}}_i$, and in many cases this density is infinite on the boundary of the cube.

(3) Extreme value theory for the tails. Empirical distribution functions are known to be poor estimators of the underlying distribution in the tails. An alternative is to use a technique from extreme value theory, described in Section 5.2.6, whereby the tails are modelled semi-parametrically using a generalized Pareto distribution (GPD); the body of the distribution may be modelled empirically.

Example 7.59. We analyse five years of daily log-return data (1996–2000) for Intel, Microsoft and General Electric stocks. The marginal distributions are estimated empirically (method (2)) and the pseudo-sample from the copula is shown in Figure 7.13. Essentially, the points are plotted at the coordinates $(\text{rank}(X_{t,i})/(n+1), \text{rank}(X_{t,j})/(n+1))$, where $\text{rank}(X_{t,i})$ denotes the rank of $X_{t,i}$ in the sample $X_{1,i}, \dots, X_{n,i}$.

7.5.3 Maximum Likelihood Estimation

Let $C_{\boldsymbol{\theta}}$ denote a parametric copula, where $\boldsymbol{\theta}$ is the vector of parameters to be estimated. The MLE is obtained by maximizing

$$\ln L(\boldsymbol{\theta}; \hat{\boldsymbol{U}}_1, \dots, \hat{\boldsymbol{U}}_n) = \sum_{t=1}^{n} \ln c_{\boldsymbol{\theta}}(\hat{\boldsymbol{U}}_t) \tag{7.55}$$

with respect to $\boldsymbol{\theta}$, where $c_{\boldsymbol{\theta}}$ denotes the copula density as in (7.18) and $\hat{\boldsymbol{U}}_t$ denotes a pseudo-observation from the copula.

Obviously, the statistical quality of the estimates of the copula parameters depends very much on the quality of the estimates of the marginal distributions used in

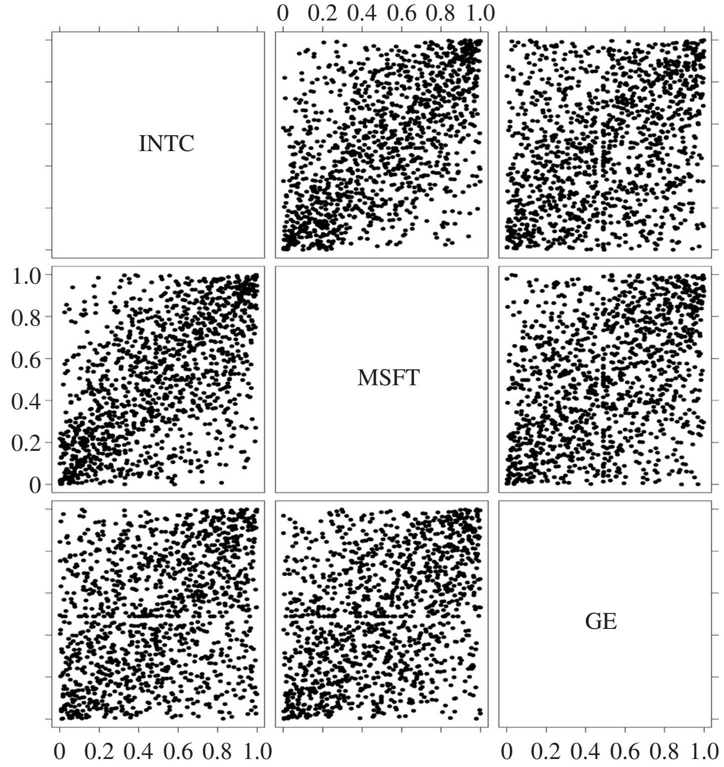

Figure 7.13. Pairwise scatterplots of a pseudo-sample from a copula for trivariate Intel, Microsoft and General Electric log-returns (see Example 7.59).

the formation of the pseudo-sample from the copula. The properties of estimates derived using the marginal estimation methods (1) and (2) in Section 7.5.2 have both been studied in more theoretical detail. When margins are estimated parametrically (method (1)), inference about the copula using (7.55) amounts to what has been termed the inference functions for margins (IFM) approach by Joe (1997). When margins are estimated non-parametrically (method (2)), the estimates of the copula parameters may be regarded as semi-parametric and the approach has been labelled pseudo-maximum likelihood by Genest and Rivest (1993) (see Notes and Comments for more references). One could envisage using the two-stage method to decide on the most appropriate copula family and then estimating all parameters (marginal and copula) in a final fully parametric round of estimation.

In practice, to implement the ML method we need to derive the copula density. This is straightforward, if tedious, for the exchangeable Archimedean copulas of Section 7.4, and these have been popular models in bivariate and trivariate applications to insurance loss data. For implicit copulas like the Gaussian and t copulas we use (7.19). The MLE is generally found by numerical maximization of the resulting log-likelihood (7.55).

Example 7.60 (fitting the Gauss copula). In the case of a Gauss copula we use (7.19) to see that the log-likelihood (7.55) becomes

$$\ln L(P; \hat{\boldsymbol{U}}_1, \dots, \hat{\boldsymbol{U}}_n) = \sum_{t=1}^{n} \ln f_P(\Phi^{-1}(\hat{U}_{t,1}), \dots, \Phi^{-1}(\hat{U}_{t,d})) - \sum_{t=1}^{n}\sum_{j=1}^{d} \ln \phi(\Phi^{-1}(\hat{U}_{t,j})),$$

where f_Σ will be used to denote the joint density of a random vector with $N_d(\mathbf{0}, \Sigma)$ distribution. It is clear that the second term is not relevant in the maximization with respect to P, and the MLE is given by

$$\hat{P} = \underset{\Sigma \in \mathcal{P}}{\arg\max} \sum_{t=1}^{n} \ln f_\Sigma(\boldsymbol{Y}_t), \tag{7.56}$$

where $Y_{t,j} = \Phi^{-1}(\hat{U}_{t,j})$ for $j = 1, \dots, d$ and $\mathcal{P}$ denotes the set of all possible linear correlation matrices. To perform this maximization in practice, note that the set $\mathcal{P}$ can be constructed as

$$\mathcal{P} = \{P = \wp(Q) \colon Q = AA',\ A \text{ lower triangular with ones on the diagonal}\},$$

where $\wp$ is defined in (6.5). In other words, we can search over the set of unrestricted lower-triangular matrices with ones on the diagonal. This search is feasible in low dimensions but very slow in high dimensions, since the number of parameters is $O(d^2)$.

An approximate solution to the maximization may be obtained easily as follows. Suppose that instead of maximizing over $\mathcal{P}$ as in (7.56) we maximize over the set of all covariance matrices. This maximization problem has the analytical solution $\hat{\Sigma} = (1/n)\sum_{t=1}^{n} \boldsymbol{Y}_t \boldsymbol{Y}_t'$, which is the MLE of the covariance matrix Σ for iid normal data with $N_d(\mathbf{0}, \Sigma)$ distribution. In practice, $\hat{\Sigma}$ is likely to be *close* to being a correlation matrix. As an approximate solution to the original problem we could take the correlation matrix $\tilde{P} = \wp(\hat{\Sigma})$.

When a Gauss copula is fitted to the trivariate data in Example 7.59 by full ML, the estimated correlation matrix has entries 0.58 (INTC-MSFT), 0.34 (INTC-GE) and 0.40 (MSFT-GE); the value of the log-likelihood at the maximum is 376.65. Using the alternative method gives estimates that are identical to two significant figures and that yield a log-likelihood value of 376.62.

A further alternative would be to use the estimation procedure in Example 7.55, based on Spearman's rank correlations. Using the Spearman method we get, respectively, 0.57, 0.34 and 0.40 for the parameter estimates; the value of the log-likelihood at this value of P is 376.50, which is also not so far from the maximum.

Example 7.61 (fitting the *t* copula). In the case of the t copula, (7.19) implies that the log-likelihood (7.55) is

$$\ln L(\nu, P; \hat{\boldsymbol{U}}_1, \dots, \hat{\boldsymbol{U}}_n) = \sum_{t=1}^{n} \ln g_{\nu,P}(t_\nu^{-1}(\hat{U}_{t,1}), \dots, t_\nu^{-1}(\hat{U}_{t,d})) - \sum_{t=1}^{n}\sum_{j=1}^{d} \ln g_\nu(t_\nu^{-1}(\hat{U}_{t,j})),$$

where $g_{\nu,P}$ denotes the joint density of a random vector with $t_d(\nu, \mathbf{0}, P)$ distribution, P is a linear correlation matrix, g_ν is the density of a univariate $t_1(\nu, 0, 1)$ distribution, and t_ν^{-1} is the corresponding quantile function.

Again, in relatively low dimensions we could search over the set of correlation matrices P and degrees of freedom parameter ν for a global maximum. For higher-dimensional work it would be easier to estimate P using Kendall's tau estimates, as in Example 7.56, and to estimate the single parameter ν by maximum likelihood.

When a t copula is fitted to the trivariate data in Example 7.59 by full ML, the estimated matrix P has entries 0.59 (INTC-MSFT), 0.36 (INTC-GE) and 0.42 (MSFT-GE); the estimate of ν is 6.5 and the value of the log-likelihood at the maximum is 420.39. Using the simpler method based on Kendall's tau gives identical parameter estimates to two significant figures and a log-likelihood value of 420.32. Clearly, the t model fits much better than a Gauss copula model; the log-likelihood is increased by over 40. This would be massively significant in a likelihood ratio test (although, strictly speaking, such a test introduces a technical difficulty, since the Gauss copula represents a boundary case of the t copula model ($\nu = \infty$), which violates standard regularity conditions (see Notes and Comments)).

Following standard statistical practice we usually fit a number of copula models to data and compare the quality of the fitted models using tools like the Akaike information criterion (see Appendix A.3.6). We may also carry out goodness-of-fit tests to assess the plausibility that the data come from any given copula. Most of the goodness-of-fit tests that have been suggested for copulas are quite computationally intensive and are limited to applications in relatively low dimensions (see Notes and Comments for some references).

Notes and Comments

The copula estimation procedure based on empirical values of Kendall's tau is discussed in detail for bivariate Archimedean copulas by Genest and Rivest (1993); they explain why the procedure may be considered to be a method-of-moments technique and show how confidence intervals for the copula parameter (in the case of single-parameter copulas) may be derived.

The method of calibrating the Gauss copula with Spearman's rank correlation in Example 7.55 is essentially due to Iman and Conover (1982). The use of this calibration method to build meta-Gaussian models with prescribed margins and the Monte Carlo simulation of data from these models are implemented in the @RISK software (Palisade 1997), which is widely used in insurance. Our Example 7.56 is intended to show that this approach can be extended to meta-t models, which may well be more interesting due to their tail dependence.

The eigenvalue method for correcting the positive definiteness of correlation matrices given in Algorithm 7.57 is described by Rousseeuw and Molenberghs (1993). An empirical comparison of the eigenvalue method with different approaches to this problem, including so-called shrinkage methods, is found in Lindskog (2000).

The inference functions for margins (IFM) approach to the estimation of copulas (method (1) of Section 7.5.2 followed by maximization of (7.55)) is described by

Joe (1997), who gives asymptotic theory; the name of the approach (IFM) follows terminology of McLeish and Small (1988).

The pseudo-likelihood approach to copula estimation (method (2) of Section 7.5.2 followed by maximization of (7.55)) is described in Genest and Rivest (1993), and the consistency and asymptotic normality of the resulting parameter estimates are demonstrated. In Monte Carlo simulations it is found that this method outperforms the Kendall's tau method for a bivariate Clayton copula (see also Genest, Ghoudi and Rivest 1995).

Frees and Valdez (1997) discuss the relevance of copulas in actuarial applications and give an example where copulas are fitted to data using the Kendall's tau method and the IFM method. Also in an insurance context, Klugman and Parsa (1999) discuss ML inference for copulas and bivariate goodness-of-fit tests, while Chen and Fan (2005) describe a likelihood-ratio test for semi-parametric copula selection.

The fitting of the t copula to data and statistical aspects of testing this copula against the Gauss copula are discussed at length in Mashal and Zeevi (2002); the technical problem that the Gauss copula is a boundary case of the t copula is addressed in this paper and a correction is suggested. The authors provide a number of financial examples suggesting that extremal dependence is a feature of financial data. Breymann, Dias and Embrechts (2003) fit various bivariate copulas to high-frequency financial return data at different timescales and provide extensive comparisons with respect to goodness-of-fit.

There is a growing literature on goodness-of-fit tests for copulas: see the survey article by Genest, Rémillard and Beaudoin (2009). For an attractive graphical test, see Hofert and Mächler (2014).

Papers that develop dynamic time-series models for financial return data using copulas include Chen and Fan (2006), Patton (2004, 2006) and Fortin and Kuzmics (2002). A change-point problem for copulas within econometrics is discussed in Dias and Embrechts (2009).

8
Aggregate Risk

This chapter is devoted to a number of theoretical concepts in quantitative risk management that fall under the broad heading of aggregate risk and integrated risk management. We understand aggregate risk as the risk of a portfolio, which could even be the entire position in risky assets of a financial institution. The material builds on general ideas in risk measurement discussed in Section 2.3 and also uses in certain places the copula theory of Chapter 7 and some facts about elliptical distributions from Section 6.3.

In Sections 8.1–8.3 we treat the issue of measuring aggregate risk. We begin with general results and we discuss, in particular, the dual representation of convex risk measures as generalized scenarios (a mathematical extension of the idea of a stress test). Next we consider certain law-invariant risk measures (risk measures that depend only on the loss distribution). Finally, we apply the representation result to the case of linear portfolios and we discuss risk measurement for the special case of portfolios that are linear combinations of elliptically distributed risk factors.

Section 8.4 is concerned with risk aggregation: we assume that risk capital numbers for sub-units of an enterprise have been computed and we discuss methods for aggregating these risk capital numbers into a capital requirement for the entire enterprise. Moreover, we consider the problem of bounding an aggregate risk if we know something about the individual risks that contribute to the whole but have only limited information about their dependence. We discuss specific difficulties that arise when risk is measured with a non-subadditive risk measure like VaR. Finally, in Section 8.5, we treat the subject of allocating risk capital for an aggregate risk back to the individual risks in the portfolio. This issue is relevant for the purposes of performance measurement, loan pricing and capital budgeting.

8.1 Coherent and Convex Risk Measures

In this section we present elements of the modern theory of risk measures. Our exposition is a simplified account of material found in Föllmer and Schied (2011). We begin by recalling the axioms characterizing coherent and convex risk measures. For the economic motivation of these axioms we refer to Section 2.3.5.

Consider a probability space $(\Omega, \mathcal{F}, P)$ and a linear space $\mathcal{M} \subset \mathcal{L}^0(\Omega, \mathcal{F}, P)$, where $\mathcal{L}^0(\Omega, \mathcal{F}, P)$ denotes the set of all random variables on $(\Omega, \mathcal{F}, P)$ that are almost surely (a.s.) finite. Each $L \in \mathcal{M}$ represents the loss incurred on a financial position over some fixed time horizon. We assume throughout that constant random variables belong to $\mathcal{M}$ and denote them by lowercase letters. In this context a *risk measure* is a mapping $\varrho\colon \mathcal{M} \to \mathbb{R}$ with the interpretation that $\varrho(L)$ gives the total amount of capital that is needed to back a position with loss L. The axioms for ϱ are as follows.

Monotonicity. $L_1 \leqslant L_2 \Rightarrow \varrho(L_1) \leqslant \varrho(L_2)$.

Translation invariance. For $m \in \mathbb{R}$, $\varrho(L + m) = \varrho(L) + m$.

Subadditivity. For $L_1, L_2 \in \mathcal{M}$, $\varrho(L_1 + L_2) \leqslant \varrho(L_1) + \varrho(L_2)$.

Positive homogeneity. For $\lambda \geqslant 0$, $\varrho(\lambda L) = \lambda\varrho(L)$.

Convexity. For $0 \leqslant \gamma \leqslant 1$, $L_1, L_2 \in \mathcal{M}$,

$$\varrho(\gamma L_1 + (1-\gamma)L_2) \leqslant \gamma\varrho(L_1) + (1-\gamma)\varrho(L_2).$$

Definition 8.1. A risk measure that satisfies the monotonicity, translation invariance and convexity axioms is called a *convex* measure of risk; a risk measure that satisfies the monotonicity, translation invariance, subadditivity and positive homogeneity axioms is called a *coherent* measure of risk.

A coherent risk measure is automatically convex; the converse implication is not true, as will be seen below. On the other hand, for a positive-homogeneous risk measure, convexity and coherence are equivalent.

8.1.1 Risk Measures and Acceptance Sets

There is an important relationship between risk measures and so-called *acceptance sets*. For a given risk measure the associated acceptance set contains the positions that are acceptable without any backing capital.

Definition 8.2. For a monotone and translation-invariant risk measure ϱ, the associated acceptance set of ϱ is the set

$$A_\varrho = \{L \in \mathcal{M}\colon \varrho(L) \leqslant 0\}. \tag{8.1}$$

Proposition 8.3. *For a monotone and translation-invariant risk measure ϱ with associated acceptance set A_ϱ, the following statements hold.*

(1) *A_ϱ is nonempty and satisfies the condition*

$$L \in A_\varrho \text{ and } \tilde{L} \leqslant L \Rightarrow \tilde{L} \in A_\varrho. \tag{8.2}$$

(2) *ϱ can be reconstructed from A_ϱ via*

$$\varrho(L) = \inf\{m \in \mathbb{R}\colon L - m \in A_\varrho\}. \tag{8.3}$$

Proof. Statement (1) is obvious. For (2) note that

$$\inf\{m\colon L - m \in A_\varrho\} = \inf\{m\varrho(L-m) \leqslant 0\} = \inf\{m\colon \varrho(L) - m \leqslant 0\},$$

and this is obviously equal to $\varrho(L)$. □

Conversely, it is sometimes useful to start with a set $A \subset \mathcal{M}$ of acceptable positions and to define an associated risk measure ϱ_A using (8.3). The properties of such a risk measure are given in the following proposition.

Proposition 8.4. *Suppose that the set A satisfies (8.2) and define ϱ_A by*

$$\varrho_A(L) = \inf\{m \in \mathbb{R} \colon L - m \in A\}. \tag{8.4}$$

Suppose, moreover, that $\varrho_A(L)$ defined in this way is finite for all $L \in \mathcal{M}$. Then ϱ_A is a monotone and translation-invariant risk measure on $\mathcal{M}$. The associated acceptance A_{ϱ_A} satisfies $A_{\varrho_A} \supseteq A$.

Proof. These properties of ϱ_A are easily checked. □

Remark 8.5. It is natural to enquire when the sets A and A_{ϱ_A} in Proposition 8.4 are equal. One result in that direction is given in Section 4.1 of Föllmer and Schied (2011) for the case where $\mathcal{M}$ contains only bounded random variables: in that case, $A = A_{\varrho_A}$ if and only if the set A is closed in the supremum norm.

In the next proposition we require some further basic ideas from convex analysis. A set $C \subset \mathcal{M}$ is said to be convex if $(1-\gamma)x + \gamma y \in C$ whenever $x \in C$, $y \in C$ and $0 < \gamma < 1$. A convex set is a convex cone if it has the additional property that it is closed under positive scalar multiplication, i.e. $\lambda x \in C$ when $x \in C$ and $\lambda > 0$.

Proposition 8.6.

(a) *Consider a monotone and translation-invariant risk measure ϱ with associated acceptance set A_ϱ defined by (8.1). Then*

(a1) *ϱ is a convex risk measure if and only if A_ϱ is convex, and*

(a2) *ϱ is coherent if and only if A_ϱ is a convex cone.*

(b) *More generally, consider a set of acceptable positions A and the associated risk measure ϱ_A defined by (8.4) (whose acceptance set may be larger than A). Then ϱ_A is a convex risk measure if A is convex and ϱ_A is coherent if A is a convex cone.*

Proof. For part (a1) it is clear that A_ϱ is convex if ϱ is convex. For the converse direction, consider arbitrary $L_1, L_2 \in \mathcal{M}$ and $0 \leqslant \gamma \leqslant 1$. Now for $i = 1, 2$, $L_i - \varrho(L_i) \in A_\varrho$ by definition of A_ϱ. Since A_ϱ is convex, we also have that $\gamma(L_1 - \varrho(L_1)) + (1-\gamma)(L_2 - \varrho(L_2)) \in A_\varrho$. By the definition of A_ϱ and translation invariance we have

$$\begin{aligned} 0 &\geqslant \varrho(\gamma(L_1 - \varrho(L_1)) + (1-\gamma)(L_2 - \varrho(L_2))) \\ &= \varrho(\gamma L_1 + (1-\gamma)L_2) - (\gamma\varrho(L_1) + (1-\gamma)\varrho(L_2)), \end{aligned}$$

which implies the convexity of ϱ.

To prove (a2) assume that $L \in A_\varrho$. As ϱ is positive homogeneous, $\varrho(\lambda L) = \lambda\varrho(L) \leqslant 0$ and hence $\lambda L \in A_\varrho$. Conversely, for $L \in \mathcal{M}$ we have $L - \varrho(L) \in A_\varrho$ and, as A_ϱ is a convex cone, also $\lambda(L - \varrho(L)) \in A_\varrho$ for all λ. Hence, by translation invariance,

$$0 \geqslant \varrho(\lambda L - \lambda\varrho(L)) = \varrho(\lambda L) - \lambda\varrho(L). \tag{8.5}$$

For the opposite inequality note that for $m < \varrho(L)$, $L - m \notin A_\varrho$ and hence also $\lambda(L - m) \notin A_\varrho$ for all $\lambda > 0$. Hence

$$0 < \varrho(\lambda L - \lambda m) = \varrho(\lambda L) - \lambda m,$$

i.e. $\varrho(\lambda L) > \lambda m$. By taking the supremum we get $\varrho(\lambda L) \geqslant \sup\{\lambda m \colon m < \varrho(L)\} = \lambda\varrho(L)$; together with (8.5) the claim follows.

The proof of (b) uses similar arguments to parts (a1) and (a2) and is omitted. □

We now give a number of examples of risk measures and acceptance sets.

Example 8.7 (value-at-risk). Given a confidence level $\alpha \in (0, 1)$, suppose we call an rv $L \in \mathcal{M}$ acceptable if $P(L > 0) \leqslant 1 - \alpha$. The associated risk measure defined by (8.4) is given by

$$\varrho_\alpha(L) := \inf\{m \in \mathbb{R} \colon P(L - m > 0) \leqslant 1 - \alpha\} = \inf\{m \in \mathbb{R} \colon P(L \leqslant m) \geqslant \alpha\},$$

which is the VaR at confidence level α.

Example 8.8 (risk measures based on loss functions). Consider a function $\ell \colon \mathbb{R} \to \mathbb{R}$ that is strictly increasing and convex and some threshold $c \in \mathbb{R}$. Assume that $E(\ell(L))$ is finite for all $L \in \mathcal{M}$ and define an acceptance set by

$$A = \{L \in \mathcal{M} \colon E(\ell(L)) \leqslant \ell(c)\}$$

and the associated risk measure by

$$\varrho_A = \inf\{m \in \mathbb{R} \colon E(\ell(L - m)) \leqslant \ell(c)\}.$$

In this context ℓ is called a *loss function* because the convexity of ℓ serves to penalize large losses; a loss function can be derived from a utility function u (a strictly increasing and concave function) by setting $\ell(x) = -u(-x)$.

The set A obviously satisfies (8.2), so that ϱ_A is translation invariant and monotone by Proposition 8.4. Furthermore, A is convex. This can be seen by considering acceptable positions L_1 and L_2 and observing that the convexity of ℓ implies

$$\begin{aligned} E(\ell(\gamma L_1 + (1-\gamma)L_2)) &\leqslant E(\gamma\ell(L_1) + (1-\gamma)\ell(L_2)) \\ &\leqslant \gamma\ell(c) + (1-\gamma)\ell(c) \\ &= \ell(c), \end{aligned}$$

where we have used the fact that $E(\ell(L_i)) \leqslant \ell(c)$ for acceptable positions. Hence $\gamma L_1 + (1-\gamma)L_2 \in A$ as required, and ϱ_A is a convex measure of risk by Proposition 8.6.

As a specific example we take $\ell(x) = \mathrm{e}^{\alpha x}$ for some $\alpha > 0$. We get

$$\begin{aligned}\varrho_{\alpha,c}(L) &:= \inf\{m\colon E(\mathrm{e}^{\alpha(L-m)}) \leqslant \mathrm{e}^{\alpha c}\} = \inf\{m\colon E(\mathrm{e}^{\alpha L}) \leqslant \mathrm{e}^{\alpha c+\alpha m}\} \\ &= \frac{1}{\alpha}\ln\{E(\mathrm{e}^{\alpha L})\} - c.\end{aligned}$$

Note that $\varrho_{\alpha,c}(0) = -c$, which could not be true for a coherent risk measure. Consider the special case where $c = 0$ and write $\varrho_\alpha := \varrho_{\alpha,0}$. We find that for $\lambda > 1$,

$$\varrho_\alpha(\lambda L) = \frac{1}{\alpha}\ln\{E(\mathrm{e}^{\alpha\lambda L})\} \geqslant \frac{1}{\alpha}\ln\{E(\mathrm{e}^{\alpha L})^\lambda\} = \lambda\varrho_\alpha(L),$$

where the inequality is strict if the rv L is non-degenerate. This shows that ϱ_α is convex but not coherent. In insurance mathematics this risk measure is known as the exponential premium principle if the losses L are interpreted as claims.

Example 8.9 (stress-test or worst-case risk measure). Given a set of stress scenarios $S \subset \Omega$, a definition of a stress-test risk measure is

$$\varrho(L) = \sup\{L(\omega)\colon \omega \in S\},$$

i.e. the worst loss when we restrict our attention to those elements of the sample space Ω that belong to S. The associated acceptance set is $A_\varrho = \{L\colon L(\omega) \leqslant 0 \text{ for all } \omega \in S\}$, i.e. the losses that are non-positive for all stress scenarios. The crucial part of defining the risk measure is the choice of the scenario set S, which is often guided by probabilistic considerations involving the underlying measure P.

Example 8.10 (generalized scenarios). Consider a set $\mathcal{Q}$ of probability measures on $(\Omega, \mathcal{F})$ and a mapping $\gamma\colon \mathcal{Q} \to \mathbb{R}$ such that $\inf\{\gamma(Q)\colon Q \in \mathcal{Q}\} > -\infty$. Suppose that $\sup_{Q\in\mathcal{Q}} E_Q(|L|) < \infty$ for all $L \in \mathcal{M}$. Define a risk measure ϱ by

$$\varrho(L) = \sup\{E_Q(L) - \gamma(Q)\colon Q \in \mathcal{Q}\}. \tag{8.6}$$

Note that a measure $Q \in \mathcal{Q}$ such that $\gamma(Q)$ is large is penalized in the maximization in (8.6), so γ can be interpreted as a penalty function specifying the relevance of the various measures in $\mathcal{Q}$. The corresponding acceptance set is given by

$$A_\varrho = \{L \in \mathcal{M}\colon \sup\{E_Q(L) - \gamma(Q)\colon Q \in \mathcal{Q}\} \leqslant 0\}.$$

A_ϱ is obviously convex so that ϱ is a convex risk measure. In fact, every convex risk measure can be represented in the form (8.6), as will be shown in Theorem 8.11 below (at least for the case of finite Ω). In the case where $\gamma(\cdot) \equiv 0$ on $\mathcal{Q}$, ϱ is obviously positive homogeneous and therefore coherent.

The stress-test risk measure of Example 8.9 is a special case of (8.6) in which the penalty function is equal to zero and in which $\mathcal{Q}$ is the set of all Dirac measures $\delta_\omega(\cdot)$, $\omega \in S$, i.e. measures such that $\delta_\omega(B) = I_B(\omega)$ for arbitrary measurable sets $B \subset \Omega$. Since in (8.6) we may choose more general sets of probability measures than simply Dirac measures, risk measures of the form (8.6) are frequently referred to as generalized scenario risk measures.

8.1.2 Dual Representation of Convex Measures of Risk

In this section we show that convex measures of risk have dual representations as generalized scenario risk measures. We state and prove a theorem in the simpler setting of a finite probability space. However, the result can be extended to general probability spaces by imposing additional continuity conditions on the risk measure. See, for instance, Sections 4.2 and 4.3 of Föllmer and Schied (2011).

Theorem 8.11 (dual representation for risk measures). *Suppose that Ω is a finite probability space with $|\Omega| = n < \infty$. Let $\mathcal{F} = \mathcal{P}(\Omega)$, the set of all subsets of Ω, and take $\mathcal{M} := \{L \colon \Omega \to \mathbb{R}\}$. Then the following hold.*

(1) *Every convex risk measure ϱ on $\mathcal{M}$ can be written in the form*

$$\varrho(L) = \max\{E_Q(L) - \alpha_{\min}(Q) \colon Q \in \mathcal{S}^1(\Omega, \mathcal{F})\}, \tag{8.7}$$

where $\mathcal{S}^1(\Omega, \mathcal{F})$ denotes the set of all probability measures on Ω, and where the penalty function $\alpha_{\min}$ is given by

$$\alpha_{\min}(Q) = \sup\{E_Q(L) \colon L \in A_\varrho\}. \tag{8.8}$$

(2) *If ϱ is coherent, it has the representation*

$$\varrho(L) = \max\{E_Q(L) \colon Q \in \mathcal{Q}\} \tag{8.9}$$

for some set $\mathcal{Q} = \mathcal{Q}(\varrho) \subset \mathcal{S}^1(\Omega, \mathcal{F})$.

Proof. The proof is divided into three steps. For simplicity we write $\mathcal{S}^1$ for $\mathcal{S}^1(\Omega, \mathcal{F})$.

Step 1. First we show that for $L \in \mathcal{M}$,

$$\varrho(L) \geqslant \sup\{E_Q(L) - \alpha_{\min}(Q) \colon Q \in \mathcal{S}^1\}. \tag{8.10}$$

To establish (8.10), set $L' := L - \varrho(L)$ and note that $L' \in A_\varrho$ so that

$$\alpha_{\min}(Q) = \sup\{E_Q(L) \colon L \in A_\varrho\} \geqslant E_Q(L') = E_Q(L) - \varrho(L).$$

This gives $\varrho(L) \geqslant E_Q(L) - \alpha_{\min}(Q)$, and taking the supremum over different measures Q gives (8.10).

Step 2. The main step of the proof is now to construct for $L \in \mathcal{M}$ a measure $Q_L \in \mathcal{S}^1$ such that $\varrho(L) \leqslant E_{Q_L}(L) - \alpha_{\min}(Q_L)$; together with (8.10) this establishes (8.7) and (8.8). This is the most technical part of the argument and we give a simple illustration in Example 8.12.

By translation invariance it is enough to construct Q_L for a loss L with $\varrho(L) = 0$; moreover, we may assume without loss of generality that $\varrho(0) = 0$. Since $|\Omega| = n < \infty$ we may identify L with the vector of possible outcomes $\boldsymbol{\ell} = (L(\omega_1), \dots, L(\omega_n))'$ in $\mathbb{R}^n$. Similarly, a probability measure $Q \in \mathcal{S}^1$ can be identified with the vector of corresponding probabilities $\boldsymbol{q} = (q(\omega_1), \dots, q(\omega_n))'$, which is an element of the unit simplex in $\mathbb{R}^n$. We may identify A_ϱ with a convex subset $\mathcal{A}_\varrho$ of $\mathbb{R}^n$. Note that a loss L with $\varrho(L) = 0$ is not in the interior of A_ϱ. Otherwise $L + \epsilon$ would belong to A_ϱ for $\epsilon > 0$ small enough, which would imply that $0 \geqslant \varrho(L + \epsilon) = \varrho(L) + \epsilon = \epsilon$, which is a contradiction. Hence $\boldsymbol{\ell}$ is not in the interior of $\mathcal{A}_\varrho$. According to the supporting hyperplane theorem (see Proposition A.8 in Appendix A.1.5) there therefore exists a vector $\boldsymbol{u} \in \mathbb{R}^n \setminus \{0\}$ such that

$$\boldsymbol{u}'\boldsymbol{\ell} \geqslant \sup\{\boldsymbol{u}'\boldsymbol{x} : \boldsymbol{x} \in \mathcal{A}_\varrho\}. \tag{8.11}$$

Below we will construct Q_L from the vector $\boldsymbol{u}$. First, we show that $u_j \geqslant 0$ for all $1 \leqslant j \leqslant n$. Since $\varrho(0) = 0$, by monotonicity and translation invariance we have $\varrho(L - 1 - \lambda I_{\{\omega_j\}}) < 0$ for all j and all $\lambda > 0$. This shows that $L - 1 - \lambda I_{\{\omega_j\}} \in A_\varrho$ or, equivalently, $\boldsymbol{\ell} - \mathbf{1} - \lambda \boldsymbol{e}_j \in \mathcal{A}_\varrho$, where we use the notation $\mathbf{1} = (1, \dots, 1)'$. Applying relation (8.11) we get that $\boldsymbol{u}'\boldsymbol{\ell} \geqslant \boldsymbol{u}'(\boldsymbol{\ell} - \mathbf{1} - \lambda \boldsymbol{e}_j)$ for all $\lambda > 0$, which implies that

$$0 \geqslant -\sum_{j=1}^{n} u_j - \lambda u_j, \quad \forall \lambda > 0.$$

This is possible only for $u_j \geqslant 0$. Since $\boldsymbol{u}$ is different from zero, as least one of the components must be strictly positive, and we can define the vector $\boldsymbol{q} := \boldsymbol{u}/(\sum_{j=1}^{n} u_j)$. Note that $\boldsymbol{q}$ belongs to the unit simplex in $\mathbb{R}^n$, and we define Q_L to be the associated probability measure. It remains to verify that Q_L satisfies the inequality $\varrho(L) \leqslant E_{Q_L}(L) - \alpha_{\min}(Q_L)$; since we assumed $\varrho(L) = 0$, we need to show that

$$E_{Q_L}(L) \geqslant \alpha_{\min}(Q). \tag{8.12}$$

We have $\alpha_{\min}(Q_L) = \sup\{E_{Q_L}(X) : X \in A_\varrho\} = \sup\{\boldsymbol{q}'\boldsymbol{x} : \boldsymbol{x} \in \mathcal{A}_\varrho\}$. It follows from (8.11) that

$$E_{Q_L}(L) = \boldsymbol{q}'\boldsymbol{\ell} \geqslant \sup\{\boldsymbol{q}'\boldsymbol{x} : \boldsymbol{x} \in \mathcal{A}_\varrho\} = \alpha_{\min}(Q_L),$$

and hence we obtain the desired relation (8.12).

Step 3. In order to establish the representation (8.9) for coherent risk measures, we recall that for a coherent risk measure ϱ the acceptance set A_ϱ is a convex cone. Hence for $\lambda > 0$ we obtain

$$\alpha_{\min}(Q) = \sup_{L \in A_\varrho} E_Q(L) = \sup_{\lambda L \in A_\varrho} E_Q(\lambda L) = \lambda \alpha_{\min}(Q),$$

which is possible only for $\alpha_{\min}(Q) \in \{0, \infty\}$. The representation (8.9) follows if we set $\mathcal{Q}(\varrho) := \{Q \in \mathcal{S}^1 : \alpha_{\min}(Q) = 0\}$. □

Example 8.12. To illustrate the construction in step (2) of the proof of Theorem 8.11 we give a simple example (see Figure 8.1 to visualize the construction).

Consider the case where $d = 2$ and where the risk measure is $\varrho(L) = \ln E(\mathrm{e}^L)$, the exponential premium principle of Example 8.8 with $\alpha = 1$ and $c = 0$. Assume that the probability measure P is given by $\boldsymbol{p} = (0.4, 0.6)$, in which case the losses L

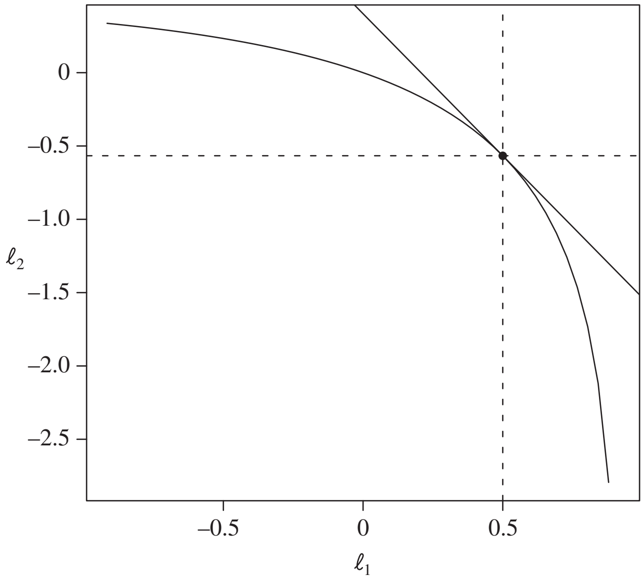

Figure 8.1. Illustration of the measure construction in step (2) of the proof of Theorem 8.11 (see Example 8.12 for details).

for which $\varrho(L) = 0$ are represented by the curve with equation $f_\varrho(\boldsymbol{\ell}) := \ln(0.4e^{\ell_1} + 0.6e^{\ell_2}) = 0$; the acceptance set A_ϱ consists of the curve and the region below it. We now construct Q_L for the loss L on the curve with $L(\omega_1) = \ell_1 = 0.5$ and $L(\omega_2) = \ell_2 \approx -0.566$. A possible choice for the vector $\boldsymbol{u}$ in (8.11) (the normal vector of the supporting hyperplane) is to take

$$\boldsymbol{u} = \nabla f_\rho(\boldsymbol{\ell}) = \left(\frac{\partial f_\varrho(\boldsymbol{\ell})}{\partial \ell_1}, \frac{\partial f_\varrho(\boldsymbol{\ell})}{\partial \ell_2}\right)'.$$

By normalization of $\boldsymbol{u}$ the measure Q_L may be identified with $\boldsymbol{q} \approx (0.659, 0.341)$, and for this measure the penalty is $\alpha_{\min}(Q_L) = \boldsymbol{q}'\boldsymbol{\ell} \approx 0.137$.

Properties of $\alpha_{\min}$. Next we discuss properties of the penalty function $\alpha_{\min}$. First we explain that $\alpha_{\min}$ is a minimal penalty function representing ϱ. Suppose that $\alpha\colon \mathcal{S}^1(\Omega, \mathcal{F}) \to \mathbb{R}$ is any function such that (8.7) holds for all $L \in M$. Then for $Q \in \mathcal{S}^1(\Omega, \mathcal{F})$, $L \in \mathcal{M}$ fixed we have $\varrho(L) \geqslant E_Q(L) - \alpha(Q)$ and hence also

$$\begin{aligned}\alpha(Q) \geqslant \sup_{L\in\mathcal{M}}\{E_Q(L) - \varrho(L)\} &\geqslant \sup_{L\in A_\varrho}\{E_Q(L) - \varrho(L)\}\\ &\geqslant \sup_{L\in A_\varrho} E_Q(L) = \alpha_{\min}(Q). \qquad (8.13)\end{aligned}$$

Next we give an alternative representation of $\alpha_{\min}$ that will be useful in the analysis of risk measures for linear portfolios. Consider some $L \in \mathcal{M}$. Since $L - \varrho(L) \in A_\varrho$ we get that

$$\sup_{L\in\mathcal{M}}\{E_Q(L) - \varrho(L)\} = \sup_{L\in\mathcal{M}}\{E_Q(L - \varrho(L))\} \leqslant \sup_{L\in A_\varrho} E_Q(L) = \alpha_{\min}(Q).$$

Combining this with the previous estimate (8.13) we obtain the relation

$$\alpha_{\min}(Q) = \sup_{L \in \mathcal{M}} \{E_Q(L) - \varrho(L)\}. \tag{8.14}$$

8.1.3 Examples of Dual Representations

In this section we give a detailed derivation of the dual representation for expected shortfall, and we briefly discuss the dual representation for the risk measure based on the exponential loss function.

Expected shortfall. Recall from Section 2.3.4 that expected shortfall is given by

$$\mathrm{ES}_\alpha(L) = \frac{1}{1-\alpha}\int_\alpha^1 q_u(L)\,\mathrm{d}u, \quad \alpha \in [0, 1),$$

for integrable losses L. The following lemma gives alternative expressions for ES_α.

Proposition 8.13. *For* $0 < \alpha < 1$ *we have*

$$\begin{aligned} \mathrm{ES}_\alpha(L) &= \frac{1}{1-\alpha}E((L - q_\alpha(L))^+) + q_\alpha(L) & (8.15) \\ &= \frac{1}{1-\alpha}(E(L; L > q_\alpha(L)) + q_\alpha(L)(1 - \alpha - P(L > q_\alpha(L)))). & (8.16) \end{aligned}$$

Proof. Recall that for $U \sim U(0, 1)$ the random variable $F^{\leftarrow}(U)$ has distribution function F_L (see Proposition 7.2). Since $q_\alpha(L) = F_L^{\leftarrow}(\alpha)$, (8.15) follows from observing that

$$\begin{aligned} \frac{1}{1-\alpha}E((L - q_\alpha(L))^+) &= \frac{1}{1-\alpha}\int_0^1 (F_L^{\leftarrow}(u) - F_L^{\leftarrow}(\alpha))^+\,\mathrm{d}u \\ &= \frac{1}{1-\alpha}\int_\alpha^1 (F_L^{\leftarrow}(u) - F_L^{\leftarrow}(\alpha))\,\mathrm{d}u \\ &= \frac{1}{1-\alpha}\int_\alpha^1 q_u(L)\,\mathrm{d}u - q_\alpha(L). \end{aligned}$$

For (8.16) we use $E((L - q_\alpha(L))^+) = E(L; L > q_\alpha(L)) - q_\alpha(L)P(L > q_\alpha(L))$. □

If F_L is continuous at $q_\alpha(L)$, then $P(L > q_\alpha(L)) = 1 - \alpha$ and

$$\mathrm{ES}_\alpha(L) = \frac{E(L; L > q_\alpha(L))}{1-\alpha} = E(L \mid L > \mathrm{VaR}_\alpha)$$

(see also Lemma 2.13).

Theorem 8.14. *For* $\alpha \in [0, 1)$, ES_α *defines a coherent measure of risk on* $\mathcal{M} = \mathcal{L}^1(\Omega, \mathcal{F}, P)$. *The dual representation is given by*

$$\mathrm{ES}_\alpha(L) = \max\{E^Q(L) \colon Q \in \mathcal{Q}_\alpha\}, \tag{8.17}$$

where $\mathcal{Q}_\alpha$ *is the set of all probability measures on* $(\Omega, \mathcal{F})$ *that are absolutely continuous with respect to* P *and for which the measure-theoretic density* $\mathrm{d}Q/\mathrm{d}P$ *is bounded by* $(1 - \alpha)^{-1}$.

Proof. For $\alpha = 0$ one has $Q_\alpha = \{P\}$ and the relation (8.17) is obvious so we consider the case where $\alpha > 0$. By translation invariance we can assume without loss of generality that $q_\alpha(L) > 0$. Since expected shortfall is only concerned with the upper tail, we can also assume that $L \geqslant 0$. (If it were not, we could simply define $\tilde{L} = \max(L, 0)$ and observe that $\mathrm{ES}_\alpha(\tilde{L}) = \mathrm{ES}_\alpha(L)$.)

Define a coherent measure ϱ_α by

$$\varrho_\alpha(L) = \sup\{E^Q(L) \colon Q \in Q_\alpha\}. \tag{8.18}$$

We want to show that $\varrho_\alpha(L)$ can be written in the form of Proposition 8.13. As a first step we transform the optimization problem in (8.18). The measures in the set $\mathcal{Q}_\alpha$ can alternatively be described in terms of their measure-theoretic density and hence by the set of random variables $\{\psi \colon 0 \leqslant \psi \leqslant 1/(1-\alpha),\ E(\psi) = 1\}$. We therefore have

$$\varrho_\alpha(L) = \sup\left\{E(L\psi) \colon 0 \leqslant \psi \leqslant \frac{1}{1-\alpha},\ E(\psi) = 1\right\}.$$

Transforming these random variables according to $\varphi = (1-\alpha)\psi$ and factoring out the expression $E(L)$ we get

$$\varrho_\alpha(L) = \frac{E(L)}{1-\alpha} \sup\left\{E\left(\frac{L}{E(L)}\varphi\right) \colon 0 \leqslant \varphi \leqslant 1,\ E(\varphi) = 1-\alpha\right\}.$$

It follows that

$$\varrho_\alpha(L) = \frac{E(L)}{1-\alpha} \sup\{\tilde{E}(\varphi) \colon 0 \leqslant \varphi \leqslant 1,\ E(\varphi) = 1-\alpha\}, \tag{8.19}$$

where the measure $\tilde{P}$ is defined by $\mathrm{d}\tilde{P}/\mathrm{d}P = L/E(L)$.

We now show that the supremum in (8.19) is attained by the random variable

$$\varphi_0 = I_{\{L > q_\alpha(L)\}} + \kappa I_{\{L = q_\alpha(L)\}}, \tag{8.20}$$

where $\kappa \geqslant 0$ is chosen such that $E(\varphi_0) = (1-\alpha)$. To verify the optimality of φ_0 consider an arbitrary $0 \leqslant \varphi \leqslant 1$ with $E(\varphi) = (1-\alpha)$. By definition of φ_0, we must have the inequality

$$0 \leqslant (\varphi_0 - \varphi)(L - q_\alpha(L)), \tag{8.21}$$

as the first factor is nonnegative for $L > q_\alpha(L)$ and nonpositive for $L < q_\alpha(L)$. Integration of (8.21) gives

$$0 \leqslant E((\varphi_0 - \varphi)(L - q_\alpha(L))) = E(L(\varphi_0 - \varphi)) - q_\alpha(L)E(\varphi_0 - \varphi).$$

The second term now vanishes as $E(\varphi_0) = E(\varphi) = 1-\alpha$, and the first term equals $E(L)\tilde{E}(\varphi_0 - \varphi)$. It follows that $\tilde{E}(\varphi_0) \geqslant \tilde{E}(\varphi)$, verifying the optimality of φ_0.

Inserting φ_0 in (8.19) gives

$$\begin{aligned}\varrho_\alpha(L) &= \frac{E(L)}{(1-\alpha)}\tilde{E}(\varphi_0)\\ &= \frac{1}{1-\alpha}E(L\varphi_0)\\ &= \frac{1}{1-\alpha}(E(L; L > q_\alpha(L)) + \kappa q_\alpha(L)P(L = q_\alpha(L))).\end{aligned} \tag{8.22}$$

The condition $E(\varphi_0) = (1-\alpha)$ yields that $P(L > q_\alpha(L)) + \kappa P(L = q_\alpha(L)) = (1-\alpha)$ and hence that

$$\kappa = \frac{(1-\alpha) - P(L > q_\alpha(L))}{P(L = q_\alpha(L))},$$

where we use the convention $0/0 = 0$. Inserting κ in (8.22) gives (8.16), which proves the theorem. □

Remark 8.15. Readers familiar with mathematical statistics may note that the construction of φ_0 in the optimization problem (8.19) is similar to the construction of the optimal test in the well-known Neyman–Pearson Lemma. The proof shows that for $\alpha \in (0,1)$ and integrable L, a measure $Q_L \in \mathcal{Q}_\alpha$ that attains the supremum in the dual representation of ES_α is given by the measure-theoretic density

$$\frac{\mathrm{d}Q_L}{\mathrm{d}P} = \frac{\varphi_0}{(1-\alpha)} = \frac{1}{1-\alpha}(I_{\{L>q_\alpha(L)\}} + \kappa I_{\{L=q_\alpha(L)\}}),$$

where

$$\kappa = \begin{cases} \dfrac{(1-\alpha) - P(L > q_\alpha(L))}{P(L = q_\alpha(L))}, & P(L = q_\alpha(L)) > 0, \\ 0, & P(L = q_\alpha(L)) = 0. \end{cases}$$

Risk measure for the exponential loss function. We end this section by giving without proof the dual representation for the risk measure derived from the acceptance set of the exponential loss function, given by

$$\varrho_\alpha(L) = \frac{1}{\alpha}\log\{E(\mathrm{e}^{\alpha L})\}$$

(see Example 8.8 for details). In this case we have a non-zero penalty function since ϱ_α is convex but not coherent. It can be shown that

$$\varrho_\alpha(L) = \max_{Q\in\mathcal{S}_1}\left\{E_Q(L) - \frac{1}{\alpha}H(Q \mid P)\right\},$$

where

$$H(Q \mid P) = \begin{cases} E_Q\left(\ln\dfrac{\mathrm{d}Q}{\mathrm{d}P}\right) & \text{if } Q \ll P, \\ \infty & \text{otherwise}, \end{cases}$$

is known as *relative entropy* between P and Q. In other words, the penalty function is proportional to the relative entropy between the two measures. The proof of this fact requires a detailed study of the properties of relative entropy and is therefore omitted (see Sections 3.2 and 4.9 of Föllmer and Schied (2011) for details).

Notes and Comments

The classic paper on coherent risk measures is Artzner et al. (1999); a non-technical introduction by the same authors is Artzner et al. (1997). Technical extensions such as the characterization of coherent risk measures on infinite probability spaces are given in Delbaen (2000, 2002, 2012). Stress testing as an approach to risk measurement is studied by Berkowitz (2000), Kupiec (1998), Breuer et al. (2009) and Rebonato (2010), among others.

The study of convex risk measures in the context of risk management and mathematical finance began with Föllmer and Schied (2002) (see also Frittelli 2002). A good treatment at advanced textbook level is given in Chapter 4 of Föllmer and Schied (2011). Cont (2006) provides an interesting link between convex risk measures and model risk in the pricing of derivatives. An alternative proof of Theorem 8.11 can be based on the duality theorem for convex functions (see, for instance, Remark 4.18 of Föllmer and Schied (2011)).

Different existing notions of expected shortfall are discussed in the very readable paper by Acerbi and Tasche (2002). Expected shortfall has been independently studied by Rockafellar and Uryasev (2000, 2002) under the name *conditional value-at-risk*; in particular, these papers develop the idea that expected shortfall can be obtained as the value of a convex optimization problem.

There has been recent interest in the subject of multi-period risk measures, which take into account the evolution of the final value of a position over several time periods and consider the effect of intermediate information and actions. Important papers in this area include Artzner et al. (2007), Riedel (2004), Weber (2006) and Cheridito, Delbaen and Kupper (2005). A textbook treatment and further references can be found in Chapter 11 of Föllmer and Schied (2011).

8.2 Law-Invariant Coherent Risk Measures

A risk measure ϱ is termed *law invariant* if $\varrho(L)$ depends on L only via its df F_L; examples of law-invariant risk measures are VaR and expected shortfall. On the other hand, the stress-test risk measures of Example 8.9 are typically not law invariant. In this section we discuss a number of law-invariant and coherent risk measures that are frequently used in financial and actuarial studies.

8.2.1 Distortion Risk Measures

The class of distortion risk measures is an important class of coherent risk measures. These risk measures are presented in many different ways in the literature, and a variety of different names are used. We begin by summarizing the more important representations before investigating the properties of distortion risk measures. Finally, we consider certain parametric families of distortion risk measures.

Representations of distortion risk measures We begin with a general definition.

Definition 8.16 (distortion risk measure).

(1) A convex distortion function D is a convex, increasing and absolutely continuous function on [0, 1] satisfying $D(0) = 0$ and $D(1) = 1$.

(2) The distortion risk measure associated with a convex distortion function D is defined by

$$\varrho(L) = \int_0^1 q_u(L)\,\mathrm{d}D(u). \tag{8.23}$$

Note that every convex distortion function is a distribution function on [0, 1]. The simplest example of a distortion risk measure is expected shortfall, which is obtained by taking the convex distortion function $D_\alpha(u) = (1-\alpha)^{-1}(u-\alpha)^+$. Clearly, a distortion risk measure is law invariant (it depends on the rv L only via the distribution of L), as it is defined as an average of the quantiles of L.

Since a convex distortion function D is absolutely continuous by definition, it can be written in the form $D(u) = \int_0^u \phi(s)\,\mathrm{d}s$ for an increasing, positive function ϕ (the right derivative of D). This yields the alternative representation

$$\varrho(L) = \int_0^1 q_u(L)\phi(u)\,\mathrm{d}u. \tag{8.24}$$

A risk measure of the form (8.24) is also known as a *spectral risk measure*, and the function ϕ is called the *spectrum*. It can be thought of as a weighting function applied to the quantiles of the distribution of L. In the case of expected shortfall, the spectrum is $\phi(u) = (1-\alpha)^{-1}I_{\{u\geqslant\alpha\}}$, showing that an equal weight is placed on all quantiles beyond the α-quantile.

A second alternative representation is derived in the following lemma.

Lemma 8.17. *The distortion risk measure ϱ associated with a convex distortion function D can be written in the form*

$$\varrho(L) = \int_{\mathbb{R}} x\,\mathrm{d}D \circ F_L(x), \tag{8.25}$$

where $D\circ F_L$ denotes the composition of the functions D and F_L, that is, $D\circ F_L(x) = D(F_L(x))$.

Proof. Let $G(x) = D \circ F_L(x)$ and note that G is itself a distribution function. The associated quantile function is given by $G^{\leftarrow} = F_L^{\leftarrow} \circ D^{\leftarrow}$, as can be seen by using Proposition A.3 (iv) to write

$$G^{\leftarrow}(v) = \inf\{x : D \circ F_L(x) \geqslant v\} = \inf\{x : F_L(x) \geqslant D^{\leftarrow}(v)\}$$

and noting that this equals $F_L^{\leftarrow} \circ D^{\leftarrow}(v)$ by definition of the generalized inverse. The right-hand side of (8.25) can therefore be written in the form

$$\int_{\mathbb{R}} x\,\mathrm{d}G(x) = \int_0^1 G^{\leftarrow}(u)\,\mathrm{d}u = \int_0^1 F_L^{\leftarrow} \circ D^{\leftarrow}(u)\,\mathrm{d}u = E(F_L^{\leftarrow} \circ D^{\leftarrow}(U)),$$

where U is a standard uniform random variable. Now introduce the random variable $V = D^{\leftarrow}(U)$, which has df D. We have shown that

$$\int_{\mathbb{R}} x\,\mathrm{d}D \circ F_L(x) = E(F^{\leftarrow}(V)) = \int_0^1 F_L^{\leftarrow}(v)\,\mathrm{d}D(v)$$

and thus established the result. □

The representation (8.25) gives more intuition for the idea of a distortion. The original df F_L is distorted by the function D. Moreover, for $u \in (0, 1)$ we note that $D(u) \leqslant u$, by the convexity of D, so that the distorted df $G = D \circ F$ places more mass on high values of L than the original df F.

Finally, we show that a distortion risk measure can be represented as a weighted average of expected shortfall over different confidence levels. To do this we fix a convex distortion D with associated distortion risk measure ϱ and spectrum ϕ (see (8.24)). From now on we work with the right-continuous version of ϕ. Since ϕ is increasing we can obtain a measure ν on $[0, 1]$ by setting $\nu([0, t]) := \phi(t)$ for $0 \leqslant t \leqslant 1$. Note that for every function $f\colon [0, 1] \to \mathbb{R}$ we have

$$\int_0^1 f(\alpha)\,\mathrm{d}\nu(\alpha) = f(0)\phi(0) + \int_0^1 f(\alpha)\,\mathrm{d}\phi(\alpha).$$

Moreover, we now define a further measure μ on $[0, 1]$ by setting

$$\frac{\mathrm{d}\mu}{\mathrm{d}\nu}(\alpha) = (1-\alpha), \quad \text{that is,} \int_0^1 f(\alpha)\,\mathrm{d}\mu(\alpha) = f(0) + \int_0^1 f(\alpha)(1-\alpha)\,\mathrm{d}\phi(\alpha). \tag{8.26}$$

Now we may state the representation result for ϱ.

Proposition 8.18. *Let ϱ be a distortion risk measure associated with the convex distortion D and define the measure μ by (8.26). Then μ is a probability measure and we have the representation*

$$\varrho(L) = \int_0^1 \mathrm{ES}_\alpha(L)\,\mathrm{d}\mu(\alpha).$$

Proof. Using integration by parts and (8.26) we check that

$$\begin{aligned}\mu([0,1]) &= \phi(0) + \int_0^1 (1-\alpha)\,\mathrm{d}\phi(\alpha) = \phi(0) + \int_0^1 \phi(\alpha)\,\mathrm{d}\alpha - \phi(0)\\ &= \int_0^1 \phi(\alpha)\,\mathrm{d}\alpha = D(1) - D(0) = 1,\end{aligned}$$

which shows that μ is a probability measure. Next we turn to the representation result for ϱ. By Fubini's Theorem we have that

$$\begin{aligned}\varrho(L) &= \int_0^1 q_u(L)\phi(u)\,\mathrm{d}u = \int_0^1 q_u(L)\int_0^u 1\,\mathrm{d}\nu(\alpha)\,\mathrm{d}u\\ &= \int_0^1\int_0^1 q_u(L)1_{\{\alpha\leqslant u\}}\mathrm{d}\nu(\alpha)\,\mathrm{d}u = \int_0^1\int_\alpha^1 q_u(L)\,\mathrm{d}u\,\mathrm{d}\nu(\alpha)\\ &= \int_0^1 (1-\alpha)\,\mathrm{ES}_\alpha(L)\,\mathrm{d}\nu(\alpha) = \int_0^1 \mathrm{ES}_\alpha(L)\,\mathrm{d}\mu(\alpha).\end{aligned}$$

□

Properties of distortion risk measures. Next we discuss certain properties of distortion risk measures. Distortion risk measures are *comonotone additive* in the following sense.

Definition 8.19 (comonotone additivity). A risk measure ϱ on a space of random variables $\mathcal{M}$ is said to be comonotone additive if

$$\varrho(L_1 + \cdots + L_d) = \varrho(L_1) + \cdots + \varrho(L_d)$$

whenever $(L_1, \ldots, L_d)$ is a vector of comonotonic risks.

The property of comonotonicity was defined in Definition 7.2.1, and it was shown in Proposition 7.20 that the quantile function (or, in other words, the value-at-risk risk measure) is additive for comonotonic risks. The comonotone additivity of the distortion risk measures follows easily from the fact that they can be represented as weighted integrals of the quantile function as in (8.23).

Moreover, distortion risk measure are coherent. Monotonicity, translation invariance and positive homogeneity are obvious. To establish that they are coherent we only need to check subadditivity, which follows immediately from Proposition 8.18 and Theorem 8.14 by observing that

$$\begin{aligned}\varrho(L_1+L_2) &= \int_0^1 \mathrm{ES}_\alpha(L_1+L_2)\,\mathrm{d}\mu(\alpha)\\ &\leqslant \int_0^1 \mathrm{ES}_\alpha(L_1)\,\mathrm{d}\mu(\alpha) + \int_0^1 \mathrm{ES}_\alpha(L_2)\,\mathrm{d}\mu(\alpha) = \varrho(L_1)+\varrho(L_2).\end{aligned}$$

In summary, we have verified that distortion risk measures are law invariant, coherent and comonotone additive. In fact, it may also be shown that, on a probability space without atoms (i.e. a space where $P(\{\omega\}) = 0$ for all ω), a law-invariant, coherent, comonotone-additive risk measure must be of the form (8.23) for some convex distortion D.

Example 8.20 (parametric families). A number of useful parametric families of distortion risk measures can be based on convex distortion functions that take the form

$$D_\alpha(u) = \Psi(\Psi^{-1}(u) + \ln(1-\alpha)), \quad 0 \leqslant \alpha < 1, \tag{8.27}$$

where Ψ is a continuous df on $\mathbb{R}$. The distortion function for expected shortfall is obtained when $\Psi(u) = 1 - \mathrm{e}^{-u}$ for $u \geqslant 0$, the standard exponential df.

There has been interest in distortion risk measures that are obtained by considering different dfs Ψ that are strictly increasing on the whole real line $\mathbb{R}$. A natural question concerns the constraints on Ψ that lead to convex distortion functions. It is straightforward to verify, by differentiating $D_\alpha(u)$ in (8.27) twice with respect to u, that a necessary and sufficient condition is that $\ln\psi(u)$ is concave, where ψ denotes the density of Ψ (see Tsukahara 2009).

A family of convex distortion functions of the form (8.27) is strictly decreasing in α for fixed u. Moreover, $D_0(u) = u$ (corresponding to the risk measure $\varrho(L) = E(L)$) and $\lim_{\alpha\to 1} D(u) = 1_{\{u=1\}}$. The fact that for $\alpha_1 < \alpha_2$ and $0 < u < 1$ we have $D_{\alpha_1}(u) > D_{\alpha_2}(u)$ means that, roughly speaking, D_{α_2} distorts the original probability measure more than D_{α_1} and places more weight on outcomes in the tail.

A particular example is obtained by taking $\Psi(u) = 1/(1+\mathrm{e}^{-u})$, the standard logistic df. This leads to the parametric family of convex distortion functions given by $D_\alpha(u) = (1-\alpha)u(1-\alpha u)^{-1}$ for $0 \leqslant \alpha < 1$. Writing $G_\alpha(x) = D_\alpha \circ F_L(x)$ we can show that

$$\left(\frac{1-G_\alpha(x)}{G_\alpha(x)}\right) = \frac{1}{1-\alpha}\left(\frac{1-F_L(x)}{F_L(x)}\right),$$

and this yields an interesting interpretation for this family. For every possibly critical loss level x, the *odds* of the tail event $\{L > x\}$ given by $(1 - F_L(x))/F_L(x)$ are multiplied by $(1-\alpha)^{-1}$ under the distorted loss distribution. For this reason the family is known as the *proportional odds family*.

8.2.2 The Expectile Risk Measure

Definition 8.21. Let $\mathcal{M} := L^1(\Omega, \mathcal{F}, P)$, the set of all integrable random variables L with $E|L| < \infty$. Then, for $\alpha \in (0,1)$ and $L \in \mathcal{M}$, the α-expectile $e_\alpha(L)$ is given by the unique solution y of the equation

$$\alpha E((L-y)^+) = (1-\alpha)E((L-y)^-), \tag{8.28}$$

where $x^+ = \max(x, 0)$ and $x^- = \max(-x, 0)$.

Recalling that $x^+ - x^- = x$, we note that $e_{0.5}(L) = E(L)$ since

$$\begin{aligned} E((L-y)^-) = E((L-y)^+) &\iff E((L-y)^+ - (L-y)^-) = 0 \\ &\iff E(L-y) = 0. \end{aligned}$$

For square-integrable losses L, the expectile $e_\alpha(L)$ can also be viewed as the minimizer in an optimization problem of the form

$$\min_{y\in\mathbb{R}} E(S(y, L)) \tag{8.29}$$

for a so-called *scoring function* $S(y, L)$. This could be relevant for the out-of-sample testing of expectile estimates (so-called backtesting), as will be explained in Section 9.3. The particular scoring function that yields the expectile is

$$S^e_\alpha(y, L) = |1_{\{L\leqslant y\}} - \alpha|(L-y)^2. \tag{8.30}$$

In fact, we can compute that

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}y}E(S^e_\alpha(y, L)) &= \frac{\mathrm{d}}{\mathrm{d}y}\int_{-\infty}^{\infty} |1_{\{y\geqslant x\}} - \alpha|(y-x)^2\,\mathrm{d}F_L(x) \\ &= \frac{\mathrm{d}}{\mathrm{d}y}\int_{-\infty}^{y} (1-\alpha)(y-x)^2\,\mathrm{d}F_L(x) + \frac{\mathrm{d}}{\mathrm{d}y}\int_{y}^{\infty} \alpha(y-x)^2\mathrm{d}F_L(x) \\ &= 2(1-\alpha)\int_{-\infty}^{y} (y-x)\,\mathrm{d}F_L(x) + 2\alpha\int_{y}^{\infty} (y-x)\,\mathrm{d}F_L(x) \\ &= 2(1-\alpha)E((L-y)^-) - 2\alpha E((L-y)^+), \end{aligned} \tag{8.31}$$

and setting this equal to zero yields the equation (8.28) that defines an expectile.

Remark 8.22. In Section 9.3.3 we will show that the α-quantile $q_\alpha(L)$ is also a minimizer in an optimization problem of the form (8.29) if we consider the scoring function

$$S^q_\alpha(y, L) = |1_{\{L\leqslant y\}} - \alpha|\,|L-y|. \tag{8.32}$$

We now show that the α-expectile of an arbitrary df F_L can be represented as the α-quantile of a related df $\tilde{F}_L$ that is strictly increasing on its support. This also shows the uniqueness of the α-expectile of a distribution. Moreover, we obtain a formula that can he helpful for computing expectiles of certain distributions and we illustrate this with a simple example.

Proposition 8.23. *Let $\alpha \in (0,1)$ and L be an rv such that $\mu := E(L) < \infty$. Then the solution $e_\alpha(L)$ of (8.28) may be written as $e_\alpha(L) = \tilde{F}_L^{-1}(\alpha)$, where*

$$\tilde{F}_L(y) = \frac{yF_L(y) - \mu(y)}{2(yF_L(y) - \mu(y)) + \mu - y} \tag{8.33}$$

is a continuous df that is strictly increasing on its support and $\mu(y) := \int_{-\infty}^{y} x \, \mathrm{d}F_L(x)$ is the lower partial moment of F_L.

Proof. Since $x^+ + x^- = |x|$, equation (8.31) shows that the expectile y must also solve

$$\alpha E(|L-y|) = E((L-y)^-) = \int_{-\infty}^{y} (y-x)\,\mathrm{d}F_L(x) = yF_L(y) - \mu(y).$$

Moreover,

$$\begin{aligned} E(|L-y|) &= \int_{-\infty}^{y} (y-x)\,\mathrm{d}F_L(x) + \int_{y}^{\infty} (x-y)\,\mathrm{d}F_L(x) \\ &= 2\int_{-\infty}^{y} (y-x)\,\mathrm{d}F_L(x) + \int_{-\infty}^{\infty} (x-y)\,\mathrm{d}F_L(x) \\ &= 2(yF_L(y) - \mu(y)) + \mu - y, \end{aligned}$$

and hence $\alpha = \tilde{F}_L(y)$ with $\tilde{F}_L$ as defined in (8.33).

Next we show that $\tilde{F}_L$ is indeed a distribution function. The derivative of $\tilde{F}_L$ can be easily computed to be

$$\tilde{f}_L(y) = \frac{\mu F_L(y) - \mu(y)}{(2(yF_L(y) - \mu(y)) + \mu - y)^2} = \frac{F_L(y)(\mu - E(L \mid L \leqslant y))}{(2(yF_L(y) - \mu(y)) + \mu - y)^2}.$$

Clearly, $\tilde{f}_L$ is nonnegative for all y and strictly positive on $D = \{y\colon 0 < F_L(y) < 1\}$, so that $\tilde{F}_L$ is increasing for all y and strictly increasing on D. Let $y_0 = \inf D$ and $y_1 = \sup D$ denote the left and right endpoints of F_L. It is then easy to check that $[y_0, y_1]$ is the support of $\tilde{F}_L$ and $\lim_{y\to y_0} \tilde{F}_L(y) = 0$ and $\lim_{y \to y_1} \tilde{F}_L(y) = 1$. □

In the following example we consider a Bernoulli distribution where the quantile is an unsatisfactory risk measure that can only take the values zero and one. In contrast, the expectile can take any value between zero and one.

Example 8.24. Let $L \sim \mathrm{Be}(p)$ be a Bernoulli-distributed loss. Then

$$F_L(y) = \begin{cases} 0, & y < 0, \\ 1-p, & 0 \leqslant y < 1, \\ 1, & y \geqslant 1, \end{cases} \qquad \mu(y) = \begin{cases} 0, & y < 1, \\ p, & y \geqslant 1, \end{cases}$$

from which it follows that

$$\tilde{F}_L(y) = \frac{y(1-p)}{y(1-2p)+p}, \quad 0 \leqslant y \leqslant 1 \quad \text{and} \quad e_\alpha(L) = \frac{\alpha p}{(1-\alpha) + p(2\alpha - 1)}.$$

Properties of the expectile.

Proposition 8.25. *Provided $\alpha \geqslant 0.5$, the expectile risk measure $\varrho = e_\alpha$ is a coherent risk measure on $\mathcal{M} = L^1(\Omega, \mathcal{F}, P)$.*

Proof. For $L \in \mathcal{M}$ and $y \in \mathbb{R}$ define the function

$$\begin{aligned} g(L, y, \alpha) &= \alpha E((L - y)^+) - (1 - \alpha) E((L - y)^-) \\ &= (2\alpha - 1) E((L - y)^+) + (1 - \alpha) E(L - y) \end{aligned}$$

and note that, for fixed L, it is a decreasing function of y and, for fixed y, it is monotonic in L, so $L_1 \leqslant L_2 \Rightarrow g(L_1, y, \alpha) \leqslant g(L_2, y, \alpha)$.

Translation invariance and positive homogeneity follow easily from the fact that if $g(L, y, \alpha) = 0$ (i.e. if $e_\alpha(L) = y$), then $g(L + m, y + m, \alpha) = 0$ for $m \in \mathbb{R}$ and $g(\lambda L, \lambda y, \alpha) = 0$ for $\lambda > 0$.

For monotonicity, fix α and let $y_1 = e_\alpha(L_1)$, $y_2 = e_\alpha(L_2)$. If $L_2 \geqslant L_1$ then $g(L_2, y_1, \alpha) \geqslant g(L_1, y_1, \alpha) = g(L_2, y_2, \alpha) = 0$. Since g is decreasing in y, it must be the case that $y_2 \geqslant y_1$.

For subadditivity, again let $y_1 = e_\alpha(L_1)$, $y_2 = e_\alpha(L_2)$. We have that

$$\begin{aligned} g(L_1 + L_2, y_1 + y_2, \alpha) &= (2\alpha - 1) E((L_1 + L_2 - y_1 - y_2)^+) \\ &\quad + (1 - \alpha) E(L_1 + L_2 - y_1 - y_2) \\ &= (2\alpha - 1) E((L_1 + L_2 - y_1 - y_2)^+) \\ &\quad + (1 - \alpha) E(L_1 - y_1) + (1 - \alpha) E(L_2 - y_2), \end{aligned}$$

and, since $(2\alpha - 1)E((L_i - y_i)^+) + (1 - \alpha)E(L_i - y_i) = 0$ for $i = 1, 2$, we get

$$\begin{aligned} g(L_1 + L_2, y_1 + y_2, \alpha) &= (2\alpha - 1)(E((L_1 + L_2 - y_1 - y_2)^+) \\ &\quad - E((L_1 - y_1)^+) - E((L_2 - y_2)^+)) \leqslant 0, \end{aligned}$$

where we have used the fact that $(x_1 + x_2)^+ \leqslant x_1^+ + x_2^+$. Since $g(L, y, \alpha)$ is decreasing in y it must be the case that $e_\alpha(L_1 + L_2) \leqslant y_1 + y_2$. □

The expectile risk measure is a law-invariant, coherent risk measure for $\alpha \geqslant 0.5$. However, it is not comonotone additive and therefore does not belong to the class of distortion risk measures described in Section 8.2.1.

If L_1 and L_2 are comonotonic random variables of the same type (so that $L_2 = kL_1 + m$ for some $m \in \mathbb{R}$ and $k > 0$), then we do have comonotone additivity (by the properties of translation invariance and positive homogeneity), but for comonotonic variables that are not of the same type we can find examples where $e_\alpha(L_1 + L_2) < e_\alpha(L_1) + e_\alpha(L_2)$ for $\alpha > 0.5$.

Notes and Comments

For distortion risk measures we use the definition of Tsukahara (2009) but restrict our attention to convex distortion functions. Using this definition, distortion risk measures are equivalent to the spectral risk measures of Kusuoka (2001), Acerbi (2002) and Tasche (2002). A parallel notion of distortion risk measures (or premium principles) has been developed in the insurance mathematics literature, where they are also known as Wang measures (see Wang 1996), although the ideas are also found in Denneberg (1990). A good discussion of these risk measures is given by Denuit and Charpentier (2004). The characterization of distortion risk measures on atomless probability spaces as law-invariant, comonotone-additive, coherent risk measures is due to Kusuoka (2001). The representation as averages of expected shortfalls is found in Föllmer and Schied (2011).

Adam, Houkari and Laurent (2008) gives examples of distortion risk measures used in the context of portfolio optimization. Further examples can be found in Tsukahara (2009), which discusses different choices of distortion function, the properties of the resulting risk measures and statistical estimation. The concept of a distortion also plays an important role in mathematical developments within prospect theory and behavioural finance: see, for instance, Zhou (2010) and He and Zhou (2011).

Expectiles have emerged as risk measures around the recent discussion related to elicitability, a statistical notion used for the comparison of forecasts (see Gneiting 2011; Ziegel 2015). This issue will be discussed in more detail in Chapter 9. Early references on expectiles include the papers by Newey and Powell (1987), Jones (1994) and Abdous and Remillard (1995); a textbook treatment is given in Remillard (2013).

8.3 Risk Measures for Linear Portfolios

In this section we consider linear portfolios in the set

$$\mathcal{M} = \{L \colon L = m + \boldsymbol{\lambda}' \boldsymbol{X},\ m \in \mathbb{R},\ \boldsymbol{\lambda} \in \mathbb{R}^d\}, \tag{8.34}$$

where $\boldsymbol{X}$ is a fixed d-dimensional random vector of risk factors defined on some probability space $(\Omega, \mathcal{F}, P)$. The case of linear portfolios is interesting for a number of reasons. To begin with, many standard approaches to risk aggregation and capital allocation are explicitly or implicitly based on the assumption that portfolio losses have a linear relationship to underlying risk factors. Moreover, as we observed in Chapter 2, it is common to use linear approximations for losses due to market risks over short time horizons.

In Section 8.3.1 we apply the dual representation of coherent risk measures to the case of linear portfolios. We show that every coherent risk measure on the set $\mathcal{M}$ in (8.34) can be viewed as a stress test in the sense of Example 8.9. In Section 8.3.2 we consider the important case where the factor vector $\boldsymbol{X}$ has an elliptical distribution, and in Section 8.3.3 we consider briefly the case of non-elliptical distributions. As well as deriving the form of the stress test in Section 8.3.2 we also collect a number of important related results concerning risk measurement for linear portfolios of elliptically distributed risks.

8.3.1 Coherent Risk Measures as Stress Tests

Given a positive-homogeneous risk measure $\varrho \colon \mathcal{M} \to \mathbb{R}$ it is convenient to define a risk-measure function $r_\varrho(\boldsymbol{\lambda}) = \varrho(\boldsymbol{\lambda}'\boldsymbol{X})$, which can be thought of as a function of portfolio weights. There is a one-to-one relationship between ϱ and r_ϱ given by

$$\varrho(m+\boldsymbol{\lambda}'\boldsymbol{X}) = m + r_\varrho(\boldsymbol{\lambda}).$$

Properties of ϱ therefore carry over to r_ϱ, and vice versa. We summarize the results in the following lemma.

Lemma 8.26. *Consider some translation-invariant risk measure $\varrho\colon \mathcal{M} \to \mathbb{R}$ with associated risk-measure function r_ϱ.*

(1) *ϱ is a positive-homogeneous risk measure if and only if r_ϱ is a positive-homogeneous function on $\mathbb{R}^d$, that is, $r_\varrho(t\boldsymbol{\lambda}) = t r_\varrho(\boldsymbol{\lambda})$ for all $t > 0$, $\boldsymbol{\lambda} \in \mathbb{R}^d$.*

(2) *Suppose that ϱ is positive homogeneous. Then ϱ is subadditive if and only if r_ϱ is a convex function on $\mathbb{R}^d$.*

The result follows easily from the definitions; a formal proof is therefore omitted.

The main result of this section shows that a coherent risk measure on the set of linear portfolios can be viewed as a stress test of the kind described in Example 8.9, where the scenario set is given by the set

$$S_\varrho := \{\boldsymbol{x} \in \mathbb{R}^d\colon \boldsymbol{u}'\boldsymbol{x} \leqslant r_\varrho(\boldsymbol{u}) \text{ for all } \boldsymbol{u} \in \mathbb{R}^d\}. \tag{8.35}$$

Proposition 8.27. *ϱ is a coherent risk measure on the set of linear portfolios $\mathcal{M}$ in (8.34) if and only if for every $L = m + \boldsymbol{\lambda}'\boldsymbol{X} \in \mathcal{M}$ we have the representation*

$$\varrho(L) = m + r_\varrho(\boldsymbol{\lambda}) = \sup\{m + \boldsymbol{\lambda}'\boldsymbol{x}\colon \boldsymbol{x} \in S_\varrho\}. \tag{8.36}$$

Proof. The risk measure given in (8.36) can be viewed as a generalized scenario: $\rho(L) = \sup\{E_Q(L)\colon Q \in \mathcal{Q}\}$, where $\mathcal{Q}$ is the set of Dirac measures $\{\delta_{\boldsymbol{x}} : \boldsymbol{x} \in S_\varrho\}$ (see Example 8.10). Such a risk measure is automatically coherent.

Conversely, suppose that ϱ is a coherent risk measure on the linear portfolio set $\mathcal{M}$. Since ϱ is translation invariant we can set $m = 0$ and consider random variables $L = \boldsymbol{\lambda}'\boldsymbol{X} \in \mathcal{M}$. Since $E_Q(\boldsymbol{\lambda}'\boldsymbol{X}) = \boldsymbol{\lambda}'E_Q(\boldsymbol{X})$, Theorem 8.11 shows that

$$\varrho(\boldsymbol{\lambda}'\boldsymbol{X}) = \sup\{\boldsymbol{\lambda}'E_Q(\boldsymbol{X})\colon Q \in \mathcal{S}^1(\Omega,\mathcal{F}),\ \alpha_{\min}(Q) = 0\}. \tag{8.37}$$

According to relation (8.14), the measures $Q \in \mathcal{S}^1(\Omega,\mathcal{F})$ for which $\alpha_{\min}(Q) = 0$ are those for which $E_Q(L) \leqslant \varrho(L)$ for all $L \in \mathcal{M}$. Hence we get

$$\begin{aligned}
\{Q \in \mathcal{S}^1&(\Omega,\mathcal{F})\colon \alpha_{\min}(Q) = 0\} \\
&= \{Q \in \mathcal{S}^1(\Omega,\mathcal{F})\colon \boldsymbol{u}'E_Q(\boldsymbol{X}) \leqslant r_\varrho(\boldsymbol{u}) \text{ for all } \boldsymbol{u} \in \mathbb{R}^d\} \\
&= \{Q \in \mathcal{S}^1(\Omega,\mathcal{F})\colon E_Q(\boldsymbol{X}) \in S_\varrho\}.
\end{aligned} \tag{8.38}$$

Now define the set $C := \{\boldsymbol{\mu} \in \mathbb{R}^d\colon \exists Q \in \mathcal{S}^1(\Omega,\mathcal{F}) \text{ with } \boldsymbol{\mu} = E_Q(\boldsymbol{X})\}$, and denote the closure of C by $\bar{C}$. Note that C and hence also $\bar{C}$ are convex subsets of $\mathbb{R}^d$. By combining (8.37) and (8.38) we therefore obtain

$$\varrho(\boldsymbol{\lambda}'\boldsymbol{X}) = \sup\{\boldsymbol{\lambda}'\boldsymbol{\mu}\colon \boldsymbol{\mu} \in C \cap S_\varrho\} = \sup\{\boldsymbol{\lambda}'\boldsymbol{\mu}\colon \boldsymbol{\mu} \in \bar{C} \cap S_\varrho\}.$$

If $S_\varrho \subset \bar{C}$, the last equation is equivalent to $\varrho(\boldsymbol{\lambda}'\boldsymbol{X}) = \sup\{\boldsymbol{\lambda}'\boldsymbol{\mu}\colon \boldsymbol{\mu} \in S_\varrho\}$, which is the result we require. Note that the key insight in this argument is the fact that a probability measure on the linear portfolio space $\mathcal{M}$ can be identified by its mean vector; it is here that the special structure of $\mathcal{M}$ enters.

To verify that $S_\varrho \subset \bar{C}$, suppose to the contrary that there is some $\boldsymbol{\mu}_0 \in S_\varrho$ that does not belong to $\bar{C}$. According to Proposition A.8 (b) (the strict separation part of the separating hyperplane theorem), there would exist some $\boldsymbol{u}^* \in \mathbb{R}^d \setminus \{0\}$ such that $\boldsymbol{\mu}_0'\boldsymbol{u}^* > \sup\{\boldsymbol{\mu}'\boldsymbol{u}^* \colon \boldsymbol{\mu} \in \bar{C}\}$. It follows that

$$\begin{aligned} r_\varrho(\boldsymbol{u}^*) = \varrho((\boldsymbol{u}^*)'\boldsymbol{X}) &\leqslant \sup\{E_Q((\boldsymbol{u}^*)'\boldsymbol{X}) \colon Q \in \mathcal{S}^1(\Omega, \mathcal{F})\} \\ &= \sup\{\boldsymbol{\mu}'\boldsymbol{u}^* \colon \boldsymbol{\mu} \in \bar{C}\} < \boldsymbol{\mu}_0'\boldsymbol{u}^*. \end{aligned}$$

This contradicts the fact that $\boldsymbol{\mu}_0 \in S_\varrho$, which requires $\boldsymbol{\mu}_0'\boldsymbol{u}^* \leqslant r_\varrho(\boldsymbol{u}^*)$. □

The scenario set S_ϱ in (8.35) is an intersection of the *half-spaces* $H_u = \{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}'\boldsymbol{x} \leqslant r_\varrho(\boldsymbol{u})\}$, so S_ϱ is a closed convex set. The precise form of S_ϱ depends on the distribution of $\boldsymbol{X}$ and on the risk measure ϱ. In the case of the quantile risk measure $\varrho = \mathrm{VaR}_\alpha$, the set S_ϱ has a probabilistic interpretation as a so-called *depth set*. Suppose that $\boldsymbol{X}$ is such that for all $\boldsymbol{u} \in \mathbb{R}^d \setminus \{0\}$ the random variable $\boldsymbol{u}'\boldsymbol{X}$ has a continuous distribution function. Then, for $H_u := \{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}'\boldsymbol{x} \leqslant \mathrm{VaR}_\alpha(\boldsymbol{u}'\boldsymbol{X})\}$ we have that $P(\boldsymbol{u}'\boldsymbol{X} \in H_u) = \alpha$, so that the set S_{VaR_α} is the intersection of all half-spaces with probability α.

8.3.2 Elliptically Distributed Risk Factors

We have seen in Chapter 6 that an elliptical model may be a reasonable approximate model for various kinds of risk-factor data, such as stock or exchange-rate returns. The next result summarizes some key results for risk measurement on linear spaces when the underlying distribution of the risk factors is elliptical.

Theorem 8.28 (risk measurement for elliptical risk factors). *Suppose that $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$ and let $\mathcal{M}$ be the space of linear portfolios (8.34). For any positive-homogeneous, translation-invariant and law-invariant risk measure ϱ on $\mathcal{M}$ the following properties hold.*

(1) *For any $L = m + \boldsymbol{\lambda}'\boldsymbol{X} \in \mathcal{M}$ we have*

$$\varrho(L) = \sqrt{\boldsymbol{\lambda}'\Sigma\boldsymbol{\lambda}}\,\varrho(Y) + \boldsymbol{\lambda}'\boldsymbol{\mu} + m, \tag{8.39}$$

where $Y \sim S_1(\psi)$, i.e. a univariate spherical distribution (a distribution that is symmetric around 0) with generator ψ.

(2) *If $\varrho(Y) \geqslant 0$ then ϱ is subadditive on $\mathcal{M}$. In particular, VaR_α is subadditive if $\alpha \geqslant 0.5$.*

(3) *If $\boldsymbol{X}$ has a finite-mean vector, then, for any $L = m + \boldsymbol{\lambda}'\boldsymbol{X} \in \mathcal{M}$, we have*

$$\varrho(L - E(L)) = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} \rho_{ij}\lambda_i\lambda_j\varrho(X_i - E(X_i))\varrho(X_j - E(X_j))}, \tag{8.40}$$

where the ρ_{ij} are elements of the correlation matrix $\wp(\Sigma)$.

(4) *If $\varrho(Y) > 0$ and $\boldsymbol{X}$ has a finite covariance matrix, then, for every $L \in \mathcal{M}$,*

$$\varrho(L) = E(L) + k_\varrho \sqrt{\operatorname{var}(L)} \tag{8.41}$$

for some constant $k_\varrho > 0$ that depends on the risk measure.

(5) *If $\varrho(Y) > 0$ and Σ is invertible, then the scenario set S_ϱ in the stress-test representation (8.36) of ϱ is given by the ellipsoid*

$$S_\varrho = \{\boldsymbol{x} \colon (\boldsymbol{x} - \boldsymbol{\mu})' \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \leqslant \varrho(Y)^2\}.$$

Proof. For any $L \in \mathcal{M}$ it follows from Definition 6.25 that we can write

$$L = m + \boldsymbol{\lambda}'\boldsymbol{X} \overset{\mathrm{d}}{=} \boldsymbol{\lambda}' A \boldsymbol{Y} + \boldsymbol{\lambda}'\boldsymbol{\mu} + m$$

for a spherical random vector $\boldsymbol{Y} \sim S_k(\psi)$, a matrix $A \in \mathbb{R}^{d \times k}$ satisfying $AA' = \Sigma$ and a constant vector $\boldsymbol{\mu} \in \mathbb{R}^d$. By Theorem 6.18 (3) we have

$$L \overset{\mathrm{d}}{=} \|A'\boldsymbol{\lambda}\| Y + \boldsymbol{\lambda}'\boldsymbol{\mu} + m, \tag{8.42}$$

where Y is a component of the random vector $\boldsymbol{Y}$ that has the symmetric distribution $Y \sim S_1(\psi)$. Every $L \in \mathcal{M}$ is therefore an rv of the same type, and the translation invariance and homogeneity of ϱ imply that

$$\varrho(L) = \|A'\boldsymbol{\lambda}\| \varrho(Y) + \boldsymbol{\lambda}'\boldsymbol{\mu} + m, \tag{8.43}$$

so that claim (1) follows.

For (2) set $L_1 = m_1 + \boldsymbol{\lambda}_1'\boldsymbol{X}$ and $L_2 = m_2 + \boldsymbol{\lambda}_2'\boldsymbol{X}$. Since $\|A'(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)\| \leqslant \|A'\boldsymbol{\lambda}_1\| + \|A'\boldsymbol{\lambda}_2\|$ and since $\varrho(Y_1) \geqslant 0$, the subadditivity of the risk measure follows easily. Since $E(L) = \boldsymbol{\lambda}'\boldsymbol{\mu} + m$ and ϱ is translation invariant, formula (8.39) implies that

$$\varrho(L - E(L)) = \left(\sum_{i=1}^{d} \sum_{j=1}^{d} \lambda_i \lambda_j \rho_{ij} \sigma_i \sigma_j \right)^{1/2} \varrho(Y), \tag{8.44}$$

where $\sigma_i = \sqrt{\Sigma_{ii}}$ for $i = 1, \dots, d$. As a special case of (8.44) we observe that

$$\varrho(X_i - E(X_i)) = \varrho(\boldsymbol{e}_i'\boldsymbol{X} - E(\boldsymbol{e}_i'\boldsymbol{X})) = \sigma_i \varrho(Y). \tag{8.45}$$

Combining (8.44) and (8.45) yields formula (8.40) and proves part (3).

For (4) assume that $\operatorname{cov}(\boldsymbol{X}) = c\Sigma$ for some positive constant c. It follows easily from (8.44) that $\varrho(L) = E(L) + \sqrt{\operatorname{var}(L)}\varrho(Y)/\sqrt{c}$ and $k_\varrho = \varrho(Y)/\sqrt{c}$.

For (5) note that part (2) implies that the risk-measure function $r_\varrho(\boldsymbol{\lambda})$ takes the form $r_\varrho(\boldsymbol{\lambda}) = \|A'\boldsymbol{\lambda}\|\varrho(Y) + \boldsymbol{\lambda}'\boldsymbol{\mu}$, so that the set S_ϱ in (8.36) is

$$\begin{aligned} S_\varrho &= \{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}'\boldsymbol{x} \leqslant \boldsymbol{u}'\boldsymbol{\mu} + \|A'\boldsymbol{u}\|\varrho(Y),\ \forall \boldsymbol{u} \in \mathbb{R}^d\} \\ &= \{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{u}'AA^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leqslant \|A'\boldsymbol{u}\|\varrho(Y),\ \forall \boldsymbol{u} \in \mathbb{R}^d\} \\ &= \left\{\boldsymbol{x} \in \mathbb{R}^d \colon \boldsymbol{v}'\frac{A^{-1}(\boldsymbol{x} - \boldsymbol{\mu})}{\varrho(Y)} \leqslant \|\boldsymbol{v}\|,\ \forall \boldsymbol{v} \in \mathbb{R}^d\right\}, \end{aligned}$$

where the last line follows because $\mathbb{R}^d = \{A'\boldsymbol{u}\colon \boldsymbol{u} \in \mathbb{R}^d\}$. By observing that the Euclidean unit ball $\{\boldsymbol{y} \in \mathbb{R}^d\colon \boldsymbol{y}'\boldsymbol{y} \leqslant 1\}$ can be written as the set $\{\boldsymbol{y} \in \mathbb{R}^d\colon \boldsymbol{v}'\boldsymbol{y} \leqslant \|\boldsymbol{v}\|,\ \forall \boldsymbol{v} \in \mathbb{R}^d\}$, we conclude that, for $\boldsymbol{x} \in S_\varrho$, the vectors $\boldsymbol{y} = A^{-1}(\boldsymbol{x} - \boldsymbol{\mu})/\varrho(Y)$ describe the unit ball and therefore

$$S_\varrho = \{\boldsymbol{x} \in \mathbb{R}^d\colon (\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) \leqslant \varrho(Y)^2\}.$$

□

The various parts of Theorem 8.28 have a number of important implications. Part (2) gives a special case where the VaR risk measure is subadditive and therefore coherent; we recall from Section 2.25 that this is not the case in general. Part (3) gives a useful interpretation of risk measures on $\mathcal{M}$ in terms of the aggregation of stress tests, as will be seen later in Section 8.4.

Part (4) is relevant to portfolio optimization. If we consider only the portfolio losses $L \in \mathcal{M}$ for which $E(L)$ is fixed at some level, then the portfolio weights that minimize ϱ also minimize the variance. The portfolio minimizing the risk measure ϱ is the same as the Markowitz variance-minimizing portfolio.

Part (5) shows that the scenario sets in the stress-test representation of coherent risk measures are ellipsoids when the distribution of risk-factor changes is elliptical. Moreover, for different examples of law-invariant coherent risk measures, we simply obtain ellipsoids of differing radius $\varrho(Y)$. Scenario sets of ellipsoidal form are often used in practice and this result provides a justification for this practice in the case of linear portfolios of elliptical risk factors.

8.3.3 Other Risk Factor Distributions

We now turn briefly to the application of Proposition 8.27 in situations where the risk factors do not have an elliptical distribution. The VaR risk measure is not coherent on the linear space $\mathcal{M}$ in general. Consider the simple case where we have two independent standard exponentially distributed risk factors X_1 and X_2.

Here it may happen that VaR_α is not coherent on $\mathcal{M}$ for some values of α. In such situations, (8.36) does not hold in general and we may find vectors of portfolio weights $\boldsymbol{\lambda}$ such that

$$\mathrm{VaR}_\alpha(\boldsymbol{\lambda}'\boldsymbol{X}) > \sup\{\boldsymbol{\lambda}'\boldsymbol{x}\colon \boldsymbol{x} \in S_\alpha\},$$

where

$$S_\alpha := S_{\mathrm{VaR}_\alpha} = \{\boldsymbol{x} \in \mathbb{R}^d\colon \boldsymbol{u}'\boldsymbol{x} \leqslant \mathrm{VaR}_\alpha(\boldsymbol{u}'\boldsymbol{X}),\ \forall \boldsymbol{u} \in \mathbb{R}^d\}.$$

Such a situation is shown in Figure 8.2 (a) for two independent standard exponential risk factors. Each line bounds a half-space with probability $\alpha = 0.7$, and the intersection of these half-spaces (the empty area in the centre) is the set $S_{0.7}$. Some lines are not supporting hyperplanes of $S_{0.7}$, meaning they do not touch it; an example is the bold diagonal line in the upper right corner of the picture. In such situations we can construct vectors of portfolio weights $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ such that $\mathrm{VaR}_\alpha((\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)'\boldsymbol{X}) > \mathrm{VaR}_\alpha(\boldsymbol{\lambda}_1'\boldsymbol{X}) + \mathrm{VaR}_\alpha(\boldsymbol{\lambda}_2'\boldsymbol{X})$. In fact, for $\alpha = 0.7$ we simply have $\mathrm{VaR}_\alpha(X_1 + X_2) > \mathrm{VaR}_\alpha(X_1) + \mathrm{VaR}_\alpha(X_2)$, as may

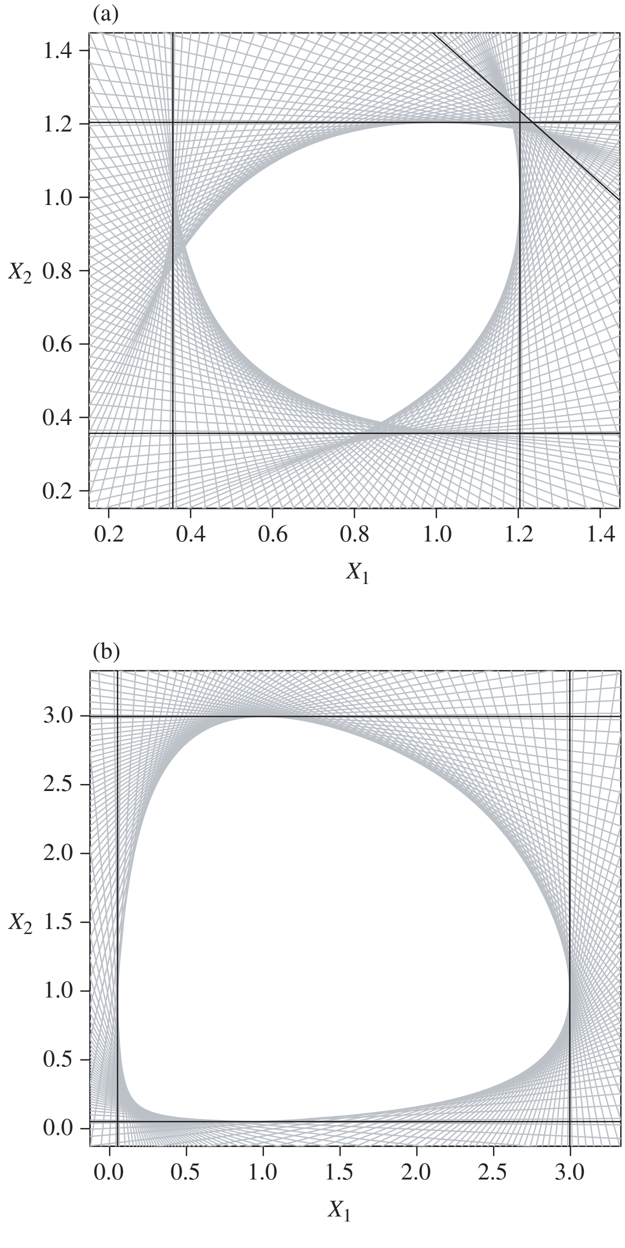

Figure 8.2. Illustration of scenario sets S_α when X_1 and X_2 are two independent standard exponential variates and the risk measure is VaR_α. (a) The case $\alpha = 0.7$, where VaR_α is not coherent on $\mathcal{M}$. (b) The case $\alpha = 0.95$, where VaR_α is coherent.

be deduced from the picture by noting that the black vertical line on the right is $x_1 = \mathrm{VaR}_\alpha(X_1)$, the upper black horizontal line is $x_2 = \mathrm{VaR}_\alpha(X_2)$, and the bold diagonal line is $x_1 + x_2 = \mathrm{VaR}_\alpha(X_1 + X_2)$. This agrees with our observation in Example 7.30.

For the value $\alpha = 0.95$, the depth set is shown in Figure 8.2 (b). In this case the depth set has a smooth boundary and there are supporting hyperplanes bounding half-spaces with probability α in every direction. We can apply Proposition 8.27 to conclude that VaR_α is a coherent risk measure on $\mathcal{M}$ for $\alpha = 0.95$.

These issues do not arise for the expected shortfall risk measure or any other coherent risk measure. The scenario set S_ϱ for these risk measures would have supporting hyperplanes with equation $\boldsymbol{u}'\boldsymbol{x} = r_\varrho(\boldsymbol{u})$ for every direction $\boldsymbol{u} \neq \boldsymbol{0}$.

Notes and Comments

The presentation of the relationship between coherent risk measures and stress tests on linear portfolio spaces is based on McNeil and Smith (2012). Our definition of a stress test coincides with the concept of the *maximum loss* risk measure introduced by Studer (1997, 1999), who also considers ellipsoidal scenario sets. Breuer et al. (2009) describe the problem of finding scenarios that are "plausible, severe and useful" and propose a number of refinements to Studer's approach based on ellipsoidal sets.

The depth sets obtained when the risk measure is VaR have an interesting history in statistics and have been studied by Massé and Theodorescu (1994) and Rousseeuw and Ruts (1999), among others. The concept has its origins in an empirical concept of *data depth* introduced by Tukey (1975) as well as in theoretical work on multivariate analogues of the quantile function by Eddy (1984) and Nolan (1992).

The treatment of the implications of elliptical distributions for risk measurement follows Embrechts, McNeil and Straumann (2002). Chapter 9 of Hult et al. (2012) contains an interesting discussion of elliptical distributions in risk management. There is an extensive body of economic theory related to the use of elliptical distributions in finance. The papers by Owen and Rabinovitch (1983), Chamberlain (1983) and Berk (1997) provide an entry to the area. Landsman and Valdez (2003) discuss the explicit calculation of the quantity $E(L \mid L > q_\alpha(L))$ for portfolios of elliptically distributed risks. This coincides with expected shortfall for continuous loss distributions (see Proposition 2.13).

8.4 Risk Aggregation

The need to aggregate risk can arise in a number of situations. Suppose that capital amounts $\text{EC}_1, \dots, \text{EC}_d$ (EC stands for economic capital) have been computed for each of d subsidiaries or business lines making up an enterprise and a method for computing the aggregate capital for the whole enterprise is required. Or, in a similar vein, suppose that capital amounts $\text{EC}_1, \dots, \text{EC}_d$ have been computed for d different asset classes on the balance sheet of an enterprise and a method is required to compute the overall capital required to back all assets.

A *risk-aggregation rule* is a mapping

$$f\colon \mathbb{R}^d \to \mathbb{R}, \quad f(\text{EC}_1, \dots, \text{EC}_d) = \text{EC},$$

which takes as input the individual capital amounts and gives as output the aggregate capital EC. Examples of commonly used rules are *simple summation*

$$\text{EC} = \text{EC}_1 + \dots + \text{EC}_d \tag{8.46}$$

and *correlation adjusted summation*

$$\mathrm{EC} = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} \rho_{ij}\,\mathrm{EC}_i\,\mathrm{EC}_j}, \tag{8.47}$$

where the ρ_{ij} are a set of parameters satisfying $0 \leqslant \rho_{ij} \leqslant 1$, which are usually referred to as correlations. Of course, (8.46) is a special case of (8.47) when $\rho_{ij} = 1$, $\forall i, j$, and the aggregate capital given by (8.46) is an upper bound for the aggregate capital given by (8.47).

The application of rules like (8.46) and (8.47) in the absence of any deeper consideration of multivariate models for the enterprise or the use of risk measures is referred to as *rules-based aggregation*. By contrast, the use of aggregation rules that can be theoretically justified by relating capital amounts to risk measures and multivariate models for losses is referred to as *principles-based aggregation*. In the following sections we give examples of the latter approach.

8.4.1 Aggregation Based on Loss Distributions

In this section we suppose that the overall loss of the enterprise over a fixed time interval is given by $L_1 + \cdots + L_d$, where $L_1, \dots, L_d$ are the losses arising from sub-units of the enterprise (such as business units or asset classes on the balance sheet). We consider a translation-invariant risk measure ϱ and define a mean-adjusted version of the risk measure by

$$\varrho^{\mathrm{mean}}(L) = \varrho(L - E(L)) = \varrho(L) - E(L). \tag{8.48}$$

ϱ^{mean} can be thought of as the capital required to cover unexpected losses.

The capital requirements for the sub-units are given by $\mathrm{EC}_i = \varrho^{\mathrm{mean}}(L_i)$ for $i = 1, \dots, d$, and the aggregate capital should be given by $\mathrm{EC} = \varrho^{\mathrm{mean}}(L_1 + \cdots + L_d)$. We require an aggregation rule f such that $\mathrm{EC} = f(\mathrm{EC}_1, \dots, \mathrm{EC}_d)$.

As an example, suppose that we take the risk measure $\varrho(L) = k\,\mathrm{sd}(L) + E(L)$, where sd denotes the standard deviation, k is some positive constant, and second moments of the loss distributions are assumed to be finite. Regardless of the underlying distribution of $L_1, \dots, L_d$ the standard deviation satisfies

$$\mathrm{sd}(L) = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} \rho_{ij}\,\mathrm{sd}(L_i)\,\mathrm{sd}(L_j)}, \tag{8.49}$$

where the ρ_{ij} are the elements of the correlation matrix of $(L_1, \dots, L_d)$ and the aggregation rule (8.47) therefore follows in this case.

When the losses are elements of the linear space $\mathcal{M}$ in (8.34) and the distribution of the underlying risk-factor changes $\boldsymbol{X}$ is elliptical with finite covariance matrix, then (8.49) and Theorem 8.28 (4) imply that (8.47) is justified for any positive-homogeneous, translation-invariant and law-invariant risk measures. We now give a more elegant proof of this fact that does not require us to assume finite second moments of $\boldsymbol{X}$.

Proposition 8.29. *Let $X \sim E_k(\boldsymbol{\mu}, \Sigma, \psi)$ with $E(X) = \boldsymbol{\mu}$. Let $\mathcal{M} = \{L \colon L = m + \boldsymbol{\lambda}' X,\ \boldsymbol{\lambda} \in \mathbb{R}^k,\ m \in \mathbb{R}\}$ be the space of linear portfolios and let ϱ be a positive-homogeneous, translation-invariant and law-invariant risk measure on $\mathcal{M}$. For $L_1, \dots, L_d \in \mathcal{M}$ let $\mathrm{EC}_i = \varrho^{\mathrm{mean}}(L_i)$ and $\mathrm{EC} = \varrho^{\mathrm{mean}}(L_1 + \cdots + L_d)$. The capital amounts $\mathrm{EC}, \mathrm{EC}_1, \dots, \mathrm{EC}_d$ then satisfy the aggregation rule (8.47), where the ρ_{ij} are elements of the correlation matrix $P = \wp(\tilde{\Sigma})$ and where $\tilde{\Sigma}$ is the dispersion matrix of the (elliptically distributed) random vector $(L_1, \dots, L_d)$.*

Proof. Let $L_i = \boldsymbol{\lambda}_i' X + m_i$ for $i = 1, \dots, d$. It follows from Theorem 8.28 (1) that

$$\mathrm{EC}_i = \varrho(L_i) - E(L_i) = \sqrt{\boldsymbol{\lambda}_i' \Sigma \boldsymbol{\lambda}_i}\,\varrho(Y),$$

where $Y \sim S_1(\psi)$, and that

$$\begin{aligned}\mathrm{EC} &= \sqrt{(\boldsymbol{\lambda}_1 + \cdots + \boldsymbol{\lambda}_d)' \Sigma (\boldsymbol{\lambda}_1 + \cdots + \boldsymbol{\lambda}_d)}\,\varrho(Y) \\ &= \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{d} \boldsymbol{\lambda}_i' \Sigma \boldsymbol{\lambda}_j}\,\varrho(Y) \\ &= \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{d} \frac{\boldsymbol{\lambda}_i' \Sigma \boldsymbol{\lambda}_j}{\sqrt{(\boldsymbol{\lambda}_i' \Sigma \boldsymbol{\lambda}_i)(\boldsymbol{\lambda}_j' \Sigma \boldsymbol{\lambda}_j)}} \mathrm{EC}_i\, \mathrm{EC}_j}.\end{aligned}$$

The dispersion matrix $\tilde{\Sigma}$ of $(L_1, \dots, L_d)$ is now given by $\tilde{\Sigma} = \Lambda \Sigma \Lambda'$, where $\Lambda \in \mathbb{R}^{d \times k}$ is the matrix with rows given by the vectors $\boldsymbol{\lambda}_i$. The correlation matrix $P = \wp(\tilde{\Sigma})$ clearly has elements given by

$$\boldsymbol{\lambda}_i' \Sigma \boldsymbol{\lambda}_j / \sqrt{(\boldsymbol{\lambda}_i' \Sigma \boldsymbol{\lambda}_i)(\boldsymbol{\lambda}_j' \Sigma \boldsymbol{\lambda}_j)}$$

and the result follows. □

Proposition 8.29 implies that the aggregation rule (8.47) can be justified when we work with the mean-adjusted value-at-risk or expected shortfall risk measures if we are prepared to make the strong assumption that the underlying multivariate loss distribution is elliptical.

Clearly the elliptical assumption is unlikely to hold in practice, so the theoretical support that allows us to view (8.47) as a principles-based approach will generally be lacking. However, even if the formula is used as a pragmatic rule, there are also practical problems with the approach.

- The formula requires the specification of pairwise correlations between the losses $L_1, \dots, L_d$. It will be difficult to obtain estimates of these correlations, since empirical data is generally available at the level of the underlying risk factors rather than the level of resulting portfolio losses.
- If, instead, the parameters are chosen by *expert judgement*, then there are compatibility requirements. In order to make sense, the ρ_{ij} must form the elements of a positive-semidefinite correlation matrix. When a correlation matrix is pieced together from pairwise estimates it is quite easy to violate this condition, and the risk of this happening increases with dimension.

- If $L_1, \dots, L_d$ are believed to have a non-elliptical distribution, then the limited range of attainable correlations for each pair (L_i, L_j), as discussed in connection with Fallacy 2 in Section 7.2.2, is also a relevant constraint
- The use of (8.47) offers no obvious way to incorporate tail dependence between the losses into the calculation of aggregate capital.

It might be supposed that use of the summation formula (8.46) would avoid these issues with correlation and yield a conservative upper bound for aggregate capital for any possible underlying distribution of $(L_1, \dots, L_d)$. While this is true if the risk measure ϱ is a coherent risk measure, it is not true in general if ϱ is a non-subadditive risk measure, such as VaR. This is an example of Fallacy 3 in Section 7.2.2.

It is possible that the underlying multivariate loss model is one where, for some value of α, $\mathrm{VaR}_\alpha(L_1+\cdots+L_d) > \mathrm{VaR}_\alpha(L_1)+\cdots+\mathrm{VaR}_\alpha(L_d)$. In this case, if we set $\mathrm{EC}_i = \mathrm{VaR}_\alpha(L_i)-E(L_i)$ and take the sum $\mathrm{EC}_1 + \cdots + \mathrm{EC}_1$, this will underestimate the actual required capital $\mathrm{EC} = \mathrm{VaR}_\alpha(L) - E(L)$, where $L = L_1 + \cdots + L_d$.

In Section 8.4.4 we examine the problem of putting upper and lower bounds on aggregate capital when marginal distributions are known and marginal capital requirements are determined by the value-at-risk measure.

8.4.2 Aggregation Based on Stressing Risk Factors

Another situation where aggregation rules of the form (8.47) are used in practice is in the aggregation of capital contributions computed by stressing individual risk factors. An example of such an application is the standard formula approach to Solvency II (see, for example, CEIOPS 2006). Capital amounts $\mathrm{EC}_1, \dots, \mathrm{EC}_d$ are computed by examining the effects on the balance sheet of extreme changes in a number of key risk factors, and (8.47) is used to compute an overall capital figure that takes into account the dependence of the risk factors.

To understand when the use of (8.47) may be considered to be a principles-based approach to aggregation, suppose we write $\boldsymbol{x} = \boldsymbol{X}(\omega)$ for a scenario defined in terms of changes in fundamental risk factors and $L(\boldsymbol{x})$ for the corresponding loss. We assume that $L(\boldsymbol{x})$ is a known function and, for simplicity, that it is increasing in each component of $\boldsymbol{x}$. Following common practice, the d risk factors are stressed one at a time by predetermined amounts $k_1, \dots, k_d$. Capital contributions for each risk factor are set by computing

$$\mathrm{EC}_i = L(k_i \boldsymbol{e}_i) - L(E(X_i)\boldsymbol{e}_i), \tag{8.50}$$

where $\boldsymbol{e}_i$ denotes the ith unit vector and where $k_i > E(X_i)$ so that $\mathrm{EC}_i > 0$. The value EC_i can be thought of as the loss incurred by stressing risk factor i by an amount k_i relative to the impact of stressing it by its expected change, while all other risk factors are held constant. One possibility is that the size of the stress event is set at the level of the α-quantile of the distribution of X_i, so that $k_i = q_\alpha(X_i)$ for α close to 1. We now prove a simple result that justifies the use of the aggregation rule (8.47) to combine the contributions $\mathrm{EC}_1, \dots, \mathrm{EC}_d$ defined in (8.50) into an aggregate capital EC.

Proposition 8.30. *Let $\boldsymbol{X} \sim E_d(\boldsymbol{\mu}, \Sigma, \psi)$, with $E(\boldsymbol{X}) = \boldsymbol{\mu}$. Let $\mathcal{M}$ be the space of linear portfolios (8.34) and let ϱ be a positive-homogeneous, translation-invariant and law-invariant risk measure on $\mathcal{M}$. Then, for any $L = L(\boldsymbol{X}) = m + \boldsymbol{\lambda}'\boldsymbol{X} \in \mathcal{M}$ we have*

$$\varrho(L - E(L)) = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} \rho_{ij}\, \mathrm{EC}_i\, \mathrm{EC}_j}, \tag{8.51}$$

where $\mathrm{EC}_i = L(\varrho(X_i)\boldsymbol{e}_i) - L(E(X_i)\boldsymbol{e}_i)$ and ρ_{ij} is an element of $\wp(\Sigma)$.

Proof. We observe that $\mathrm{EC}_i = \lambda_i \varrho(X_i) - \lambda_i E(X_i) = \lambda_i \varrho(X_i - E(X_i))$, and (8.51) follows by application of Theorem 8.28 (3). □

Proposition 8.30 shows that, under a strong set of assumptions, we can aggregate the effects of single-risk-factor stresses to obtain an aggregate capital requirement that corresponds to application of any positive-homogeneous, translation-invariant and law-invariant risk measure to the distribution of the unexpected loss; this would apply to VaR, expected shortfall or one of the distortion risk measures of Section 8.2.1. It is this idea that underscores the use of (8.47) in Solvency II. However, the key assumptions are, once again, the linearity of losses in the risk-factor changes and the elliptical distribution of risk-factor changes, both of which are simplistic in real-world applications.

We can of course regard the use of (8.47) as a pragmatic, rules-based approach. The correlation parameters are defined at the level of the risk factors and, for typical market-risk factors such as returns on prices or rates, the data may be available to permit estimation of these parameters. For other risk factors, such as mortality and policy lapse rates in Solvency II applications, parameters may be set by expert judgement and the same issues mentioned in Section 8.4.1 apply. In particular, the matrix with components ρ_{ij} must be positive definite in order for the procedure to make any kind of sense.

The summation rule may once again appear to be a conservative rule that avoids the problems related to estimating and setting correlations. However, it should be noted that in the presence of non-linear relationships between losses and risk factors, there can be complex interactions between risk factors that would require even higher capital than indicated by the sum of losses due to single-risk-factor stresses (see Notes and Comments).

8.4.3 Modular versus Fully Integrated Aggregation Approaches

The approaches discussed in Sections 8.4.1 and 8.4.2 can be described as *modular* approaches to risk capital. The risk is computed in modules or *silos* and then aggregated. In Section 8.4.1 the modules are defined in terms of business units or asset classes; in Section 8.4.2 the modules are defined in terms of individual risk factors. The former approach is arguably more natural because the losses across asset classes and business units are additive and it is possible to remove risks from the enterprise by selling parts of the business. The risks due to fundamental underlying risk factors are more pervasive and may manifest themselves in different parts of the balance sheet; typically, their effects can be non-linear and they can only be reduced by hedging.

Regardless of the nature of the underlying silos the aggregation approaches we have described involve the specification of correlations and the use of (8.47) or its special case (8.46). We have observed that there are practical problems associated with choosing correlations, and in Chapter 7 we have argued that correlation gives only a partial description of a multivariate distribution and that copulas are a better approach to multivariate dependence modelling. It is natural to consider using copulas in aggregation.

In the set-up of Section 8.4.1, where the total loss is given by $L = L_1 + \cdots + L_d$ and the L_i are losses due to business units, suppose that we know, or can accurately estimate, the marginal distributions $F_1, \dots, F_d$ for each of the modules. This is a necessary prerequisite for computing the marginal capital requirements $\mathrm{EC}_i = \varrho(L_i) - E(L_i)$. Instead of aggregating these marginal capital figures with correlation, we could attempt to choose a suitable copula C and build a multivariate loss distribution $F(\boldsymbol{x}) = C(F_1(x_1), \dots, F_d(x_d))$ using the converse of Sklar's Theorem (7.3). This is referred to as the *margins-plus-copula* approach. Computation of aggregate capital would typically proceed by generating large numbers of multivariate losses from F, summing them to obtain simulated overall losses L, and then applying empirical quantile or shortfall estimation techniques.

The problem with this approach is the specification of C. Multivariate loss data from the business units may be sparse or non-existent, and expert judgment may have to be employed. This might involve deciding whether the copula should have a degree of tail dependence, taking a view on plausible levels of rank correlation between the pairs (L_i, L_j) and then using the copula calibration methods based on rank correlation described in Section 7.5.1. Clearly, this approach has as many, if not more, problems than choosing a correlation matrix to use in (8.47). It remains a modular approach in which we start with models for the individual L_i and add dependence assumptions as an overlay. In Section 8.4.4 we will address the issue of *dependence uncertainty* in such a margins-plus-copula approach; in particular, we will quantify the "best-to-worst" gaps in VaR and ES estimation if only the marginal dfs of the losses are known.

A more appealing approach, which we describe as a *fully integrated* approach, is to build multivariate models for the changes in underlying risk factors $\boldsymbol{X} = (X_1, \dots, X_k)'$ and for the functionals $g_i : \mathbb{R}^k \mapsto \mathbb{R}$ that give the losses $L_i = g_i(\boldsymbol{X})$, $i = 1, \dots, d$, for the different portfolios, desks or business units that make up the enterprise. It is generally easier to build multivariate models for underlying risk factors because more data exist at the level of the risk factors. The models for $\boldsymbol{X}$ may range in sophistication from margins-plus-copula distributional models to more dynamic, financial econometric models. They are often referred to as *economic scenario generators*. In the fully integrated approach, aggregate capital is derived by applying risk measures to the distribution of $L = g_1(\boldsymbol{X}) + \cdots + g_d(\boldsymbol{X})$, and the losses in business units L_i and L_j are implicitly dependent through their mutual dependence on $\boldsymbol{X}$.

8.4.4 Risk Aggregation and Fréchet Problems

In the margins-plus-copula approach to risk aggregation described in Section 8.4.3 a two-step procedure for the construction of a model for the total loss $L = L_1 + \cdots + L_d$ is followed.

(1) Find appropriate models (dfs) $F_1, \dots, F_d$ for the marginal risks $L_1, \dots, L_d$. These can be obtained by statistical fitting to historical data, or postulated a priori in a stress-testing exercise.

(2) Choose a suitable copula C resulting in a joint model $C(F_1, \dots, F_d)$ for the random vector $\boldsymbol{L} = (L_1, \dots, L_d)'$ from which the df for the total portfolio loss L can be derived.

Based on steps (1) and (2), any law-invariant risk measure $\varrho(L)$ can, in principle, be calculated. The examples we will concentrate on in this section are $\varrho = \mathrm{VaR}_\alpha$ and $\varrho = \mathrm{ES}_\alpha$. Note that there is nothing special about the sum structure of the portfolio L; more general portfolios (or financial positions) $L = \Psi(L_1, \dots, L_d)$ for suitable functions $\Psi : \mathbb{R}^d \to \mathbb{R}$ could also be considered.

As mentioned in Section 8.4.3, there is a lot of model uncertainty surrounding the choice of the appropriate copula in step (2). In this section we will therefore drop step (2) while retaining step (1). Clearly, this means that quantities such as $\mathrm{VaR}_\alpha(L) = \mathrm{VaR}_\alpha(L_1 + \cdots + L_d)$ can no longer be computed precisely due to the lack of a fully specified model for the vector $\boldsymbol{L}$ and hence the aggregate loss L.

Instead we will try to find bounds for $\mathrm{VaR}_\alpha(L)$ given *only* the marginal information from step (1). Problems of this type are known as *Fréchet problems* in the literature. In this section we refer to the situation where only marginal information is available as *dependence uncertainty*.

In order to derive bounds we introduce the class of rvs

$$\begin{aligned}\mathcal{S}_d &= \mathcal{S}_d(F_1, \dots, F_d) \\ &= \left\{ L = \sum_{i=1}^{d} L_i : L_1, \dots, L_d \text{ rvs with } L_i \sim F_i,\ i = 1, \dots, d \right\}.\end{aligned}$$

Clearly, every element of $\mathcal{S}_d$ is a feasible risk position satisfying step (1). The problem of finding VaR bounds under dependence uncertainty now reduces to finding

$$\overline{\mathrm{VaR}}_\alpha(\mathcal{S}_d) = \sup\{\mathrm{VaR}_\alpha(L) : L \in \mathcal{S}_d(F_1, \dots, F_d)\}$$

and

$$\underline{\mathrm{VaR}}_\alpha(\mathcal{S}_d) = \inf\{\mathrm{VaR}_\alpha(L) : L \in \mathcal{S}_d(F_1, \dots, F_d)\}.$$

We will use similar notation if VaR_α is replaced by another risk measure ϱ; for instance, we will write $\overline{\mathrm{ES}}_\alpha$ and $\underline{\mathrm{ES}}_\alpha$. In our main case of interest, when $L = L_1 + \cdots + L_d$ with the L_i variables as in step (1), we will often write $\varrho(\mathcal{S}_d) = \varrho(L)$ and use similar notation for the corresponding upper and lower bounds. For expected shortfall, which is a coherent and comonotone-additive risk measure (see Definition 8.19), we have

$$\overline{\mathrm{ES}}_\alpha(L) = \sum_{i=1}^{d} \mathrm{ES}_\alpha(L_i),$$

and we see that the upper bound is achieved under comonotonicity. We often refer to $\underline{\varrho}$ as the *best* and $\overline{\varrho}$ as the *worst* ϱ; this interpretation depends, of course, on the context.

The calculation of $\overline{\mathrm{VaR}}_\alpha(\mathcal{S}_d)$, $\underline{\mathrm{VaR}}_\alpha(\mathcal{S}_d)$ and $\underline{\mathrm{ES}}_\alpha(\mathcal{S}_d)$ is difficult in general. We will review some of the main results without proof; further references on this very active research area can be found in Notes and Comments. The available results very much depend on the dimension ($d = 2$ versus $d > 2$) and whether the portfolio is homogeneous ($F_1 = \cdots = F_d$) or not. We begin with a result for the case $d = 2$.

Proposition 8.31 (VaR, $d = 2$). *Under the set-up above,* $\forall \alpha \in (0, 1)$,

$$\overline{\mathrm{VaR}}_\alpha(\mathcal{S}_2) = \inf_{x\in[0,1-\alpha]} \{F_1^{-1}(\alpha + x) + F_2^{-1}(1 - x)\}$$

and

$$\underline{\mathrm{VaR}}_\alpha(\mathcal{S}_2) = \inf_{x\in[0,\alpha]} \{F_1^{-1}(x) + F_2^{-1}(\alpha - x)\}.$$

Proof. See Makarov (1981) and Rüschendorf (1982). □

From the above proposition we already see that the optimal *couplings*—the dependence structures achieving the VaR bounds—combine large outcomes in one risk with small outcomes in the other. Next we give VaR bounds for higher dimensions, assuming a homogeneous portfolio.

Proposition 8.32 (VaR, $d \geqslant 2$, homogeneous case). *Suppose that* $F := F_1 = \cdots = F_d$ *and that for some* $b \in \mathbb{R}$ *the density function* f *of* F *(assumed to exist) is decreasing on* $[b, \infty)$. *Then, for* $\alpha \in [F(b), 1)$ *and* $X \sim F$,

$$\overline{\mathrm{VaR}}_\alpha(\mathcal{S}_d) = dE(X \mid X \in [F^{-1}(\alpha + (d - 1)c, F^{-1}(1 - c)]), \tag{8.52}$$

where c *is the smallest number in* $[0, (1 - \alpha)/d]$ *such that*

$$\int_{\alpha+(d-1)c}^{1-c} F^{-1}(t)\,\mathrm{d}t \geqslant \frac{1 - \alpha - dc}{d}((d - 1)F^{-1}(\alpha + (d - 1)c) + F^{-1}(1 - c)).$$

If the density f *of* F *is decreasing on its support, then for* $\alpha \in (0, 1)$ *and* $X \sim F$,

$$\underline{\mathrm{VaR}}_\alpha(\mathcal{S}_d) = \max\{(d - 1)F^{-1}(0) + F^{-1}(\alpha), dE(X \mid X \leqslant F^{-1}(\alpha))\}. \tag{8.53}$$

Proof. For the proof of (8.52) see Wang, Peng and Yang (2013). The case (8.53) follows by symmetry arguments (see Bernard, Jiang and Wang 2014). □

Remark 8.33. First of all note the extra condition on the density f of F that is needed to obtain the sharp bound (8.53) for $\underline{\mathrm{VaR}}_\alpha(\mathcal{S}_d)$: we need f to be decreasing on its *full* support rather than only on a certain tail region, which is sufficient for (8.52). As a consequence, both (8.52) and (8.53) can be applied, for instance, to the case where F is Pareto, but for lognormal rvs only (8.52) applies.

Though the results (8.52) and (8.53) look rather involved, they exhibit an interesting structure. As in the case where $d = 2$, the extremal couplings combine large and small values of the underlying df F. More importantly, if $c = 0$, then (8.52) reduces to

$$\overline{\mathrm{VaR}}_\alpha(\mathcal{S}_d) = \overline{\mathrm{ES}}_\alpha(\mathcal{S}_d). \tag{8.54}$$

The extremal coupling for VaR is rather special and differs from the extremal coupling for ES, which is of course comonotonicity. The condition $c = 0$ corresponds to the crucial notion of d-mixability (see Definition 8.35).

The observation (8.54) is relevant to a discussion of the pros and cons of value-at-risk versus expected shortfall and the regulatory debate surrounding these risk measures. Although the upper bounds coincide in the case $c = 0$, it is much easier to compute $\overline{\mathrm{ES}}_\alpha(\mathcal{S}_d)$ due to the comonotone additivity of expected shortfall (see Embrechts et al. (2014) and Notes and Comments).

Similar to Proposition 8.32, a sharp bound for the best ES case for a homogeneous portfolio can be given, and this also requires a strong monotonicity condition for the underlying density. Here the *lower expected shortfall* risk measure LES_α enters; for $\alpha \in (0, 1)$ this is defined to be

$$\mathrm{LES}_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha \mathrm{VaR}_u(X)\,\mathrm{d}u = -\mathrm{ES}_{1-\alpha}(-X).$$

Proposition 8.34 (ES, $d \geqslant 2$, homogeneous case). *Suppose that $F = F_1 = \cdots = F_d$, that F has a finite first moment and that the density function of F (which is assumed to exist) is decreasing on its support. Then, for $\alpha \in [1 - dc, 1)$, $\beta = (1-\alpha)/d$ and $X \sim F$,*

$$\begin{aligned}\underline{\mathrm{ES}}_\alpha(\mathcal{S}_d) &= \frac{1}{\beta}\int_0^\beta ((d-1)F^{-1}((d-1)t) + F^{-1}(1-t))\,\mathrm{d}t \\ &= (d-1)^2\,\mathrm{LES}_{(d-1)\beta}(X) + \mathrm{ES}_{1-\beta}(X),\end{aligned} \tag{8.55}$$

where c is the smallest number in $[0, 1/d]$ such that

$$\int_{(d-1)c}^{1-c} F^{-1}(d)\,\mathrm{d}t \geqslant \frac{1-dc}{d}((d-1)F^{-1}((d-1)c) + F^{-1}(1-c)).$$

Proof. See Bernard, Jiang and Wang (2014). □

An important tool in proofs of these results is a general concept of multivariate *negative dependence* known as *mixability*, which is introduced next.

Definition 8.35. A df F on $\mathbb{R}$ is called *d-completely mixable* (d-CM) if there exist d rvs $X_1, \dots, X_d \sim F$ such that for some $k \in \mathbb{R}$,

$$P(X_1 + \dots + X_d = dk) = 1. \tag{8.56}$$

F is *completely mixable* if F is d-CM for all $d \geqslant 2$. The dfs $F_1, \dots, F_d$ on $\mathbb{R}$ are called *jointly mixable* if there exist d rvs $X_i \sim F_i$, $i = 1, \dots, d$, such that for some $c \in \mathbb{R}$,

$$P(X_1 + \dots + X_d = c) = 1.$$

Clearly, if F has finite-mean μ, we must have $k = \mu$ in (8.56). Complete mixability is a concept of strong negative dependence. It is indeed this dependence structure that yields the extremal couplings in Proposition 8.32. The above definition of d-complete mixability and its link to dependence-uncertainty problems can be found in Wang and Wang (2011). Examples of completely mixable dfs are the normal, Student t, Cauchy and uniform distributions. In Rüschendorf and Uckelmann (2002) it was shown that any continuous distribution function with a symmetric and unimodal density is d-completely mixable for any $d \geqslant 2$. See Notes and Comments for a historical perspective and further references.

In contrast to the above analytic results for the homogeneous case, very little is known for non-homogeneous portfolios, i.e. for portfolios where the condition $F_1 = \dots = F_d$ does not hold. In general, however, there is a fast and efficient numerical procedure for solving dependence-uncertainty problems that is called the *rearrangement algorithm* (RA) (see Embrechts, Puccetti and Rüschendorf 2013). The RA was originally worked out for the calculation of best/worst VaR bounds; it can be generalized to other risk measures like expected shortfall. Mathematically, the RA is based on the above idea of mixability. For instance, for the calculation of $\overline{\mathrm{VaR}}_\alpha(L)$, one discretizes the $(1-\alpha)100\%$ upper tail of the underlying factor dfs $F_1, \dots, F_d$, using $N = 100\,000$ bins, say. For the dimension d, values around and above 1000, say, can easily be handled by the RA. It can similarly be used for the calculation of $\underline{\mathrm{VaR}}_\alpha(L)$ and $\underline{\mathrm{ES}}_\alpha(L)$ in both the homogeneous and non-homogeneous cases.

With the results discussed above, including the RA, we can now calculate several quantities related to diversification, (non-)coherence, and model and dependence uncertainty. We restrict our attention to the additive portfolio $L = L_1 + \dots + L_d$ under the set-up in step (1). It will be useful to consider several functions $\mathcal{X} \colon \mathbb{R}^2 \to \mathbb{R}$ that compare risk measures under different dependence assumptions. Examples encountered in the literature are the following.

Super/subadditivity indices. $a = \mathrm{VaR}_\alpha(L)$, $b = \mathrm{VaR}_\alpha^+(L)$, and $\mathcal{X}_1(a,b) = a/b$, $\mathcal{X}_2(a,b) = 1 - (a/b)$, $\mathcal{X}_3(a,b) = b - a$. $\mathrm{VaR}_\alpha^+(L)$ denotes the comonotonic case, i.e. $\mathrm{VaR}_\alpha^+(L) = \sum_{i=1}^d \mathrm{VaR}_\alpha(L_i)$.

Worst superadditivity ratio. $a = \overline{\mathrm{VaR}}_\alpha(L)$, $b = \mathrm{VaR}_\alpha^+(L)$, and $\mathcal{X}_4(a,b) = a/b$; the case is similar for the *best superadditivity ratio*, replacing $\overline{\mathrm{VaR}}_\alpha(L)$ by $\underline{\mathrm{VaR}}_\alpha(L)$.

Dependence-uncertainty spread. Either $a = \underline{\mathrm{VaR}}_{\alpha}(L)$, $b = \overline{\mathrm{VaR}}_{\alpha}(L)$ or $a = \underline{\mathrm{ES}}_{\alpha}(L)$, $b = \overline{\mathrm{ES}}_{\alpha}(L)$ and $\mathcal{X}_5(a, b) = b - a$. A further interesting measure compares the VaR dependence-uncertainty spread with the ES dependence-uncertainty spread.

Best/worst (VaR, ES) ratios. Either $a = \underline{\mathrm{VaR}}_{\alpha}(L), b = \underline{\mathrm{ES}}_{\alpha}(L)$ or $a = \overline{\mathrm{VaR}}_{\alpha}(L)$, $b = \overline{\mathrm{ES}}_{\alpha}(L)$ and $\mathcal{X}_6(a, b) = b/a$, say.

As we already observed in (8.54), whenever $c = 0$ in Proposition 8.32 we have that $\overline{\mathrm{VaR}}_{\alpha}(\mathcal{S}_d) = \overline{\mathrm{ES}}_{\alpha}(\mathcal{S}_d)$. The following result extends this observation in an asymptotic way.

Proposition 8.36 (asymptotic equivalence of $\overline{\mathbf{ES}}$ and $\overline{\mathbf{VaR}}$). *Suppose that* $L_i \sim F_i$, $i \geqslant 1$, *and that*

(i) for some $k > 1$, $E(|L_i - E(L_i)|^k)$ *is uniformly bounded, and*

(ii) for some $\alpha \in (0, 1)$,

$$\liminf_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \mathrm{ES}_{\alpha}(L_i) > 0.$$

Then, as $d \to \infty$,

$$\frac{\overline{\mathrm{ES}}_{\alpha}(\mathcal{S}_d)}{\overline{\mathrm{VaR}}_{\alpha}(\mathcal{S}_d)} = 1 + O(d^{(1/k)-1}). \tag{8.57}$$

Proof. See Embrechts, Wang and Wang (2014). □

Proposition 8.36 shows that under very general assumptions typically encountered in QRM practice, we have that for d large, $\overline{\mathrm{VaR}}_{\alpha}(L) \approx \overline{\mathrm{ES}}_{\alpha}(L)$. The proposition also provides a rate of convergence. From numerical examples it appears that these asymptotic results hold fairly accurately even for small to medium values of the portfolio dimension d (see Example 8.40). From the same paper (Embrechts, Wang and Wang 2014) we add a final result related to the VaR and ES dependence-uncertainty spreads.

Proposition 8.37 (dependence-uncertainty spread of VaR versus ES). *Take* $0 < \alpha_1 \leqslant \alpha_2 < 1$ *and assume that the dfs* F_i, $i \geqslant 1$, *satisfy condition (i) of Proposition 8.36 as well as*

(iii) $\displaystyle\liminf_{d\to\infty} \frac{1}{d} \sum_{i=1}^{d} \mathrm{LES}_{\alpha_1}(X_i) > 0$ *and*

(iv) $\displaystyle\limsup_{d\to\infty} \frac{\sum_{i=1}^{d} E(X_i)}{\sum_{i=1}^{d} \mathrm{ES}_{\alpha_1}(X_i)} < 1.$

Then

$$\liminf_{d\to\infty} \frac{\overline{\mathrm{VaR}}_{\alpha_2}(\mathcal{S}_d) - \underline{\mathrm{VaR}}_{\alpha_2}(\mathcal{S}_d)}{\overline{\mathrm{ES}}_{\alpha_1}(\mathcal{S}_d) - \underline{\mathrm{ES}}_{\alpha_1}(\mathcal{S}_d)} \geqslant 1. \tag{8.58}$$

Proof. See Embrechts, Wang and Wang (2014). □

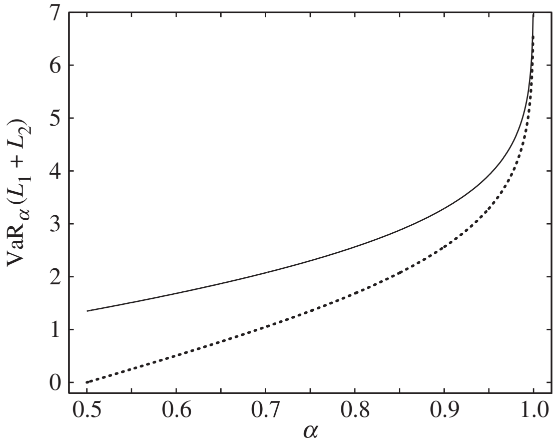

Figure 8.3. The worst-case $\overline{\mathrm{VaR}}_\alpha(L)$ (solid line) plotted against α for two standard normal risks; the case of comonotonic risks $\mathrm{VaR}_\alpha^+(L)$ is shown as a dotted line for comparison.

Remark 8.38. Propositions 8.36 and 8.37 are relevant to the ongoing discussion of risk measures for the calculation of regulatory capital. Recall that under the Basel framework for banking and also the Solvency II framework for insurance, VaR-based capital requirements are to be compared and contrasted with those based on expected shortfall. In particular, comparisons are made between VaR and ES at different quantiles, e.g. between $\mathrm{VaR}_{0.99}$ and $\mathrm{ES}_{0.975}$. The above propositions add a component of dependence uncertainty to these discussions. In particular, the dependence-uncertainty spread of VaR is generally larger than that of ES. For a numerical illustration of this, consider Example 8.40 below.

Examples. We consider examples where the aggregate loss is given by $L = L_1 + \cdots + L_d$. For any risk measure ϱ we denote by $\varrho^+(L)$ the value of ϱ when $L_1, \dots, L_d$ are comonotonic, and we write $\varrho^\perp(L)$ when they are independent. In a first example we consider the case when $d = 2$ and $F_1 = F_2 = \Phi$, the standard normal df. In Example 8.40, higher-dimensional portfolios with Pareto margins are considered.

Example 8.39 (worst VaR for a portfolio with normal margins). For $i = 1, 2$ let $F_i = \Phi$. In Figure 8.3 we have plotted $\overline{\mathrm{VaR}}_\alpha(L)$ calculated using Proposition 8.31 as a function of α together with the curve corresponding to the comonotonic case $\mathrm{VaR}^+(L)$ calculated using Proposition 7.20. The fact that the former lies above the latter implies the existence of portfolios with normal margins for which VaR is not subadditive. For example, for $\alpha = 0.95$, the upper bound is 3.92, whereas $\mathrm{VaR}_\alpha(L_i) = 1.645$, so, for the worst VaR portfolio, $\overline{\mathrm{VaR}}_{0.95}(L) = 3.92 > 3.29 = \mathrm{VaR}_{0.95}(L_1) + \mathrm{VaR}_{0.95}(L_2)$. The density function of the distribution of (L_1, L_2) that leads to the $\overline{\mathrm{VaR}}_\alpha(L)$ is shown in Figure 8.4 (see Embrechts, Höing and Puccetti (2005) for further details).

Example 8.40 (VaR and ES bounds for Pareto margins). In Tables 8.1 and 8.2 we have applied the various results to a homogeneous Pareto case where $L_i \sim \mathrm{Pa}(\theta, 1)$,

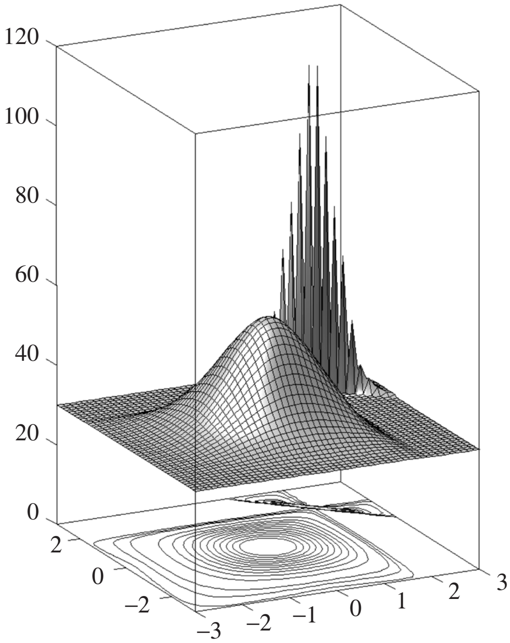

Figure 8.4. Contour and perspective plots of the density function of the distribution of (L_1, L_2) leading to the worst-case $\overline{\mathrm{VaR}}_\alpha(L)$ for $L = L_1 + L_2$ at the $\alpha = 0.95$ level when the L_i are standard normal.

$i = 1, \dots, d$, so that the common df is $F(x) = 1 - (1 + x)^{-\theta}$, $x \geqslant 0$ (see also Section A.2.8 in the appendix). In Table 8.1 we consider cases where $\theta \geqslant 3$, corresponding to finite-variance distributions; in Table 8.2 we consider cases where $\theta \leqslant 2$, corresponding to infinite-variance distributions. The values $d = 8$ and $d = 56$ are chosen with applications to operational risk in mind. In that context $d = 8$ corresponds to a relatively low-dimensional aggregation problem and $d = 56$ to a moderately high-dimensional aggregation problem (see Chapter 13).

We use the analytic results from Propositions 8.32 and 8.34, as well as the RA. For the independent case, simulation is used. The figures given are appropriately rounded. For the homogeneous case we only report the analytic bounds (with a numerical root search for c in the propositions). The RA bounds are close to identical to their analytical counterparts, with only very small deviations for heavy-tailed dfs, i.e. for small θ. We note that for heavy-tailed risks, the RA requires a fine discretization, and hence considerably more time is needed to calculate $\underline{\mathrm{ES}}_\alpha(L)$.

Both tables confirm the results discussed above, i.e. $\overline{\mathrm{ES}}_\alpha(L)/\overline{\mathrm{VaR}}_\alpha(L)$ is close to 1 even for $d = 8$; the dependence-uncertainty spreads behave as stated in Proposition 8.37, and finally, in the $\mathrm{Pa}(\theta, 1)$, $\theta > 1$, case we have that

$$\lim_{\alpha\uparrow 1} \frac{\overline{\mathrm{ES}}_\alpha(L)}{\mathrm{VaR}^+_\alpha(L)} = \frac{\theta}{\theta - 1},$$

which can be observed in the examples given above, though the convergence is much slower here. The latter result holds more generally for distributions with regularly

Table 8.1. ES and VaR bounds for $L = L_1 + \cdots + L_d$, where $L_i \sim \mathrm{Pa}(\theta, 1), i = 1, \dots, d$, with df $F(x) = 1 - (1 + x)^{-\theta}$, $x \geqslant 0$, for $\theta \geqslant 3$. See Example 8.40 for a discussion.

	$d = 8$			$d = 56$		
$\theta = 10$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$
$\overline{\mathrm{VaR}}_\alpha$	4.0	6.1	9.7	28.0	42.6	68.1
VaR_α^+	2.8	4.7	8.0	19.6	32.8	55.7
$\mathrm{VaR}_\alpha^\perp$	1.5	1.9	2.5	7.8	8.6	9.6
$\underline{\mathrm{VaR}}_\alpha$	0.7	0.8	1.0	5.1	5.9	6.2
$\overline{\mathrm{ES}}_\alpha$	4.0	6.1	9.7	28.0	42.6	68.1
$\mathrm{ES}_\alpha^\perp$	1.8	2.2	2.8	8.3	9.1	10.0
$\underline{\mathrm{ES}}_\alpha$	0.9	1.2	1.7	6.2	6.2	6.2

	$d = 8$			$d = 56$		
$\theta = 5$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$
$\overline{\mathrm{VaR}}_\alpha$	10.2	17.1	31.7	71.4	119.8	222.7
VaR_α^+	6.6	12.1	23.8	46.0	84.7	166.9
$\mathrm{VaR}_\alpha^\perp$	3.7	5.0	7.2	18.3	20.7	24.1
$\underline{\mathrm{VaR}}_\alpha$	1.6	1.8	3.0	11.0	12.9	13.8
$\overline{\mathrm{ES}}_\alpha$	10.2	17.1	31.8	71.4	119.8	222.7
$\mathrm{ES}_\alpha^\perp$	4.5	5.9	8.7	19.8	22.2	26.0
$\underline{\mathrm{ES}}_\alpha$	2.5	3.8	6.5	14.0	14.0	14.3

	$d = 8$			$d = 56$		
$\theta = 3$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$
$\overline{\mathrm{VaR}}_\alpha$	24.1	46.9	110.2	171.9	333.7	783.7
VaR_α^+	13.7	29.1	72.0	96.0	203.9	504.0
$\mathrm{VaR}_\alpha^\perp$	8.1	12.3	23.0	39.1	47.6	67.2
$\underline{\mathrm{VaR}}_\alpha$	2.9	3.6	9.0	20.4	24.9	27.2
$\overline{\mathrm{ES}}_\alpha$	24.6	47.7	112.0	172.0	333.9	784.0
$\mathrm{ES}_\alpha^\perp$	11.0	17.0	32.9	44.9	56.2	85.5
$\underline{\mathrm{ES}}_\alpha$	7.2	12.9	29.0	28.6	31.3	56.4

varying tails (see Definition 5.7 and Karamata's Theorem (Appendix A.1.4) and recall that this includes distributions like the Student t and loggamma distributions).

Table 8.2 also includes the case $\theta = 0.8$, i.e. an infinite-mean case (for which ES is not defined). Here we note that $\mathrm{VaR}_\alpha^\perp(L) > \mathrm{VaR}_\alpha^+(L)$, so this gives an example of superadditivity of VaR_α in the case of independence (see the discussion following Example 2.25).

Notes and Comments

The use of a standard formula approach based on the kind of aggregation embodied in (8.47) is permitted under Solvency II; see CEIOPS (2006), a document produced

Table 8.2. ES and VaR bounds for $L = L_1 + \cdots + L_d$, where $L_i \sim \mathrm{Pa}(\theta, 1), i = 1, \dots, d$, with df $F(x) = 1 - (1 + x)^{-\theta}, x \geqslant 0$, for $\theta \leqslant 2$. See Example 8.40 for a discussion.

	$d = 8$			$d = 56$		
$\theta = 2$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$
$\overline{\mathrm{VaR}}_\alpha$	59	142	465	440	1 054	3 454
VaR_α^+	28	72	245	194	504	1 715
$\mathrm{VaR}_\alpha^\perp$	18	35	96	89	132	293
$\underline{\mathrm{VaR}}_\alpha$	5	9	31	36	46	53
$\overline{\mathrm{ES}}_\alpha$	64	152	498	445	1 064	3 486
$\mathrm{ES}_\alpha^\perp$	31	63	184	123	205	518
$\underline{\mathrm{ES}}_\alpha$	24	56	178	75	149	472
	$d = 8$			$d = 56$		
$\theta = 1.5$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$
$\overline{\mathrm{VaR}}_\alpha$	135	409	1 928	1 100	3 323	15 629
VaR_α^+	51	164	792	357	1 150	5 544
$\mathrm{VaR}_\alpha^\perp$	39	98	413	207	421	1 574
$\underline{\mathrm{VaR}}_\alpha$	8	21	99	56	77	99
$\overline{\mathrm{ES}}_\alpha$	169	509	2 392	1 182	3 563	16 744
$\mathrm{ES}_\alpha^\perp$	98	265	1 159	419	1 016	4 126
$\underline{\mathrm{ES}}_\alpha$	88	258	1 199	323	945	4 390
	$d = 8$			$d = 56$		
$\theta = 0.8$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$	$\alpha = 95\%$	$\alpha = 99\%$	$\alpha = 99.9\%$
$\overline{\mathrm{VaR}}_\alpha$	2 250	16 873	300 182	35 168	263 301	4 683 172
VaR_α^+	330	2 522	44 979	2 313	17 653	314 855
$\mathrm{VaR}_\alpha^\perp$	620	4 349	75 877	7 318	49 858	862 855
$\underline{\mathrm{VaR}}_\alpha$	41	315	5 622	207	433	5 622

by the Committee of European Insurance and Operational Pensions Supervisors (now EIOPA).

In the banking context the summation approach in (8.46) is commonly used, particularly for the aggregation of capital requirements for market and credit risk. As explained by Breuer et al. (2010) this is commonly justified by assuming that credit risk arises from the banking book and market risk from the trading book, but they point out that, for derivative instruments depending on both market and credit risks, it can potentially lead to underestimation of risk. In contrast, Alessandri and Drehmann (2010) study integration of credit risk and interest-rate risk in the banking book and conduct simulations suggesting that summation of capital for the two risk types is likely to be too conservative; Drehmann, Sorenson and Stringa (2010) argue that credit and interest-rate risk must be assessed jointly in the banking book.

Kretzschmar, McNeil and Kirchner (2010) make similar points about the importance of developing fully integrated models rather than modular approaches.

A summary of methodological practice in economic capital models before the 2007 crisis is presented in a comprehensive survey by the International Financial Risk Institute that included both banks and insurance companies. In this survey, the prevailing approach to integration is reported to be the use of correlation matrices (see IFRI Foundation and CRO Forum 2007). This approach was favoured by over 75% of the surveyed banks, with the others using simulation approaches based on scenario generation or hybrid approaches. In the insurance industry there was more diversity in the approaches used for integration: around 35% of respondents used the correlation approach and about the same number used simulation; the remainder reported the use of copulas or hybrid approaches.

There is a large literature on Fréchet problems: see, for instance, Chapter 2 in Rüschendorf (2013). From a QRM perspective, Embrechts and Puccetti (2006) gave the field a considerable boost. The latter paper also contains the most important references to the early literature. Historically, the question of bounding the df of a sum of rvs with given marginals goes back to Kolmogorov and was answered by Makarov (1981) for $d = 2$. Frank, Nelsen and Schweizer (1987) restated Makarov's result using the notion of a copula. Independently, Rüschendorf (1982) gave a very elegant proof of the same result using duality. Williamson and Downs (1990) introduced the use of dependence information. Numerous other authors (especially in analysis and actuarial mathematics) have contributed to this area. Besides the comprehensive book by Müller and Stoyan (2002), several other texts in actuarial mathematics contain interesting contributions on dependence modelling: for an introduction, see Chapter 10 in Kaas et al. (2001). A rich set of optimization problems within an actuarial context are to be found in De Vylder (1996); see especially "Part II: Optimization Theory", where the author "shows how to obtain best upper and lower bounds on functionals $T(F)$ of the df F of a risk, under moment or other integral constraints". An excellent account is to be found in Denuit and Charpentier (2004). The definitive account from an actuarial point of view is Denuit et al. (2005). A wealth of actuarial examples is to be found in the two extensive articles Hürlimann (2008a) and Hürlimann (2008b).

The rearrangement algorithm (RA) for VaR appeared in Embrechts, Puccetti and Rüschendorf (2013) and was based on earlier work by Puccetti and Rüschendorf (2012). Full details on the RA are collected by Giovanni Puccetti at https://sites.google.com/site/rearrangementalgorithm/. The interested reader may also search the literature for probability box (or p-box) and the related *Dempster–Shafer Theory*. These search items lead to well-established theory and numerous examples in the realm of engineering, computer science and economics. For expected shortfall, the RA was worked out in Puccetti (2013). For the analytical results and a discussion of the sharpness of the various bounds for VaR and ES, the papers cited for the corresponding propositions give an excellent introduction. Some further interesting papers are Bernard, Jiang and Wang (2014), Bernard et al. (2013) and Bernard, Rüschendorf and Vanduffel (2013).

The notion of complete mixability leads to a condition of negative dependence for multivariate ($d \geqslant 2$) random vectors. Recent developments in this field are summarized in Puccetti and Wang (2015), Puccetti and Wang (2014) and Wang and Wang (2014).

Rosenberg and Schuermann (2006) gives some idea of the applicability of aggregation ideas used in this chapter. The authors construct the joint risk distribution for a typical, large, internationally active bank using the method of copulas and aggregate risk measures across the categories of market, credit and operational risk.

For an illustration of the ideas and results of Section 8.4.4 in the practical environment of a Norwegian financial group, see Dimakos and Aas (2004) and Aas and Puccetti (2014). The latter paper contains a discussion of the best/worst couplings. See also Embrechts, Puccetti and Rüschendorf (2013) on this topic.

8.5 Capital Allocation

The final section of this chapter essentially looks at the converse problem to Section 8.4. Given a model for aggregate losses we now consider how the overall capital requirement may be disaggregated into additive contributions attributable to the different sub-portfolios or assets that make up the overall portfolio.

8.5.1 The Allocation Problem

As in Section 8.4.1 let the rvs $L_1, \dots, L_d$ represent the losses (or negative P&Ls) arising from d different lines of business, or the losses corresponding to d different asset classes on the balance sheet of a firm. In this section we will refer to these sub-units of a larger portfolio simply as investments. The allocation problem can be motivated by considering the question of how we might measure the risk-adjusted performance of different investments within a portfolio.

The performance of investments is usually measured using a RORAC (return on risk-adjusted capital) approach, i.e. by considering a ratio of the form

$$\frac{\text{expected profit of investment } i}{\text{risk capital for investment } i}. \tag{8.59}$$

The general approach embodied in (8.59) raises the question of how we should calculate the risk capital for an investment that is part of a larger portfolio. It should not simply be the stand-alone risk capital for that investment considered in isolation; this would neglect the issue of diversification and give an inaccurate measure of the performance of an investment within the larger portfolio. Instead, the risk capital for an investment within a portfolio should reflect the contribution of that investment to the overall riskiness of the portfolio. A two-step procedure for determining these contributions is used in practice.

(1) Compute the overall risk capital $\varrho(L)$, where $L = \sum_{i=1}^{d} L_i$ and ϱ is a particular risk measure such as VaR, ES or a mean-adjusted version of one of these (see (8.48)); note that at this stage we are not stipulating that ϱ must be coherent.

(2) Allocate the capital $\varrho(L)$ to the individual investments according to some mathematical *capital allocation principle* such that, if AC_i denotes the capital allocated to the investment with potential loss L_i (the so-called *risk contribution* of unit i), the sum of the risk contributions corresponds to the overall risk capital $\varrho(L)$.

In this section we are interested in step (2) of the procedure; loosely speaking, we require a mapping that takes as input the individual losses $L_1, \dots, L_d$ and the risk measure ϱ and yields as output the vector of risk contributions $(\mathrm{AC}_1, \dots, \mathrm{AC}_d)$ such that

$$\varrho(L) = \sum_{i=1}^{d} \mathrm{AC}_i, \tag{8.60}$$

and such a mapping will be called a capital allocation principle. The relation (8.60) is sometimes called the *full allocation property* since all of the overall risk capital $\varrho(L)$ (not more, not less) is allocated to the investments; we consider this property to be an integral part of the definition of an allocation principle. Of course, there are other properties of a capital allocation principle that are desirable from an economic viewpoint; we first make some formal definitions and give examples of allocation principles before discussing further properties.

The formal set-up. Let $L_1, \dots, L_d$ be rvs on a common probability space $(\Omega, \mathcal{F}, P)$ representing losses (or profits) for d investments. For our discussion it will be useful to consider portfolios where the weights of the individual investments are varied with respect to our basic portfolio $(L_1, \dots, L_d)$, which is regarded as a fixed random vector. That is, we consider an open set $\Lambda \subset \mathbb{R}^d \setminus \{\mathbf{0}\}$ of portfolio weights such that $\mathbf{1} \in \Lambda$ and define for $\boldsymbol{\lambda} \in \Lambda$ the loss $L(\boldsymbol{\lambda}) = \sum_{i=1}^{d} \lambda_i L_i$; the loss of our actual portfolio is of course $L(\mathbf{1})$. Let ϱ be some risk measure defined on a set $\mathcal{M}$ that contains the rvs $\{L(\boldsymbol{\lambda}) \colon \boldsymbol{\lambda} \in \Lambda\}$. As in Section 8.3.1 we use the associated risk-measure function $r_\varrho \colon \Lambda \to \mathbb{R}$ with $r_\varrho(\boldsymbol{\lambda}) = \varrho(L(\boldsymbol{\lambda}))$.

8.5.2 The Euler Principle and Examples

From now on we restrict our attention to risk measures that are positive homogeneous. This may be a coherent risk measure, or a mean-adjusted version of a coherent risk measure as in (8.48); it may also be VaR (or a mean-corrected version of VaR) or the standard deviation risk measure. Obviously, the associated risk-measure function must satisfy $r_\varrho(t\boldsymbol{\lambda}) = t r_\varrho(\boldsymbol{\lambda})$ for all $t > 0$, $\boldsymbol{\lambda} \in \Lambda$, so $r_\varrho \colon \Lambda \to \mathbb{R}$ is a positive-homogeneous function of a vector argument. Recall Euler's well-known rule that states that if r_ϱ is positive homogeneous and differentiable at $\boldsymbol{\lambda} \in \Lambda$, we have

$$r_\varrho(\boldsymbol{\lambda}) = \sum_{i=1}^{d} \lambda_i \frac{\partial r_\varrho}{\partial \lambda_i}(\boldsymbol{\lambda}). \tag{8.61}$$

If we apply this at $\boldsymbol{\lambda} = \mathbf{1}$, we get, using that $\varrho(L) = r_\varrho(\mathbf{1})$,

$$\varrho(L) = \sum_{i=1}^{d} \frac{\partial r_\varrho}{\partial \lambda_i}(\mathbf{1}).$$

This suggests the following definition.

Definition 8.41 (Euler capital allocation principle). If r_ϱ is a positive-homogeneous risk-measure function, which is differentiable at $\boldsymbol{\lambda} = \mathbf{1}$, then the Euler capital allocation principle associated with ϱ has risk contributions

$$\mathrm{AC}_i^\varrho = \frac{\partial r_\varrho}{\partial \lambda_i}(\mathbf{1}), \quad 1 \leqslant i \leqslant d.$$

The Euler principle is sometimes called *allocation by the gradient*, and it obviously gives a full allocation of the risk capital. We now look at a number of specific examples of Euler allocations corresponding to different choices of risk measure ϱ.

Standard deviation and the covariance principle. Consider the risk-measure function $r_{\mathrm{SD}}(\boldsymbol{\lambda}) = \sqrt{\mathrm{var}(L(\boldsymbol{\lambda}))}$ and write Σ for the covariance matrix of $(L_1, \dots, L_d)$. Then we have $r_{\mathrm{SD}}(\boldsymbol{\lambda}) = (\boldsymbol{\lambda}'\Sigma\boldsymbol{\lambda})^{1/2}$, from which it follows that

$$\mathrm{AC}_i^\varrho = \frac{\partial r_{\mathrm{SD}}}{\partial \lambda_i}(\mathbf{1}) = \frac{(\Sigma\mathbf{1})_i}{r_{\mathrm{SD}}(\mathbf{1})} = \frac{\sum_{j=1}^d \mathrm{cov}(L_i, L_j)}{r_{\mathrm{SD}}(\mathbf{1})} = \frac{\mathrm{cov}(L_i, L)}{\sqrt{\mathrm{var}(L)}}.$$

This formula is known as the *covariance principle*. If we consider more generally a risk measure of the form $\varrho(L) = E(L) + \kappa\,\mathrm{SD}(L)$ for some $\kappa > 0$, we get $r_\varrho(\boldsymbol{\lambda}) = \boldsymbol{\lambda}'E(\boldsymbol{L}) + \kappa r_{\mathrm{SD}}(\boldsymbol{\lambda})$ and hence

$$\mathrm{AC}_i^\varrho = E(L_i) + \kappa\frac{\mathrm{cov}(L_i, L)}{\sqrt{\mathrm{var}(L)}}.$$

VaR and VaR contributions. Suppose that $r_{\mathrm{VaR}}^\alpha(\boldsymbol{\lambda}) = q_\alpha(L(\boldsymbol{\lambda}))$. In this case it can be shown that, subject to technical conditions,

$$\mathrm{AC}_i^\varrho = \frac{\partial r_{\mathrm{VaR}}^\alpha}{\partial \lambda_i}(\mathbf{1}) = E(L_i \mid L = q_\alpha(L)), \quad 1 \leqslant i \leqslant d. \tag{8.62}$$

The derivation of (8.62) is more involved than that of the covariance principle, and we give a justification following Tasche (2000) under the simplifying assumption that the loss distribution of $(L_1, \dots, L_d)$ has a joint density f. In the following lemma we denote by $\phi(u, l_2, \dots, l_d) = f_{L_1 \mid L_2,\dots,L_d}(u \mid l_2, \dots, l_d)$ the conditional density of L_1.

Lemma 8.42. *Assume that $d \geqslant 2$ and that $(L_1, \dots, L_d)$ has a joint density. Then, for any vector $(\lambda_1, \dots, \lambda_d)$ of portfolio weights such that $\lambda_1 \neq 0$, we find that*

(i) *$L(\boldsymbol{\lambda})$ has density*

$$f_{L(\boldsymbol{\lambda})}(t) = |\lambda_1|^{-1}E\biggl(\phi\biggl(\lambda_1^{-1}\biggl(t - \sum_{j=2}^d \lambda_j L_j\biggr), L_2, \dots, L_d\biggr)\biggr);$$

and

(ii) *for $i = 2, \dots, d$,*

$$E(L_i \mid L(\boldsymbol{\lambda}) = t) = \frac{E(L_i\phi(\lambda_1^{-1}(t - \sum_{j=2}^d \lambda_j L_j), L_2, \dots, L_d))}{E(\phi(\lambda_1^{-1}(t - \sum_{j=2}^d \lambda_j L_j), L_2, \dots, L_d))}, \quad \text{a.s.}$$

Proof. For (i) consider the case $\lambda_1 > 0$ and observe that we can write

$$\begin{aligned}P(L(\boldsymbol{\lambda}) \leqslant t) &= E(P(L(\boldsymbol{\lambda}) \leqslant t \mid L_2, \dots, L_d)) \\ &= E\biggl(P\biggl(L_1 \leqslant \lambda_1^{-1}\biggl(t - \sum_{j=2}^{d} \lambda_j L_j\biggr) \biggm| L_2, \dots, L_d\biggr)\biggr) \\ &= E\biggl(\int_{-\infty}^{\lambda_1^{-1}(t-\sum_{j=2}^{d}\lambda_j L_j)} \phi(u, L_2, \dots, L_d)\,\mathrm{d}u\biggr).\end{aligned}$$

The assertion follows by differentiating under the expectation.

For (ii) observe that we can write

$$E(L_i \mid L(\boldsymbol{\lambda}) = t) = \lim_{\delta \to 0} \frac{\delta^{-1} E(L_i I_{\{t < L(\boldsymbol{\lambda}) \leqslant t+\delta\}})}{\delta^{-1} P(t < L(\boldsymbol{\lambda}) \leqslant t + \delta)} = \frac{(\partial/\partial t) E(L_i I_{\{L(\boldsymbol{\lambda}) \leqslant t\}})}{f_{L(\boldsymbol{\lambda})}(t)},$$

provided $f_{L(\boldsymbol{\lambda})}(t) \neq 0$. The result follows by applying a similar conditioning technique to the ones used in the proof of (i) to the numerator. □

We now explain why (8.62) follows from Lemma 8.42. Since the rv $L(\boldsymbol{\lambda})$ has a density, we have $P(L(\boldsymbol{\lambda}) \leqslant q_\alpha(L(\boldsymbol{\lambda}))) = \alpha$. Writing $k(t) = \lambda_1^{-1}(t - \sum_{j=2}^{d} \lambda_j L_j)$, we have

$$\alpha = P(L(\boldsymbol{\lambda}) \leqslant r_{\mathrm{VaR}}^{\alpha}(\boldsymbol{\lambda})) = E\biggl(\int_{-\infty}^{k(r_{\mathrm{VaR}}^{\alpha}(\boldsymbol{\lambda}))} \phi(u, L_2, \dots, L_d)\,\mathrm{d}u\biggr). \tag{8.63}$$

We take derivatives of (8.63) with respect to λ_i for $i = 2, \dots, d$ to get

$$0 = \lambda_1^{-1} E\biggl(\biggl(\frac{\partial r_{\mathrm{VaR}}^{\alpha}(\boldsymbol{\lambda})}{\partial \lambda_i} - L_i\biggr)\phi(k(r_{\mathrm{VaR}}^{\alpha}(\boldsymbol{\lambda})), L_2, \dots, L_d)\biggr).$$

Solving this expression for $\partial r_{\mathrm{VaR}}^{\alpha}(\boldsymbol{\lambda})/\partial \lambda_i$, using part (ii) of Lemma 8.42 and substituting $\boldsymbol{\lambda} = \mathbf{1}$ yields (8.62), as desired. Analogous calculations can be done for $i = 1$ and $\lambda_1 < 0$. Tasche (2000) makes the derivations mathematically rigorous by using the implicit function theorem and giving all necessary conditions.

Expected shortfall and shortfall contributions. Now consider using the risk-measure function $r_{\mathrm{ES}}^{\alpha}(\boldsymbol{\lambda}) = E(L \mid L \geqslant q_\alpha(L(\boldsymbol{\lambda})))$ corresponding to expected shortfall. It follows from Definition 2.12 that we can write

$$r_{\mathrm{ES}}^{\alpha}(\boldsymbol{\lambda}) = \frac{1}{1-\alpha} \int_{\alpha}^{1} r_{\mathrm{VaR}}^{u}(\boldsymbol{\lambda})\,\mathrm{d}u,$$

where we make use of the notation $r_{\mathrm{VaR}}^{\alpha}(\boldsymbol{\lambda}) = q_\alpha(L(\boldsymbol{\lambda}))$ as above. We apply the Euler principle by again computing the derivative with respect to λ_i. Assuming the differentiability of $r_{\mathrm{VaR}}^{u}(\boldsymbol{\lambda})$, we have, with $L = L(\mathbf{1})$,

$$\frac{\partial r_{\mathrm{ES}}^{\alpha}}{\partial \lambda_i}(\mathbf{1}) = \frac{1}{1-\alpha} \int_{\alpha}^{1} \frac{\partial r_{\mathrm{VaR}}^{u}}{\partial \lambda_i}(\mathbf{1})\,\mathrm{d}u = \frac{1}{1-\alpha} \int_{\alpha}^{1} E(L_i \mid L = q_u(L)\,\mathrm{d}u.$$

Now we assume that the density f_L of L is strictly positive so that the df of L has a differentiable inverse and we can make the change of variables $v = q_u(L) = F_L^{\leftarrow}(u)$. Since $\mathrm{d}v/\mathrm{d}u = (f_L(v))^{-1}$, we get

$$\frac{\partial r_{\mathrm{ES}}^{\alpha}}{\partial \lambda_i}(\mathbf{1}) = \frac{1}{1-\alpha}\int_{q_\alpha(L)}^{\infty} E(L_i \mid L = v) f_L(v)\,\mathrm{d}v = \frac{1}{1-\alpha}E(L_i;\, L \geqslant q_\alpha(L))).$$

Hence the Euler capital allocation takes the form

$$\mathrm{AC}_i^{\varrho} = E(L_i \mid L \geqslant \mathrm{VaR}_\alpha(L)), \quad L := L(\mathbf{1}), \tag{8.64}$$

where AC_i^{ϱ} is known as the *expected shortfall contribution* of investment possibility (or line of business) i. This is a popular allocation principle in practice, and is often considered to be preferable to the covariance principle and the principle based on VaR contributions. See Notes and Comments for literature on its use in practice in the context of credit portfolios.

Euler allocation for elliptical loss distributions. In the following corollary to Theorem 8.28 we consider the special case of an elliptical loss distribution for the vector $(L_1, \dots, L_d)$. We consider this distribution to be centred at zero so that it really represents fluctuations of the loss around its mean; centring $(L_1, \dots, L_d)$ is of course equivalent to working with the mean-adjusted version of some translation-invariant risk measure ϱ. We find that the relative amounts of capital allocated to each investment opportunity are always the same, regardless of whether we base an Euler allocation on the standard deviation, VaR or expected shortfall risk measures, or indeed any positive-homogeneous risk measure. Allocation is therefore very simple in this case: depending on our choice of risk measure we calculate the total risk capital to be allocated and then use a simple partitioning formula given in (8.65) below.

Corollary 8.43. *Assume that $r_\varrho \colon \Lambda \to \mathbb{R}$ is the risk-measure function of a positive-homogeneous and law-invariant risk measure ϱ. Let $\boldsymbol{L} \sim E_d(\mathbf{0}, \Sigma, \psi)$. Then, under an Euler allocation, the relative capital allocation is given by*

$$\frac{\mathrm{AC}_i^{\varrho}}{\mathrm{AC}_j^{\varrho}} = \frac{\sum_{k=1}^{d} \Sigma_{ik}}{\sum_{k=1}^{d} \Sigma_{jk}}, \quad 1 \leqslant i, j \leqslant d. \tag{8.65}$$

Proof. From the proof of Theorem 8.28 we deduce that, by the positive homogeneity of the risk measure, we have

$$r_\varrho(\boldsymbol{\lambda}) = \varrho(L(\boldsymbol{\lambda})) = \varrho\biggl(\sum_{i=1}^{d} \lambda_i L_i\biggr) = \sqrt{\boldsymbol{\lambda}'\Sigma\boldsymbol{\lambda}}\,\varrho(Y_1),$$

where Y_1 is the first component of a spherical random vector with characteristic generator ψ. For the Euler allocation we get

$$\mathrm{AC}_i^{\varrho} = \frac{\partial r_\varrho}{\partial \lambda_i}(\mathbf{1}) = \frac{\sum_{k=1}^{d} \Sigma_{ik}}{\sqrt{\mathbf{1}'\Sigma\mathbf{1}}}\varrho(Y_1),$$

from which the result follows. □

8.5.3 Economic Properties of the Euler Principle

In this section we show that the Euler principle has a number of good economic properties. As in the previous section we consider a positive-homogeneous risk measure ϱ and we assume that the corresponding risk-measure function r_ϱ is continuously differentiable in $\mathbb{R}^d \setminus \{0\}$ (a positive-homogeneous function is typically not differentiable in $\boldsymbol{\lambda} = 0$). By $\mathrm{AC}_i^\varrho = \partial r_\varrho(\mathbf{1})/\partial\lambda_i$ we then denote the associated risk contributions under the Euler principle.

Compatibility with a RORAC approach. We define the RORAC of the overall loss by $\mathrm{RORAC}(L) := E(-L)/\rho(L)$; the portfolio-related RORAC of investment unit i is defined as

$$\mathrm{RORAC}(L_i \mid L) := \frac{E(-L_i)}{\mathrm{AC}_i^\varrho},$$

where it is tacitly assumed that the denominator is strictly positive. The Euler principle is then compatible with a RORAC approach in the following sense: if investment opportunity i performs better than the overall portfolio L in the RORAC metric, then the RORAC of the overall portfolio is increased if one increases slightly the weight of unit i. The Euler principle therefore gives correct signals for investment decisions. In mathematical terms, RORAC compatibility means that there is some $\varepsilon > 0$ such that for all $0 < h \leqslant \varepsilon$ it holds that

$$(\mathrm{RORAC}(L_i \mid L) > \mathrm{RORAC}(L)) \Rightarrow (\mathrm{RORAC}(L+hL_i) > \mathrm{RORAC}(L)). \quad (8.66)$$

In order to establish (8.66) it suffices to show that $\mathrm{RORAC}(L_i \mid L) > \mathrm{RORAC}(L)$ implies that $(\mathrm{d}/\mathrm{d}h)\,\mathrm{RORAC}(L + hL_i)|_{h=0} > 0$. Denote by $\boldsymbol{e}_i$ the ith unit vector. Then it holds that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}h}\mathrm{RORAC}(L + hL_i)\bigg|_{h=0} &= \frac{\mathrm{d}}{\mathrm{d}h}\frac{E(-(L + hL_i))}{r_\varrho(\mathbf{1} + h\boldsymbol{e}_i)}\bigg|_{h=0} \\ &= \frac{1}{r_\varrho(\mathbf{1})^2}\bigg(E(-L_i)r_\varrho(\mathbf{1}) - E(-L)\frac{\partial r_\varrho(\mathbf{1})}{\partial\lambda_i}\bigg).\end{aligned}$$

Recall that $\varrho(L) = r_\varrho(\mathbf{1})$ and that $\mathrm{AC}_i^\varrho = \partial r_\varrho(\mathbf{1})/\partial\lambda_i$. Hence the last expression is strictly positive if $E(-L_i)/\mathrm{AC}_i^\varrho > E(-L)/\varrho(L)$, as claimed.

In fact, it can be shown that for a positive-homogeneous ϱ the Euler principle is the only capital allocation principle that satisfies the RORAC compatibility (8.66) (see Tasche (1999) for details).

Diversification benefit. Suppose that the risk measure ϱ is positive homogeneous and subadditive, as is the case for a coherent risk measure or a mean-adjusted version thereof. In that case, since $\varrho(L) \leqslant \sum_{i=1}^d \varrho(L_i)$, the overall risk capital required for the portfolio is smaller than the sum of the risk capital required for the business units on a stand-alone basis. In practice, the difference $\sum_{i=1}^d \varrho(L_i) - \varrho(L)$ is known as the *diversification benefit*. It is reasonable to require that each business unit profits from the diversification benefit in the sense that the individual risk contribution of unit i does not exceed the stand-alone capital charge $\varrho(L_i)$ (otherwise there would be an incentive for unit i to leave the firm, at least in theory). We now show that the Euler principle does indeed satisfy the inequality

$$\mathrm{AC}_i^{\varrho} \leqslant \varrho(L_i), \quad 1 \leqslant i \leqslant d. \tag{8.67}$$

The key is the following inequality: for a convex and positive-homogeneous function $f\colon \mathbb{R}^d \to \mathbb{R}$ that is continuously differentiable in $\mathbb{R}^d \setminus \{0\}$, it holds for all $\boldsymbol{\lambda}$, $\tilde{\boldsymbol{\lambda}}$ with $\boldsymbol{\lambda} \neq -\tilde{\boldsymbol{\lambda}}$ that

$$f(\boldsymbol{\lambda}) \geqslant \sum_{i=1}^{d} \lambda_i \frac{\partial f(\boldsymbol{\lambda} + \tilde{\boldsymbol{\lambda}})}{\partial \lambda_i}. \tag{8.68}$$

If we apply this inequality with $f = r_{\varrho}$ (which is convex as ϱ is positive homogeneous and subadditive), $\boldsymbol{\lambda} = \boldsymbol{e}_i$ and $\tilde{\boldsymbol{\lambda}} = \mathbf{1} - \boldsymbol{e}_i$, we get the inequality $r_{\varrho}(\boldsymbol{e}_i) \geqslant \partial r_{\varrho}(\mathbf{1})/\partial \lambda_i$ and hence (8.67).

It remains to establish the inequality (8.68). Since f is convex it holds for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\boldsymbol{x} \neq 0$, that

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \sum_{i=1}^{d} (y_i - x_i) \frac{\partial f(\boldsymbol{x})}{\partial x_i}.$$

Moreover, by Euler's rule we have $f(\boldsymbol{x}) = \sum_{i=1}^{d} x_i \partial f(\boldsymbol{x})/\partial x_i$ and hence

$$f(\boldsymbol{y}) \geqslant \sum_{i=1}^{d} y_i \frac{\partial f(\boldsymbol{x})}{\partial x_i}.$$

Substituting $\boldsymbol{y} = \boldsymbol{\lambda}$, $\boldsymbol{x} = \boldsymbol{\lambda} + \tilde{\boldsymbol{\lambda}}$ gives the result.

The work of Kalkbrener (2005) and Denault (2001) takes this analysis one step further. In these papers it is shown that under suitable technical conditions the Euler principle is the only capital allocation principle that satisfies a slight strengthening of the diversification-benefit inequality (8.67). Obviously, this gives additional support for using the Euler principle if one works in the realm of coherent risk measures. From a practical point of view, the use of expected shortfall and expected shortfall contributions might be a reasonable choice in many application areas, particularly for credit risk management and loan pricing (see Notes and Comments, where this issue is discussed further).

Notes and Comments

A broad, non-technical discussion of capital allocation and performance measurement is to be found in Matten (2000) (see also Klaassen and van Eeghen 2009). The term "Euler principle" seems to have first been used in Patrik, Bernegger and Rüegg (1999). The result (8.62) is found in Gouriéroux, Laurent and Scaillet (2000) and Tasche (2000); the former paper assumes that the losses have a joint density and the latter gives a slightly more general result as well as technical details concerning the differentiability of the VaR and ES risk measures with respect to the portfolio composition. Differentiability of the coherent premium principle of Section 2.3.5 is discussed in Fischer (2003). The derivation of allocation principles from properties of risk measures is also to be found in Goovaerts, Dhaene and Kaas (2003) and Goovaerts, van den Borre and Laeven (2005).

For the arguments concerning suitability of risk measures for performance measurement, see Tasche (1999) and Tasche (2008). An axiomatic approach to capital allocation is found in Kalkbrener (2005) and Denault (2001). For an early contribution on game theory applied to cost allocation in an insurance context, see Lemaire (1984).

Applications to credit risk are found in Kalkbrener, Lotter and Overbeck (2004) and Merino and Nyfeler (2004); these make strong arguments in favour of the use of expected shortfall contributions. On the other hand, Pfeifer (2004) contains some compelling examples to show that expected shortfall as a risk measure and expected shortfall contributions as an allocation method may have some serious deficiencies when used in non-life insurance. The existence of rare, extreme events may lead to absurd capital allocations when based on expected shortfall. The reader is therefore urged to reflect carefully before settling on a specific risk measure and allocation principle. It may also be questionable to base a "coherent" risk-sensitive capital allocation on formal criteria only; for further details on this from a non-life insurance perspective see Koryciorz (2004).

Risk-adjusted performance measures are widely used in industry in the context of capital budgeting and performance measurement. A good overview of current practice is given in Chapter 14 of Crouhy, Galai and Mark (2001) (see also Klaassen and van Eeghen 2009). An analysis of risk management and capital budgeting for financial institutions from an economic viewpoint is given in Froot and Stein (1998).

Part III

Applications

9
Market Risk

In this chapter we look at methods for measuring the market risk in portfolios of traded instruments. We emphasize the use of statistical models and techniques introduced in Part II of the book. While we draw on material from most of the foregoing chapters, essential prerequisites are Chapter 2, in which the basic risk measurement problem was introduced, and Chapter 4 on financial time series. The material is divided into three sections.

In Section 9.1 we revisit the topic of risk factors and mappings, first described in very general terms in Section 2.2. We develop the modelling framework in more detail in this chapter for the specific problem of modelling market risk in a bank's trading book, where derivative positions are common and the regulator requires risk to be measured over short time horizons such as one day or two trading weeks.

Section 9.2 is devoted to the topic of market-risk measurement. Assuming that the portfolio has been mapped to risk factors, we describe the various statistical approaches that are used in industry to estimate loss distributions and risk measures like VaR or expected shortfall. These methods include the variance–covariance (delta-normal), historical simulation and Monte Carlo methods.

The subject of backtesting the performance of such methods is treated in Section 9.3. We describe commonly used model-validation procedures based on VaR violations as well as more recent proposals for comparing methods using scoring functions based on elicitability theory.

9.1 Risk Factors and Mapping

The key idea in this section is that of a *loss operator*, which is introduced in Section 9.1.1. This is a function that relates portfolios losses to changes in the risk factors and is effectively the function that a bank must evaluate in order to determine the P&L of its trading book under scenarios for future risk-factor changes. Since the time to maturity or expiry has an impact on the value of many market instruments, we consider the issue of different timescales for risk measurement and valuation in detail. In Section 9.1.2 we show how the typically non-linear loss operator can be approximated over short time intervals by linear (delta) and quadratic (delta–gamma) functions.

The methodology is applied to a portfolio of zero-coupon bonds in Section 9.1.3, and it is shown that the linear and quadratic approximations to the loss operator have interpretations in terms of the classical bond pricing concepts of duration and convexity. Since the mapping of fixed-income portfolios is typically a high-dimensional problem, we consider factor modelling strategies for reducing the complexity of the mapping exercise in Section 9.1.4.

9.1.1 The Loss Operator

Consider a portfolio of assets subject to market risk, such as a collection of stocks and bonds or a book of derivatives. The value of the portfolio is given by the continuous-time stochastic process $(V(t))_{t\in\mathbb{R}}$, where it is assumed that $V(t)$ is *known* at time t; this means that the instruments in the portfolio can either be marked-to-market or marked to an appropriate model (see Section 2.2.2 for discussion of these concepts).

For a given time horizon Δt, such as one or ten days in a typical market-risk application, the P&L of the portfolio over the period $[t, t + \Delta t]$ is given by $V(t + \Delta t) - V(t)$. We find it convenient to consider the negative P&L $-(V(t+\Delta t)-V(t))$ and to represent the risk by the right tail of this quantity, which we refer to simply as the loss. It is assumed that the portfolio composition remains constant over this period and that there is no intermediate income or fees (the so-called clean or no-action P&L).

In transforming the problem of analysing the loss distribution to a problem in financial time-series analysis, it is convenient to measure time in units of Δt and to introduce appropriate time-series notation. In a number of places in this chapter we move from a generic continuous-time process $Y(t)$ to the time series $(Y_t)_{t\in\mathbb{Z}}$ by setting

$$Y_t := Y(\tau_t), \quad \tau_t := t(\Delta t). \tag{9.1}$$

Using this notation the loss is written as

$$L_{t+1} := -(V(\tau_{t+1}) - V(\tau_t)) = -(V_{t+1} - V_t). \tag{9.2}$$

In market-risk management we often work with valuation models (such as Black–Scholes) where calendar time is measured in years and interest rates and volatilities are quoted on an annualized basis. In this case, if we are interested in daily losses, we set $\Delta t = 1/365$ or $\Delta t \approx 1/250$; the latter convention is mainly used in markets for equity derivatives since there are approximately 250 trading days per year. The rvs V_t and V_{t+1} then represent the portfolio value on days t and $t + 1$, respectively, and L_{t+1} is the loss from day t to day $t + 1$.

As explained in Section 2.2.1 the value V_t is modelled as a function of time and a d-dimensional random vector $\boldsymbol{Z}_t = (Z_{t,1}, \dots, Z_{t,d})'$ of risk factors. This procedure is referred to as *mapping*. Using the canonical units of time for the valuation model (typically years), mapping leads to an equation of the form

$$V_t = g(\tau_t, \boldsymbol{Z}_t) \tag{9.3}$$

for some measurable function $g\colon \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}$ and some vector of appropriate risk factors $\boldsymbol{Z}_t$. The choice of the function g and risk factors $\boldsymbol{Z}_t$ reflects the structure of the portfolio and also the desired level of precision in the modelling of risk.

Note that by introducing $f(t, \boldsymbol{Z}_t) := g(\tau_t, \boldsymbol{Z}_t)$ we can get the simpler version of the mapping formula used in (2.2). However, the use of the $g(\tau_t, \boldsymbol{Z}_t)$ notation allows us more flexibility to map positions while preserving market conventions with respect to timescale; see, for example, the mapping of a zero-coupon bond portfolio in Section 9.1.3.

Recall that the risk-factor changes $(\boldsymbol{X}_t)_{t\in\mathbb{Z}}$ are given by $\boldsymbol{X}_t := \boldsymbol{Z}_t - \boldsymbol{Z}_{t-1}$. Using the mapping (9.3) the portfolio loss can be written as

$$L_{t+1} = -(g(\tau_{t+1}, \boldsymbol{Z}_t + \boldsymbol{X}_{t+1}) - g(\tau_t, \boldsymbol{Z}_t)). \tag{9.4}$$

Since $\boldsymbol{Z}_t$ is known at time t, the loss distribution at time t is determined by the distribution of the risk-factor change $\boldsymbol{X}_{t+1}$. We therefore introduce a new piece of notation in this chapter, namely the *loss operator* at time t, written $l_{[t]}\colon \mathbb{R}^d \to \mathbb{R}$, which maps risk-factor changes into losses. It is defined by

$$l_{[t]}(\boldsymbol{x}) := -(g(\tau_{t+1}, \boldsymbol{z}_t + \boldsymbol{x}) - g(\tau_t, \boldsymbol{z}_t)), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{9.5}$$

where $\boldsymbol{z}_t$ denotes the realized value of $\boldsymbol{Z}_t$ at time t; we obviously have $L_{t+1} = l_{[t]}(\boldsymbol{X}_{t+1})$ at time t. The loss operator will facilitate our discussion of statistical approaches to measuring market risk in Section 9.2.

Note that, while we use lowercase $\boldsymbol{z}_t$ in (9.5) to emphasize that the loss operator is a function of known risk-factor values at time t, we will not apply this convention strictly in later examples.

9.1.2 Delta and Delta–Gamma Approximations

If the mapping function g is differentiable and Δt is relatively small, we can approximate g with a first-order Taylor series approximation

$$g(\tau_t + \Delta t, \boldsymbol{z}_t + \boldsymbol{x}) \approx g(\tau_t, \boldsymbol{z}_t) + g_\tau(\tau_t, \boldsymbol{z}_t)\Delta t + \sum_{i=1}^{d} g_{z_i}(\tau_t, \boldsymbol{z}_t)x_i, \tag{9.6}$$

where the τ subscript denotes the partial derivative with respect to the time argument of the mapping and the z_i subscripts denote the partial derivatives with respect to the risk factors. This allows us to approximate the loss operator in (9.5) by the *linear loss operator* at time t, which is given by

$$l^{\Delta}_{[t]}(\boldsymbol{x}) := -\biggl(g_\tau(\tau_t, \boldsymbol{z}_t)\Delta t + \sum_{i=1}^{d} g_{z_i}(\tau_t, \boldsymbol{z}_t)x_i\biggr). \tag{9.7}$$

Note that, when working with a short time horizon Δt, the term $g_\tau(\tau_t, \boldsymbol{z}_t)\Delta t$ is very small and is sometimes omitted in practice.

We can also develop a second-order Taylor series, or so-called *delta–gamma*, approximation. Suppose we introduce vector notation

$$\boldsymbol{\delta}(\tau_t, \boldsymbol{z}_t) = (g_{z_1}(\tau_t, \boldsymbol{z}_t), \dots, g_{z_d}(\tau_t, \boldsymbol{z}_t))'$$

for the first-order partial derivatives of the mapping with respect to the risk factors. For the second-order partial derivatives let

$$\boldsymbol{\omega}(\tau_t, \boldsymbol{z}_t) = (g_{z_1\tau}(\tau_t, \boldsymbol{z}_t), \dots, g_{z_d\tau}(\tau_t, \boldsymbol{z}_t))'$$

denote the vector of mixed partial derivatives with respect to time and the risk factors and let $\Gamma(\tau_t, z_t)$ denote the matrix with (i, j)th element given by $g_{z_i z_j}(\tau_t, z_t)$; this matrix contains *gamma sensitivities* to individual risk factors on the diagonal and *cross gamma sensitivities* to pairs of risk factors off the diagonal. The full second-order approximation of g is

$$\begin{aligned} g(\tau_t + \Delta t, z_t + \boldsymbol{x}) \approx{}& g(\tau_t, z_t) + g_\tau(\tau_t, z_t)\Delta t + \boldsymbol{\delta}(\tau_t, z_t)'\boldsymbol{x} \\ &+ \tfrac{1}{2}(g_{\tau\tau}(\tau_t, z_t)(\Delta t)^2 + 2\boldsymbol{\omega}(\tau_t, z_t)'\boldsymbol{x}\Delta t + \boldsymbol{x}'\Gamma(\tau_t, z_t)\boldsymbol{x}). \end{aligned} \tag{9.8}$$

In practice, we would usually omit terms of order $o(\Delta t)$ (terms that tend to zero faster than Δt). In the above expression this is the term in $(\Delta t)^2$ and, if we assume that risk factors follow a standard continuous-time financial model such as Black–Scholes or many generalizations thereof, the term in $\boldsymbol{x}\Delta t$.

To understand better why the last statement is true, consider the case of the Black–Scholes model. The log stock price at time t is given by $\ln S_t = (\mu - \frac{1}{2}\sigma^2)\tau_t + \sigma W_{\tau_t}$, where μ is the drift, σ is the volatility, $\tau_t = t(\Delta t)$ as usual and W_{τ_t} denotes Brownian motion. It follows that the risk-factor change satisfies

$$X_{t+1} = \ln\left(\frac{S_{t+1}}{S_t}\right) \sim N((\mu - \tfrac{1}{2}\sigma^2)\Delta t, \sigma^2\Delta t).$$

Clearly, $X_{t+1}/(\sigma\sqrt{\Delta t})$ converges in distribution to a standard normal variable as $\Delta t \to 0$. Risk-factor changes x in this model are therefore of order $O(\sqrt{\Delta t})$, meaning they tend to zero at the same rate as $\sqrt{\Delta t}$. It follows that the term $x\Delta t$ tends to zero at the same rate as $(\Delta t)^{3/2}$ and is therefore a term of order $o(\Delta t)$.

Omitting terms of order $o(\Delta t)$ in (9.8) leaves us with the *quadratic loss operator*

$$l_{[t]}^{\Delta\Gamma}(\boldsymbol{x}) := -(g_\tau(\tau_t, z_t)\Delta t + \boldsymbol{\delta}(\tau_t, z_t)'\boldsymbol{x} + \tfrac{1}{2}\boldsymbol{x}'\Gamma(\tau_t, z_t)\boldsymbol{x}), \tag{9.9}$$

and this typically provides a more accurate approximation to (9.5) than the linear loss operator. In Example 9.1 below we give an application of the delta–gamma approximation (9.9).

Example 9.1 (European call option). The set-up and notation in this example are similar to those of Example 2.2 but we now consider a European call option that has been sold by a bank and *delta-hedged* to remove some of the risk. This means that the bank has bought a quantity of stock equivalent to the delta of the option so that the first-order sensitivity of the hedged position to stock price changes is 0. To simplify the analysis of risk factors, we assume that the interest rate r is constant.

Using the time-series notation in (9.1), the value of the hedged position at time t is

$$V_t = S_t h_t - C^{\mathrm{BS}}(\tau_t, S_t; r, \sigma_t, K, T), \tag{9.10}$$

where S_t and σ_t are the stock price and implied volatility at t, K is the strike price, T is the maturity and $h_t = C_S^{\mathrm{BS}}(\tau_t, S_t; r_t, \sigma_t, K, T)$ is the delta of the option. The time horizon of interest is one day and the natural time unit in the Black–Scholes formula is years, so $\Delta t = 1/250$ and $\tau_t = t/250$.

The valuation formula (9.10) is of the form (9.3) with risk factors $\boldsymbol{Z}_t = (\ln S_t, \sigma_t)'$. The linear loss operator (9.7) is given by

$$l^{\Delta}_{[t]}(\boldsymbol{x}) = C^{\mathrm{BS}}_{\tau}\Delta t + C^{\mathrm{BS}}_{\sigma}x_2,$$

since $g_{z_1}(\tau_t, z_t) = (h_t - C^{\mathrm{BS}}_{S})S_t = 0$.

Consider the situation where the time to expiry is $T - \tau_t = 1$, the strike price is 100 and the interest rate is $r = 0.02$. Moreover, assume that the current stock price is $S_t = 110$, so that the option is in the money, and the current implied volatility is $\sigma_t = 0.2$. The values of the Greeks in the Black–Scholes model may be calculated using well-known formulas (see Notes and Comments): they are $C^{\mathrm{BS}}_{\tau} \approx -4.83$ and $C^{\mathrm{BS}}_{\sigma} \approx 34.91$. Suppose we consider the effect of risk-factor changes $\boldsymbol{x} = (0.05, 0.02)'$ representing a stock return of (approximately) 5% and an increase in implied volatility of 2%. The stock return obviously makes no contribution to the linearized loss, which is given by

$$l^{\Delta}_{[t]}(\boldsymbol{x}) = C^{\mathrm{BS}}_{\tau} \cdot (1/250) + C^{\mathrm{BS}}_{\sigma} \cdot 0.02 \approx -0.019 + 0.698 = 0.679.$$

On the other hand, if we use full revaluation of the option at time $t + 1$, the loss would be given by $l_{[t]}(\boldsymbol{x}) \approx 0.812$. So, for risk-factor changes of this order of $\boldsymbol{x}$, there is a 16% underestimate involved in linearization.

To make a second-order approximation in this case we need to compute *gamma*, the second derivative C^{BS}_{SS} with respect to stock price, the second derivative $C^{\mathrm{BS}}_{\sigma\sigma}$ with respect to volatility, and the mixed derivative $C^{\mathrm{BS}}_{S\sigma}$ with respect to stock price and volatility. This gives the quadratic loss operator

$$l^{\Delta\Gamma}_{[t]}(\boldsymbol{x}) = C^{\mathrm{BS}}_{\tau}\Delta t + C^{\mathrm{BS}}_{\sigma}x_2 + \tfrac{1}{2}C^{\mathrm{BS}}_{SS}S_t^2x_1^2 + C^{\mathrm{BS}}_{S\sigma}S_tx_1x_2 + \tfrac{1}{2}C^{\mathrm{BS}}_{\sigma\sigma}x_2^2,$$

where we note that the S_t^2 and S_t factors enter the third and fourth terms because the risk factor is $\ln S_t$ rather than S_t. In the numerical example,

$$\begin{aligned} l^{\Delta\Gamma}_{[t]}(\boldsymbol{x}) &= l^{\Delta}_{[t]}(\boldsymbol{x}) + \tfrac{1}{2}C^{\mathrm{BS}}_{SS}S_t^2x_1^2 + C^{\mathrm{BS}}_{S\sigma}S_tx_1x_2 + \tfrac{1}{2}C^{\mathrm{BS}}_{\sigma\sigma}x_2^2 \\ &\approx 0.679 + 0.218 - 0.083 + 0.011 = 0.825. \end{aligned}$$

This is less than a 2% overestimate of the true loss, which is a substantially more accurate assessment of the impact of $\boldsymbol{x}$. The inclusion of the gamma of the option C^{BS}_{SS} is particularly important.

This example shows that the additional complexity of second-order approximations may often be warranted. Note, however, that delta–gamma approximations can give very poor results when applied to longer time horizons with large risk-factor changes.

9.1.3 Mapping Bond Portfolios

In this section we apply the ideas of Section 9.1.2 to the mapping of a portfolio of bonds and relate this to the classical concepts of duration and convexity in the risk management of bond portfolios.

Basic definitions for bond pricing. In standard bond pricing notation, $p(t, T)$ denotes the price at time t of a default-free zero-coupon bond with maturity T. While zero-coupon bonds of long maturities are relatively rare in practice, many other fixed-income instruments such as coupon bonds or standard swaps can be viewed as portfolios of zero-coupon bonds, and zero-coupon bonds are therefore fundamental building blocks for studying interest-rate risk. We follow a standard convention in modern interest-rate theory and normalize the face value $p(T, T)$ of the bond to 1, and we measure time in years.

The mapping $T \to p(t, T)$ for different maturities is one way of describing the so-called *term structure* of interest rates at time t. An alternative description is based on yields. The *continuously compounded yield* of a zero-coupon bond is defined to be $y(t, T) = -(1/(T - t)) \ln p(t, T)$, so that we have the relationship

$$p(t, T) = \exp(-(T - t)y(t, T)).$$

The mapping $T \mapsto y(t, T)$ is referred to as the continuously compounded *yield curve* at time t. Yields are a popular way of describing the term structure because they are comparable across different times to maturity due to the rescaling by $(T - t)$; they are generally expressed on an annualized basis.

We now consider the mapping of a portfolio of zero-coupon bonds. Note that the same mapping structure would be obtained for a single coupon bond, a portfolio of coupon bonds or any portfolio of promised cash flows at fixed future times.

Detailed mapping of a bond portfolio. We consider a portfolio of d default-free zero-coupon bonds with maturities T_i and prices $p(t, T_i)$, $1 \leqslant i \leqslant d$. By λ_i we denote the number of bonds with maturity T_i in the portfolio.

In a detailed analysis of the change in value of the bond portfolio, one takes all yields $y(t, T_i)$, $1 \leqslant i \leqslant d$, as risk factors. The value of the portfolio at time t is given by

$$V(t) = \sum_{i=1}^{d} \lambda_i p(t, T_i) = \sum_{i=1}^{d} \lambda_i \exp(-(T_i - t)y(t, T_i)). \tag{9.11}$$

Switching to a discrete-time set-up using (9.1), the mapping (9.3) of the bond portfolio can be written as

$$V_t = g(\tau_t, \mathbf{Z}_t) = \sum_{i=1}^{d} \lambda_i \exp(-(T_i - \tau_t)Z_{t,i}), \tag{9.12}$$

where $\tau_t = t(\Delta t)$, Δt is the time horizon expressed in years, and the risk factors are the yields $Z_{t,i} = y(\tau_t, T_i)$, $1 \leqslant i \leqslant d$. The risk-factor changes are the changes in yields $X_{t+1,i} = y(\tau_{t+1}, T_i) - y(\tau_t, T_i)$, $1 \leqslant i \leqslant d$.

From (9.12) the loss operator $l_{[t]}$ and its linear and quadratic approximations can easily be computed. The first derivatives of the mapping function are

$$g_\tau(\tau_t, z_t) = \sum_{i=1}^{d} \lambda_i p(\tau_t, T_i) z_{t,i},$$

$$g_{z_i}(\tau_t, z_t) = -\lambda_i (T_i - \tau_t) \exp(-(T_i - \tau_t) z_{t,i}).$$

Inserting these into (9.7) and reverting to standard bond pricing notation we obtain

$$l^{\Delta}_{[t]}(\boldsymbol{x}) = -\sum_{i=1}^{d} \lambda_i p(\tau_t, T_i)(y(\tau_t, T_i)\Delta t - (T_i - \tau_t)x_i), \tag{9.13}$$

where x_i represents the change in yield of the ith bond.

For the second-order approximation we need the second derivatives with respect to yields, which are

$$g_{z_i z_i}(\tau_t, z_t) = \lambda_i (T_i - \tau_t)^2 \exp(-(T_i - \tau_t)z_{t,i})$$

and $g_{z_i z_j}(\tau_t, z_t) = 0$ for $i \neq j$. Using standard bond pricing notation, the quadratic loss operator in (9.9) is

$$l^{\Delta\Gamma}_{[t]}(\boldsymbol{x}) = -\sum_{i=1}^{d} \lambda_i p(\tau_t, T_i)(y(\tau_t, T_i)\Delta t - (T_i - \tau_t)x_i + \tfrac{1}{2}(T_i - \tau_t)^2 x_i^2). \tag{9.14}$$

Relationship to duration and convexity. The approximations (9.13) and (9.14) can be interpreted in terms of the classical notions of the duration and convexity of bond portfolios. To make this connection consider a very simple model for the yield curve at time t in which

$$y(\tau_{t+1}, T_i) = y(\tau_t, T_i) + x \tag{9.15}$$

for all maturities T_i. In this model we assume that a *parallel shift in level* takes place along the entire yield curve, an assumption that is unrealistic but that is frequently made in practice.

Obviously, when (9.15) holds, the loss operators in (9.13) and (9.14) are functions of a scalar variable x (the size of the shift). We can express (9.13) in terms of the classical concept of the *duration* of a bond portfolio by writing

$$l^{\Delta}_{[t]}(x) = -V_t(A_t \Delta t - D_t x), \tag{9.16}$$

where

$$D_t := \sum_{i=1}^{d} \frac{\lambda_i p(\tau_t, T_i)}{V_t}(T_i - \tau_t), \qquad A_t := \sum_{i=1}^{d} \frac{\lambda_i p(\tau_t, T_i)}{V_t} y(\tau_t, T_i).$$

The term that interests us here is D_t, which is usually called the (Macaulay) *duration* of the bond portfolio. It is a weighted sum of the times to maturity of the different cash flows in the portfolio, the weights being proportional to the discounted values of the cash flows.

Over short time intervals the Δt term in (9.16) will be negligible and losses of value in the bond portfolio will be determined by $l_{[t]}(x) \approx v_t D_t x$, so that increases in the level of the yield curve lead to losses and decreases lead to gains (assuming all positions are long so that $\lambda_i > 0$ for all i). The duration D_t can be thought of as the bond pricing analogue of the delta of an option; to a first-order approximation, losses will be governed by D_t. Any two bond portfolios with equal value and duration will be subject to similar losses when there is a small parallel shift of the yield curve, regardless of differences in the exact composition of the portfolios.

Duration is an important tool in traditional bond-portfolio or asset-liability management. The standard duration-based strategy to manage the interest-rate risk of a bond portfolio is called *immunization*. Under this strategy an asset manager, who has a certain amount of funds to invest in various bonds and who needs to make certain known payments in the future, allocates these funds to various bonds in such a way that the duration of the overall portfolio consisting of bond investments and liabilities is equal to zero. As we have just seen, duration measures the sensitivity of the portfolio value with respect to shifts in the level of the yield curve. A zero duration therefore means that the position has been immunized against changes in level. However, the portfolio is still exposed to other types of yield-curve changes, such as changes in slope and curvature.

It is possible to get more accurate approximations for the loss in a bond portfolio by considering second-order effects. The analogue of the gamma of an option is the concept of *convexity*. Under our model (9.15) for changes in the level of yields, the expression for the quadratic loss operator in (9.14) becomes

$$l_{[t]}^{\Delta\Gamma}(x) = -V_t(A_t\Delta t - D_t x + \tfrac{1}{2}C_t x^2), \tag{9.17}$$

where

$$C_t := \sum_{i=1}^{d} \frac{\lambda_i p(\tau_t, T_i)}{V_t}(T_i - \tau_t)^2$$

is the convexity of the bond portfolio. The convexity is a weighted average of the squared times to maturity and is the negative of the derivative of the duration with respect to yield. Consider two portfolios (1) and (2) with identical values V_t and durations D_t. Assume that the convexity of portfolio (1) is greater than that of portfolio (2), so that $C_t^{(1)} > C_t^{(2)}$. Ignoring terms in Δt, the difference in loss operators satisfies

$$l_{[t]}^{\Delta\Gamma(1)}(x) - l_{[t]}^{\Delta\Gamma(2)}(x) \approx -\tfrac{1}{2}V_t(C_t^{(1)} - C_t^{(2)})x^2 < 0.$$

In other words, an increase in the level of yields will lead to smaller losses for portfolio (1), and a decrease in the level of yields will lead to larger gains (since $-l_{[t]}^{\Delta\Gamma(1)}(x) > -l_{[t]}^{\Delta\Gamma(2)}(x)$). For this reason portfolio managers often take steps to construct portfolios with relatively high convexity. Roughly speaking, this is done by spreading out the cash flows as much as possible (see Notes and Comments).

9.1.4 Factor Models for Bond Portfolios

For large portfolios of fixed-income instruments, such as the overall fixed-income position of a major bank, modelling changes in the yield for every cash flow maturity date becomes impractical. Moreover, the statistical task of estimating a distribution for $\boldsymbol{X}_{t+1}$ is difficult because the yields are highly dependent for different times to maturity. A pragmatic approach is therefore to build a factor model for yields that captures the main features of the evolution of the yield curve. Three-factor models of the yield curve in which the factors typically represent *level*, *slope* and *curvature* are often used in practice.