UTX

Jürgen Franke
Wolfgang Karl Härdle
Christian Matthias Hafner

Statistics of Financial Markets

An Introduction

Fourth Edition

QUANTLETS

Springer

Universitext

More information about this series at
http://www.springer.com/series/223

Jürgen Franke • Wolfgang Karl Härdle •
Christian Matthias Hafner

Statistics of Financial Markets

An Introduction

Fourth edition

Jürgen Franke
TU Kaiserslautern
FB Mathematik
Kaiserslautern
Germany

Christian Matthias Hafner
Université Catholique de Louvain
Inst. Statistique
Leuven-la-Neuve
Belgium

Wolfgang Karl Härdle
Ladislaus von Bortkiewicz Chair of
    Statistics
C.A.S.E. Centre for Applied Statistics and
    Economics
School of Business and Economics
Humboldt-Universität zu Berlin
    Unter den Linden 6
Berlin, Germany
and
Sim Kee Boon Institute for Financial
    Economics
Singapore Management University
Singapore

Printed on acid-free paper

Fig. 1 Notes of a student for the exam of a course based on this book

Fig. 2 Notes of a student for the exam of a course based on this book

Fig. 3 Notes of a student for the exam of a course based on this book

Fig. 4 Notes of a student for the exam of a course based on this book

Fig. 5 Notes of a student for the exam of a course based on this book

Preface to the Fourth Edition

The Global Financial Crisis (2007–2009) and the following European Sovereign Debt Crisis (2009–) have caused tremendous changes in the business, economy and governance. The spillover effect and globalization of the crisis affected millions and stated a genuine challenge for modern societies. Thus it was a task of extreme importance for us to revise our book and present actualized research in financial statistics and econometrics. In particular, we felt obliged to discuss a collateralized debt obligation (CDO), studied in Chap. 22. This financial product became extremely popular at the brink of the Global Financial Crisis and therefore captured the mainstream media attention as one of the roots of the crisis. Observing recent rapid changes in the interest rate market, we restructured and updated Chap. 10. In Chaps. 18 and 22 we updated the data analysis with more recent data. Moreover, exercise sections in all chapters are actualized and correspond to the new edition of the exercises and solutions book by S. Borak, W.K. Härdle and B. Lopez-Cabrera (Springer Verlag, Heidelberg, ISBN: 978-3-642-33929-5). In addition to these changes, we have eliminated a small number of errors in the third edition and added notation and definitions section. Finally, we would like to thank Piotr Majer for the editing work of this edition.

All Quantlets downloadable from the Springer.com page.

Kaiserslautern, Germany	Jürgen Franke
Berlin, Germany	Wolfgang K. Härdle
Louvain-la-Neuve, Belgium	Christian M. Hafner
January 2015	

Contents

Part III   Selected Financial Applications

Part I
Option Pricing

Chapter 1
Derivatives

Classic financial mathematics deals first and foremost with basic financial instruments such as stocks, foreign currencies and bonds. A *derivative* (*derivative security* or *contingent claim*) is a financial instrument whose value depends on the value of other, more basic *underlying* variables. In this chapter we consider forward contracts, futures contracts and options as well as some combinations.

Simple derivatives have been known on European stock exchanges since the turn of the nineteenth century. While they lost popularity between World War I and II, they revived in the 1970s with the help of work by Black, Scholes and Merton, who developed a theoretical foundation to price such instruments. Their entrepreneurial approach—which is not only applied to price derivatives but everywhere in finance where the risk of complex financial instruments is measured and controlled received a Nobel prize for economics in 1997. At the same time, it triggered the development of modern financial mathematics the basics of which is described in Chap. 6 of this book. Since we have only concentrated on the mathematical modelling ideas, relevant financial terminology is only introduced when required. Numerous details which are of practical importance but which are of no interest for mathematical modelling have been left out; we refer to, for example, Hull (2006), Welcker, Kloy and Schindler (1992).

Particularly simple derivative securities are *forward* and *future contracts*. Both contracts are agreements involving two parties and call for future delivery of an asset at an agreed price. Stocks, currencies and bonds, as well as agricultural products (grain, meat) and raw materials (oil, copper, electric energy) are underlying in the contract.

Definition 1.1 (Forward Contract) A *forward contract* is an agreement between two parties in which one of the parties assumes a long position (the other party assumes a short position) and obliges to purchase (sell) the underlying asset at a specified future date $T > t$, (*expiration date* or *maturity*) for a specified price K (*delivery strike price*).

At time t, the value $V_{K,T}(S_t, \tau)$ of such a contract depends on the current value of the underlying S_t, the time to maturity $\tau = T - t$ and of the parameters K, T specified in the contract.

Futures contracts closely resemble forward contracts. While the latter do not entail any further payments until maturity, once the agreement is signed, futures contracts are traded on an exchange and marked to the market on a daily basis. Under certain circumstances forward and futures prices are identical.

Example 1.1 An investor enters into a long-term forward contract on September 1, 2003, which obliges him to buy 1,000,000 EUR at a specified exchange rate of 1.2 USD/EUR in 90 days. The investor gains if the exchange rate is up to 1.3 USD/EUR on November 30 2003, since he can sell the 1,000,000 EUR for USD 1,300,000.

In this case t = September 1, 2003 τ = 90 days, T = November 30, and K = USD 1,200,000.

Definition 1.2 (Spot Price, Forward Price, Future Price) The current price of the underlying (stock, currency, raw material) S_t is often referred to as the *spot price*. The delivery price giving a forward contract a value of zero is called the *forward price* and denoted F_t. That is, F_t solves $V_{F_t,T}(S_t, \tau) = 0$. The *future price* is defined accordingly.

Later we will compute the value of a forward contract, which determines the forward price. Since under certain circumstances forward and future contracts have the same value, their prices are equal. When such a contract is initiated in time $t = 0$, often the delivery price is set to $K = F_0$. The contract has a value of zero for both the seller and the buyer, i.e. no payments occur. Over time, as additional transactions take place on the exchange, the delivery price K and the forward price F_t can be different.

Contrary to forward and futures contracts where both parties are obliged to carry out the transaction, an option gives one party the right to buy or sell the security. Obviously, it's important to distinguish whether the buyer or seller of the option has the right to transact. There are two types of options: call options and put options. Furthermore, European options are delimited from American options. While European options are like forward contracts, American options can be exercised at any date before maturity. These terms are derived from historical, not geographical roots.

Definition 1.3 (Call Option, Put Option) A *European call option* is an agreement which gives the holder the right to buy the underlying asset at a specified date $T > t$, (*expiry date* or *maturity*), for a specified price K, (*strike price* or *exercise price*). If the holder does not exercise, the option expires as worthless.

A *European put option* is an agreement which gives the holder the right to sell the underlying asset at a specified date T for a specified price K.

The holder of an *American call* or *put option* has the right to exercise the option at any time between t and T.

The option types defined above are also called *plain vanilla options*. In practice, many more complex derivatives exist and numerous new financial instruments are still emerging. *Over-the-counter (OTC) derivatives* are tailor made instruments designed by banking institutions to satisfy a particular consumer need. A compound option, for example, is such an OTC-derivative. It gives the holder the right to buy or sell at time T an underlying option which matures in $T' > T$. The mathematical treatment of these *exotic options* is particularly difficult, since the current value of this instrument does not only depend on the value of the underlying S_t but also on the entire path of the underlying, $S_{t'}, 0 \leq t' \leq t$.

Asian, lookback and knock-out options are path-dependent derivatives. While the delivery price K of an Asian option depends on the average value of the security of a certain period of time, it depends, in the case of a lookback option, on the minimum or maximum value of the security for a certain period of time. Knock-out options expire as worthless if the price level ever reaches a specified level.

To get used to forward and futures contracts, plain vanilla options and simple combinations of them, it is convenient to have a look at the *payoff* of an instrument, i.e. the value of the derivative at maturity T. The payoff of a long position in a forward contract is just $S_T - K$, with S_T the security's spot price at expiration date T. The holder of the contract pays K for the security and can sell it for S_T. Thus, he makes a profit if the value of the security S_T at expiration is greater than the delivery price K. Being short in a forward contract implies a payoff $K - S_T$. Both payoff functions are depicted in Fig. 1.1.

The call option payoff function is denoted:

$$\max\{S_T - K, 0\} = (S_T - K)^+.$$

Thus, the option holder only exercises if the delivery price K is less than the value of the security S_T at the expiry date T. In this case, he would receive the same cash amount as in the case of a forward or future contract. If $K > S_T$, he will clearly
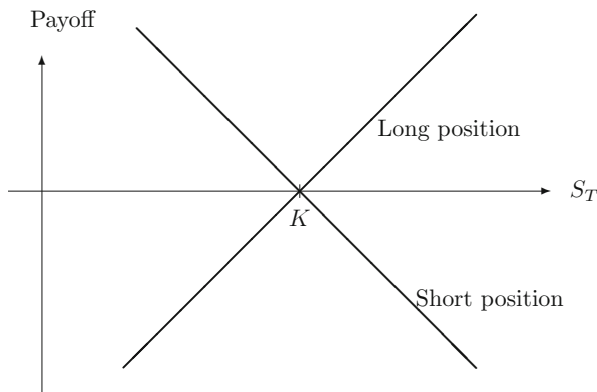
Fig. 1.1 Value of forward contract at maturity

choose not to exercise and the option expires as worthless. The put option payoff function is:

$$\max\{K - S_T, 0\} = (K - S_T)^+.$$

In contrast to forward and future contracts, options need to be bought for a positive amount $C(S_0, T)$, called the *option price* or *option prime*. Often, the options profit function is defined as $(S_T - K)^+ - C(S_0, T)$. However, this definition adds cash flows of different points in time. The correct profit is obtained by compounding the cash outflow in time $t = 0$ up to time $t = T$, since the investor could have invested the option at the risk-free interest rate r. Assuming continuous compounding at a constant interest rate r, the profit function of a call option is denoted: $(S_T - K)^+ - C(S_0, T)e^{rT}$.

Example 1.2 Consider a short call option with delivery price K and option price C_0 in time $t = 0$. The payoff and profit function are given in Figs. 1.2 and 1.3, respectively.
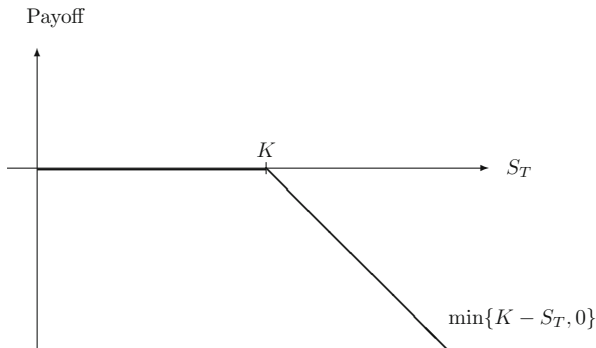
Fig. 1.2 Payoff of a short position in a call option
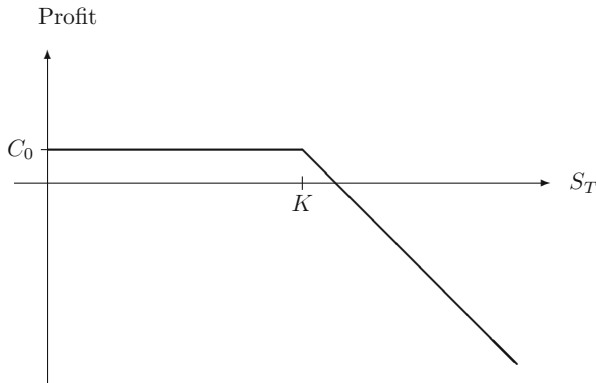
Fig. 1.3 Profit of a short position in a call option
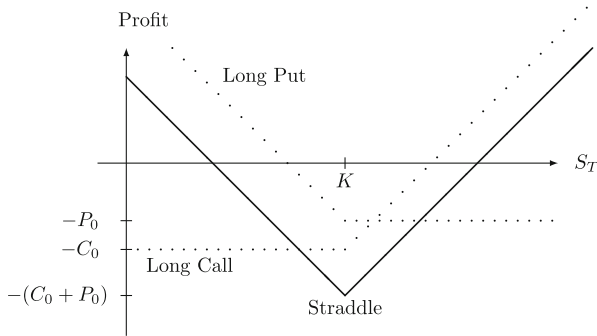
Fig. 1.4 Profit of a straddle

Example 1.3 Combining a long call and a long put with the same delivery price, K is called a *straddle*. Figure 1.4 shows the straddle profit function. C_0 and P_0 denote the call and put option price respectively.

Another fundamental financial instrument which is used in option pricing is a *bond*. Apart from interest yields, the bond holder could receive coupon payments at fixed points in time. In particular, we will consider zero-coupon bonds, i.e. bonds which promise a single payment at a fixed future date.

Definition 1.4 (Zero Coupon Bond, Discount Bond) A *zero coupon bond* or *discount bond* is a bond without coupon payments which pays interest r. The investor pays in time 0 an amount B_0 and receives at maturity T the amount B_T which is the sum of B_0 and the interest earned on B_0. The bonds' value at maturity is termed *face (nominal) value*.

Buying a zero-coupon bond corresponds to lending money at a fixed interest rate for a fixed period of time. Conversely, selling a zero-coupon bond is equivalent to borrowing money at rate r. Since bonds are traded on an exchange, they can be sold prior to maturity at price B_t, i.e. B_0 plus accrued interest up to time t.

In practice, interest rates are compounded at discrete points in time, for example annually, semiannually or monthly. If the interest rate r is compounded annually, the initial investment B_0 has (n years later) a value of $B_n^{(1)} = B_0(1 + r)^n$. If it is compounded k times per annum (p.a.), the investment pays an interest rate of $\frac{r}{k}$ each $\frac{1}{k}$ years and has a terminal value of $B_n^{(k)} = B_0(1 + \frac{r}{k})^{nk}$ after n years. However, when options and other complex derivatives are priced, continuous compounding is used, which denoted for $k \to \infty$. In this case, the initial investment B_0 grows in n years to $B_n = B_0 \cdot e^{nr}$, and r is called *short (spot) rate*. The difference between discrete and continuous compounding is small when k is large. While an investment of $B_0 = 1{,}000$ EUR at a yearly rate $r = 10\%$ grows to 1,100 EUR within a year when annually compounded, it grows to 1,105.17 EUR when continuously compounded.

In light of this, the continuous compounded rate r can be modified to account for these deviations. Assuming annual compounding at rate r_1, for both continuous and annual compounding, a continuous compounded rate $r = \log(1 + r_1)$ has to be applied, in order to obtain the same terminal value $B_n = B_n^{(1)}$.

If not stated otherwise, continuous compounding will be assumed in this chapter from here on. For comparing cash flows occurring at different points in time, they need to be compounded or discounted to the same point in time. That is, interest payments are added or subtracted. With continuous compounding, an investment of B in time t in $\Delta t > 0$ is

compounded   to time $t + \Delta t$: $B\,e^{r\Delta t}$
discounted   to time $t - \Delta t$: $B\,e^{-r\Delta t}$.

Some more financial terms will be introduced before the end of the chapter. A *portfolio* is a combination of one or more financial instruments—its value is considered as an individual financial instrument. One element of a portfolio is also called a *position*. An investor assumes a *long position* when he buys an instrument, and a *short position* when he sells it. A *long call* results from buying a call option, a *long put* from buying a put option and *a short forward* from selling a forward contract.

An investor closes out a position of his portfolio by making the future portfolio performance independent of the instrument. If the latter is traded on an exchange, he can sell (e.g. a stock or a bond) or buy (e.g. borrowed money) it. However, if the instrument is not traded, the investor can close out the position by adding to the portfolio the inverse instrument. Thus, both add up to zero and do not influence the portfolio performance any further.

Example 1.4 Consider an investor who on 1 February bought a 1,000,000 USD forward contract with a delivery price of 1,200,000 EUR with maturity within 1 year. On 1 June, he wishes to close out the position. He can *sell* another forward contract of the same size with the same delivery price and the maturity date, namely 31 January. The long and the short positions add up to zero at any point in time.

Short selling is a trading strategy that involves selling financial instruments, for example stocks, which an investor does not own. At a later stage, the investor buys back these objects. In practice, this requires the intervention of a broker who mediates with another client who owns the objects and is willing to lend them to the investor. The short selling investor commits to pay the client any foregone income, (dividends for example), that is earned in the interim.

Example 1.5 An investor selling short 1,000 stocks, borrows them from the owner and sells them immediately for $1{,}000\,S_0$ in the market (S_t denotes the stock price at time t). Later, at time $t > 0$, he closes out the position, by buying back the stocks for $1{,}000\,S_t$ and returns them to the owner. The strategy is profitable if S_t is clearly below S_0. If in time t_0, $0 < t_0 < t$, a dividend D per share is paid, the investor pays $1{,}000\,D$ to the owner. Short selling is in practice subject to numerous restrictions. In this example, it is only the possibility of short selling that will be of any interest.

1.1 Recommended Literature

Basic textbooks on derivatives are, among others, Hull (2006), Jarrow (1992) and Cox and Rubinstein (1985). Neftci (1996) and Duffie (1996) are more advanced on a mathematical level. A rather practical but still theoretically well-founded introduction is provided by Briys, Bellalah, Mai and de Varenne (1998).

1.2 Exercises

Exercise 1.1 *A Bull Call Price Spread consists of the purchase of one expensive call with a low exercise price and the sale of one cheap call with a high exercise price. Both call options have the same time to expiration. Consider a European call with an exercise price, K1 = 100 USD and another European call with an exercise price, K2 = 120 USD. Draw the payoff graph of this strategy.*

Exercise 1.2 *Think of a strategy:*

– *The purchase of one stock*
– *The purchase of one European call with exercise price K*
– *The sale of one European call with exercise price K*

Calculate the payoff and explain the risk of this strategy.

Exercise 1.3 *Consider the option spread strategy known as the* butterfly spread. *The aim of this strategy is to profit from the speculation that the price of the asset underlying the options will remain very close to strike price K_0 upon the expiration of the option contracts. The butterfly spread can be implemented using either put or call contracts. If put contracts are used, the strategy would necessitate two long put contracts, one P_1 with strike price $K_0 - (K_0/2)$ and a second, P_2 with strike price $K_0 + (K_0 * 2)$. In addition, two identical short put contracts P_3 would be required with an identical strike price K_0. If the cost of $P_1 = $ USD 5, $P_2 = $ USD 11, $P_3 = $ USD 7 and $K_0 = $ USD 17, draw the payoff diagram for this option transaction.*

Exercise 1.4 *Consider the option combination strategy known as the strangle. The strangle consists of a put and a call, each with a different strike price. The put strike price K_1 must be higher than the call strike price K_2. The aim of the strangle strategy is to profit from an anticipated upward or downward movement in the price of the underlying asset while minimizing the risk that the asset's price will remain at a level between K_2 and K_1. If the price P_1 of the long put is 4 USD and its strike price K_1 is USD, and the price C_1 of the long call is 2 USD with a strike price K_2 is 35 USD, plot the resulting payoff diagram.*

Exercise 1.5 *Consider the option combination strategy known as a strip. A strip consists of both long calls and puts with strike price K_0, and with the number of puts being greater than the number of calls. The aim of the strip is to profit from a*

large anticipated decline in the price of the underlying asset below the strike price K_0. Consider a strip strategy in which two put contracts with price $P = 3$ USD and a single call contract with price $C = 4$ USD are purchased simultaneously with strike price $K_0 = 25$ USD. Draw the payoff diagram for this option strategy.

Exercise 1.6 *Consider the option strategy known as a strap. A strap could be intuitively interpreted as the reverse of a strip. A strap consists of both long calls and puts with strike prices K_0, with the number of calls being greater than the number of puts. The aim of the strap is to profit from a large anticipated rise in the price of the underlying asset above strike price K_0. Consider a strip strategy with two long call contracts with price $C = 6$ USD purchased simultaneously along with one long put contract with price $P = 8$ USD. The strike price for both contracts $K_0 = 25$ USD. Draw the payoff diagram for this option strategy.*

Chapter 2
Introduction to Option Management

2.1 Arbitrage Relations

In this section we consider the fundamental notion of no-arbitrage. An *arbitrage opportunity* arises if it is possible to make a riskless profit. In an ideal financial market, in which all investors dispose of the same pieces of information and in which all investors can react instantaneously, there should not be any arbitrage opportunity. Since otherwise each investor would try to realize the riskless profit instantaneously. The resulting transactions would change the prices of the involved financial instruments such that the arbitrage opportunity disappears.

Additionally to no-arbitrage we presume in the remaining chapter that the financial market fulfills further simplifying assumptions which are in this context of minor importance and solely serve to ease the argumentation. If these assumptions hold we speak of a perfect financial market.

Assumption (Perfect Financial Market) *There are no arbitrage opportunities, no transaction costs, no taxes, and no restrictions on short selling. Lending rates equal borrowing rates and all securities are perfectly divisible.*

The assumption of a perfect financial market is sufficient to determine the value of future and forward contracts as well as some important relations between the prices of some types of options. Above all no mathematical model for the price of the financial instrument is needed. However, in order to determine the value of options more than only economic assumptions are necessary. A detailed mathematical modelling becomes inevitable. Each mathematical approach though has to be in line with certain fundamental arbitrage relations being developed in this chapter. If the model implies values of future and forward contracts or option prices which do not fulfil these relations the model's assumptions must be wrong.

An important conclusion drawn from the assumption of a perfect financial market and thus from no-arbitrage will be used frequently in the proofs to come. It is the fact that two portfolios, which have at a certain time T the same value, must

have the same value at a prior time $t < T$ as well. Due to its importance we will further illustrate this reasoning. We proceed from two portfolios A and B consisting of arbitrary financial instruments. Their value in time t will be denoted by $W_A(t)$ and $W_B(t)$, respectively. For any fixed point of time T, we assume that $W_A(T) = W_B(T)$ independently of the prior time T values of each financial instrument contained in A and B. For any prior point of time $t < T$ we assume without loss of generality that $W_A(t) \leq W_B(t)$. In time t an investor can construct without own financial resources a portfolio which is a combination of A and B by buying one unit of every instrument of A, selling one unit of every instrument of B (short selling) and by investing the difference $\Delta(t) = W_B(t) - W_A(t) \geq 0$ at a fixed rate r. The combined portfolio has at time t a value of

$$W_A(t) - W_B(t) + \Delta(t) = 0,$$

i.e. the investor has no initial costs. At time T the part of the combined portfolio which is invested at rate r has the compounded value $\Delta(T) = \Delta(t)e^{r(T-t)}$, and hence the combined portfolio has a value of

$$W_A(T) - W_B(T) + \Delta(T) = \Delta(t)e^{r(T-t)} > 0,$$

if $\Delta(t) > 0$. The investor made a riskless gain by investing in the combined portfolio which contradicts the no-arbitrage assumption. Therefore, it must hold $\Delta(t) = 0$, i.e. $W_A(t) = W_B(t)$.

The previous reasoning can be used to determine the unknown value of a financial derivative. For this, a portfolio A is constructed which contains instruments with known price along with one unit of the derivative under investigation. Portfolio A will be compared to another portfolio B, called the *duplicating portfolio*, which contains exclusively instruments with known prices. Since the duplicating portfolio B is constructed such that for certain it has the same value at a fixed point of time T as portfolio A the no-arbitrage assumption implies that both portfolios must have the same value at any prior point of time. The value of the financial derivative can thus be computed at any time $t \leq T$. We illustrate this procedure in the following example of a forward contract.

Theorem 2.1 *We consider a long forward contract to buy an object which has a price of S_t at time t. Let K be the delivery price, and let T be the maturity date. $V(s, \tau)$ denotes the value of the long forward contract at time t as a function of the current price $S_t = s$ and the time to maturity $\tau = T - t$. We assume constant interest rates r during the time to maturity.*

1. *If the underlying object does not pay any dividends and does not involve any costs during the time to maturity τ, then it holds*

$$V(S_t, \tau) = V_{K,T}(S_t, \tau) = S_t - Ke^{-r\tau} \tag{2.1}$$

The forward price is equal to $F_t = S_t e^{r\tau}$.

2. *If during the time to maturity the underlying pays at discrete time points dividends or involves any costs whose current time t discounted total value is equal to D_t, then it holds*

$$V(S_t, \tau) = V_{K,T}(S_t, \tau) = S_t - D_t - Ke^{-r\tau} \qquad (2.2)$$

The forward price is equal to $F_t = (S_t - D_t)e^{r\tau}$.
3. *If the underlying involves continuous costs at rate b, then it holds*

$$V(S_t, \tau) = V_{K,T}(S_t, \tau) = S_t e^{(b-r)\tau} - Ke^{-r\tau} \qquad (2.3)$$

The forward price is equal to $F_t = S_t e^{b\tau}$.

Proof For simplicity we assume the underlying object to be a stock paying either discrete dividend yields whose value discounted to time t is D_t or paying a continuous dividend yield at rate b. In the latter case the stock involves continuous costs equal to $b = r - d$. The investor having a long position in the stock gains dividends (as negative costs) at rate d but simultaneously loses interests at rate r since he invested his capital in the stock instead of in a bond with a fixed interest rate. In place of stocks, bonds, currencies or other simple instruments can be considered as well.

1. We consider at time t the following two portfolios A and B:

   Portfolio A: One long forward contract on a stock with delivery price K, maturing in time T.
   One long zero bond with face value K, maturing in time T.
   Portfolio B: A long position in one unit of the stock.

   At maturity T portfolio A contains a zero bond of value K. Selling this zero bond for K the obligation to buy the stock for K can be fulfilled. Following these transactions portfolio A consists as well as portfolio B of one unit of the stock. Thus both portfolios have at time T the same value and must therefore, due to the no-arbitrage assumption, have the same value at any time t prior to T :

$$V(S_t, \tau) + Ke^{-r\tau} = S_t , \qquad (2.4)$$

   since the value of the zero bond at time t is given by discounting K at rate r, $Ke^{-r\tau}$. The forward price is by definition the solution of

$$0 = V_{F_t,T}(S_t, \tau) = S_t - F_t e^{-r\tau}.$$

2. We consider at time t the two portfolios A and B as given above and add one position to portfolio B :

   Portfolio B: A long position in one unit of the stock and one short position of size D_t in a zero bond with interest rate r (lending an amount of money of D_t).

At maturity T the dividend yields of the stock in portfolio B, which compounded to time T amount to $D_t e^{r\tau}$, are used to pay back the bond. Thus, both portfolios A and B consist again of one unit of the stock, and therefore they must have the same value at any time $t < T$:

$$V(S_t, \tau) + Ke^{-r\tau} = S_t - D_t . \tag{2.5}$$

The forward price results as in Part I from the definition.

3. If the stock pays dividends continuously at a rate d, then the reasoning is similar as in Part II. Once again, we consider at time t two portfolios A and B. And again, A is left unchanged, B is now composed of the following position:

Portfolio B: A long position in $e^{-d\tau}$ stocks.

Reinvesting the dividends yields continuously in the stock portfolio B consists again of exactly one stock at time T. Heuristically, this can be illustrated as follows: In the time interval $[t, t + \delta]$ the stock pays approximately, for a small δ, a dividend of $d \cdot \delta \cdot S_t$. Thus, the current total amount of stocks in the portfolio, $e^{-d\tau} = e^{-d(T-t)}$, pays a total dividend of $d \cdot \delta \cdot S_t \cdot e^{-d(T-t)}$, which is reinvested in the stock. Assuming that the stock price does not change significantly in the interval $[t, t + \delta]$, i.e. $S_{t+\delta} \approx S_t$, portfolio B contains in time $t + \delta$

$$(1 + d \cdot \delta) \cdot e^{-d(T-t)} \approx e^{d\delta} \cdot e^{-d(T-t)} = e^{-d(T-t-\delta)}$$

stocks. The above reasoning can be done exactly by taking the limit $\delta \to 0$, and it can be shown that portfolio B contains at any time s between t and T exactly $e^{-d(T-s)}$ stocks. That is, for $s = T$ portfolio B is composed of exactly one stock. The same reasoning as in Part I leads to the conclusion that portfolio A and B must have the same value at any time t. Thus, we have

$$V(S_t, \tau) + Ke^{-r\tau} = e^{-d\tau} S_t . \tag{2.6}$$

where we have to set $b = r - d$. The forward price results as in Part I from the definition.   □

Example 2.1 We consider a long forward contract on a 5-year bond which is currently traded at a price of 900 EUR. The delivery price is 910 EUR, the time to maturity of the forward contract is 1 year. The coupon payments of the bond of 60 EUR occur after 6 and 12 months (the latter shortly before maturity of the forward contract). The continuously compounded annual interest rates for 6 and 12 months are 9 and 10 %, respectively. In this example we have at $t = 0$

$$S_t = 900 , \ K = 910 , \ r = 0.10 , \ \tau = 1 , \ D_t = 60e^{-0.09 \cdot \frac{1}{2}} + 60e^{-0.10} = 111.65 \tag{2.7}$$

Thus, the value of the forward contract is given by

$$V(S_t, \tau) = 900 - 111.65 - 910e^{-0.10} = -35.05. \tag{2.8}$$

The value of the respective short position in the forward contract is $+35.05$. The price F_t of the forward contract is equal to $F_t = (S_t - D_t)e^{r\tau} = 871.26$.

Example 2.2 Consider a long forward contract to buy \$1,000. If the investor buys the \$1,000 and invests this amount in an American bond, the American interest rate can be interpreted as a dividend yield d which is continuously paid. Let r be the home interest rate. The investment involves costs $b = r - d$, which are the difference between the American and the home interest rate. Denoting the dollar exchange rate by S_t the price of the forward contract is then given by

$$F_t = S_t e^{b\tau} = S_t e^{(r-d)\tau}. \tag{2.9}$$

While for $r > d$ a contango $S_t < F_t$ results, for $r < d$ a backwardation $S_t > F_t$ results. If $r > d$ and the delivery price is chosen to equal the current exchange rate, i.e. $K = S_t$, then the value of the forward contract is

$$V_{S_t,T}(S_t, \tau) = S_t(e^{-d\tau} - e^{-r\tau}) > 0.$$

Buying the forward contract at a price of S_t is thus more expensive than buying the dollars immediately for the same price since in the former case the investor can invest the money up to time T in a domestic bond paying an interest rate which is higher than the American interest rate.

The following result states that forward and future contracts with the same delivery price and the same time to maturity are equal, if interest rates are constant during the contract period. We will use the fact that by definition forward and future contracts do not cost anything if the delivery price is chosen to be equal to the current price of the forward contract, respectively, the price of the future contract.

Theorem 2.2 *If interest rates are constant during contract period, then forward and future prices are equal.*

Proof We proceed from the assumption that the future contract is agreed on at time 0, and that it has a time to maturity of N days. We assume that profits and losses are settled (marked to market) on a daily basis at a daily interest rate of ρ. While the forward price at the end of day 0 is denoted by F, the future price at the end of day t, $t = 0, 1, \cdots, N$ is denoted by F_t. The goal is to show that $F = F_0$. For that we construct two portfolios again:

Portfolio A: A long position in $e^{N\rho}$ forward contracts with delivery price F and maturity date N.
A long position in a zero bond with face value $Fe^{N\rho}$ maturing in N days.

Portfolio B: A long position in futures contracts with delivery price F_t and maturity date N. The contracts are bought daily such that the portfolio contains at the end of the t-th day exactly $e^{(t+1)\rho}$ future contracts ($t = 0, 1, \cdots, N$).
A long position in a zero bond with face value $F_0 e^{N\rho}$ maturing in N days.

Purchasing a forward or a future contract does not cost anything since their delivery prices are set to equal the current forward or future price. Due to the marking to market procedure the holder of portfolio B receives from day $t-1$ to day t for each future contract an amount of $F_t - F_{t-1}$ which can possibly be negative (i.e. he has to pay).

At maturity, i.e. at the end of day N, the zero bond of portfolio A is sold at the face value $Fe^{N\rho}$ to fulfil the terms of the forward contract and to buy $e^{N\rho}$ stocks at the delivery price F. Then A contains exclusively these stocks and has a value of $S_N e^{N\rho}$. Following, we show that portfolio B has the same value.

At the beginning of day t portfolio B contains $e^{t\rho}$ future contracts, and the holder receives due to the marking to market procedure the amount $(F_t - F_{t-1})e^{t\rho}$ which can possibly be negative. During the day he increases his long position in the future contracts at zero costs such that the portfolio contains $e^{(t+1)\rho}$ future contracts at the end of the day. The earnings at day t compounded to the maturity date have a value of:

$$(F_t - F_{t-1})e^{t\rho} \cdot e^{(N-t)\rho} = (F_t - F_{t-1})e^{N\rho}. \tag{2.10}$$

At maturity the terms of the future contracts are fulfilled due to the marking to market procedure. All profits and losses compounded to day N have a value of:

$$\sum_{t=1}^{N}(F_t - F_{t-1})e^{N\rho} = (F_N - F_0)e^{N\rho}. \tag{2.11}$$

Together with the zero bond portfolio B has at day N a value of

$$(F_N - F_0)e^{N\rho} + F_0 e^{N\rho} = F_N e^{N\rho} = S_N e^{N\rho},$$

since at maturity the future price F_N and the price S_N of the underlying are obviously equal.

Hence, both portfolios have at day N the same value and thus due to the no-arbitrage assumption their day 0 values must be equal as well. Since the forward contract with delivery price F has a value of 0 at day 0 due to the definition of the forward price, the value of portfolio A is equal to the value of the zero bond, i.e. F (face value $Fe^{N\rho}$ discounted to day 0). Correspondingly, the e^ρ futures contained in portfolio B have at the end of day 0 a value of 0 due to the definition of the future price. Again, the value of portfolio B reduces to the value of the zero bond. The latter has a value of F_0 (face value $F_0 e^{N\rho}$ discounted to day 0). Putting things together, we conclude that $F = F_0$.   □

Now, we want to prove some relationship between option prices using similar methods. The most elementary properties are summarized in the following remark without a proof. For that, we need the notion of the *intrinsic value* of an option.

Definition 2.1 (Intrinsic Value) The *intrinsic value* of a call option at time t is given by $\max(S_t - K, 0)$, the intrinsic value of a put option is given by $\max(K - S_t, 0)$. If the intrinsic value of an option is positive we say that the option is *in the money*. If $S_t = K$, then the option is *at the money*. If the intrinsic value is equal to 0, then the option is said to be *out of the money*.

Remark 2.1 Options satisfy the following elementary relations. $C(s, \tau) = C_{K,T}(s, \tau)$ and $P(s, \tau) = P_{K,T}(s, \tau)$ denote the time t value of a call and a put with delivery price K and maturity date T, if $\tau = T - t$ is the time to maturity and the price of the underlying is s, i.e. $S_t = s$.

1. Option prices are non-negative since an exercise only takes place if it is in the interest of the holder. An option gives the right to exercise. The holder is not obligated to do so.
2. American and European options have the same value at maturity T since in T they give the same rights to the holder. At maturity T the value of the option is equal to the intrinsic value:

$$C_{K,T}(S_T, 0) = \max(S_T - K, 0), \quad P_{K,T}(S_T, 0) = \max(K - S_T, 0).$$

3. An American option must be traded at least at its intrinsic value since otherwise a riskless profit can be realized by buying and immediately exercising the option. This relation does not hold in general for European options. The reason is that a European option can be exercised only indirectly by means of a future contract. The thereby involved discounting rate can possibly lead to the option being worth less than its intrinsic value.
4. The value of two American options which have different time to maturities, $T_1 \leq T_2$, is monotonous in time to maturity:

$$C_{K,T_1}(s, T_1 - t) \leq C_{K,T_2}(s, T_2 - t), \quad P_{K,T_1}(s, T_1 - t) \leq P_{K,T_2}(s, T_2 - t).$$

This follows, for calls, say, using 2., 3. from the inequality which holds at time $t = T_1$ with $s = S_{T_1}$

$$C_{K,T_2}(s, T_2 - T_1) \geq \text{intrinsic value} = \max(s - K, 0) = C_{K,T_1}(s, 0) \qquad (2.12)$$

Due to the no-arbitrage assumption the inequality must hold for any point in time $t \leq T_1$. For European options this result does not hold in general.
5. An American option is at least as valuable as the identically specified European option since the American option gives more rights to the holder.
6. The value of a call is a monotonously decreasing function of the delivery price since the right to buy is the more valuable the lower the agreed upon delivery price. Accordingly, the value of a put is a monotonously increasing function of the delivery price.

$$C_{K_1,T}(s, \tau) \geq C_{K_2,T}(s, \tau), \quad P_{K_1,T}(s, \tau) \leq P_{K_2,T}(s, \tau)$$

for $K_1 \leq K_2$. This holds for American as well as for European options.

The value of European call and put options on the same underlying with the same time to maturity and delivery price are closely linked to each other without using a complicated mathematical model.

Theorem 2.3 (Put–Call Parity for European Options) *For the value of a European call and put option which has the same maturity date T, the same delivery price K, the same underlying the following holds (where r denotes the continuous interest rate):*

1. *If the underlying pays a dividend yield with a time t discounted total value of D_t during the time to maturity $\tau = T - t$ then it holds*

$$C(S_t, \tau) = P(S_t, \tau) + S_t - D_t - Ke^{-r\tau} \tag{2.13}$$

<div align="right">🔍 SFEPutCall</div>

2. *If the underlying involves continuous costs of carry at rate b during the time to maturity $\tau = T - t$ then it holds*

$$C(S_t, \tau) = P(S_t, \tau) + S_t e^{(b-r)\tau} - Ke^{-r\tau} \tag{2.14}$$

Proof For simplicity, we again assume the underlying to be a stock. We consider a portfolio A consisting of one call which will be duplicated by a suitable portfolio B containing a put among others.

1. In the case of discrete dividend yields we consider at time t the following portfolio B :

   (a) Buy the put.
   (b) Sell a zero bond with face value K maturing T.
   (c) Buy one stock.
   (d) Sell a zero bond at the current price D_t.

   The stock in portfolio B pays dividends whose value discounted to time t is D_t. At time T these dividend yields are used to pay back the zero bond of position (d). Hence this position has a value of zero at time T. Table 2.1 shows the value of portfolio B at time T where we distinguished the situations where the put is exercised ($K \geq S_T$) and where it is not exercised. At time T portfolio B has thus

Table 2.1 Value of portfolio B at time T (Theorem 2.3)

Position	Value at time T	
	$K < S_T$	$K \geq S_T$
Long put	0	$K - S_T$
Short bond	$-K$	$-K$
Long stock	S_T	S_T
Short bond (D_t)	0	0
Sum	$S_T - K$	0

the same value $\max(S_T - K, 0)$ as the call. To avoid arbitrage opportunities both portfolios A and B must have the same value at any time t prior T, that is it holds

$$C(S_t, \tau) = P(S_t, \tau) - Ke^{-r\tau} + S_t - D_t \qquad (2.15)$$

2. In the case of continuous dividends at rate d and corresponding costs of carry $b = r - d$ we consider the same portfolio B as in Part I. but this time without position (d). Instead we buy $e^{-d\tau}$ stocks in position (c) whose dividends are immediately reinvested in the same stock. If d is negative, then the costs are financed by selling stocks. Thus, portfolio B contains exactly one stock at time T, and we conclude as in Part I that the value of portfolio B is at time t equal to the value of the call.   □

The proof of the put–call parity holds only for European options. For American options it may happen that the put or call is exercised prior maturity and that both portfolios are not hold until maturity.

The following result makes it possible to check whether prices of options on the same underlying are consistent. If the convexity formulated below is violated, then arbitrage opportunities arise as we will show in the example following the proof of the next theorem.

Theorem 2.4 *The price of a (American or European) Option is a convex function of the delivery price.*

Proof It suffices to consider calls since the proof is analogous for puts. The put–call parity for European options is linear in the term which depends explicitly on K. Hence, for European options it follows immediately that puts are convex in K given that calls are convex in K.

For $0 \leq \lambda \leq 1$ and $K_1 < K_0$ we define $K_\lambda \stackrel{\text{def}}{=} \lambda K_1 + (1 - \lambda)K_0$. We consider a portfolio A which at time $t < T$ consists of one call with delivery price K_λ and maturity date T. At time t we duplicate this portfolio by the following portfolio B:

1. A long position in λ calls with delivery price K_1 maturing in T.
2. A long position in $(1 - \lambda)$ calls delivery price K_0 maturing in T.

By liquidating both portfolios at an arbitrary point of time $t', t \leq t' \leq T$ we can compute the difference in the values of portfolio A and B which is given in Table 2.2.

Since $\lambda(S_{t'} - K_1) \geq 0$ and $(1 - \lambda)(K_0 - S_{t'})) \geq 0$ in the last row of Table 2.2 the difference in the values of portfolio A and B at time t' and thus for any point of time $t < t'$ is greater than or equal to zero. Hence, denoting $\tau = T - t$ it holds

$$\lambda C_{K_1,T}(S_t, \tau) + (1 - \lambda)C_{K_0,T}(S_t, \tau) - C_{K_\lambda,T}(S_t, \tau) \geq 0 \qquad (2.16)$$

□

Example 2.3 We consider three European call options on the MD*TECH A.G. having all the same time to maturity and delivery prices $K_1 = 190$, $K_\lambda = 200$,

Table 2.2 Difference in the values of portfolios B and A at time t' (Theorem 2.4)

	Value at time t'			
Position	$S_{t'} \leq K_1$	$K_1 < S_{t'} \leq K_\lambda$	$K_\lambda < S_{t'} \leq K_0$	$K_0 < S_{t'}$
B 1	0	$\lambda(S_{t'} - K_1)$	$\lambda(S_{t'} - K_1)$	$\lambda(S_{t'} - K_1)$
B 2	0	0	0	$(1 - \lambda)(S_{t'} - K_0)$
$-A$	0	0	$-(S_{t'} - K_\lambda)$	$-(S_{t'} - K_\lambda)$
Sum	0	$\lambda(S_{t'} - K_1)$	$(1 - \lambda)(K_0 - S_{t'})$	0

Table 2.3 Data of Example 2.3

Delivery price	Option price (EUR)
$K_1 = 190$	30.6
$K_\lambda = 200$	26.0
$K_0 = 220$	14.4

Table 2.4 Portfolio value at time T of Example 2.3

	Value at time T			
Position	$S_T \leq 190$	$190 < S_T \leq 200$	$200 < S_T \leq 220$	$220 < S_T$
1	0	$\frac{2}{3}(S_T - 190)$	$\frac{2}{3}(S_T - 190)$	$\frac{2}{3}(S_T - 190)$
2	0	0	0	$\frac{1}{3}(S_T - 220)$
3	0	0	$-(S_T - 200)$	$-(S_T - 200)$
Sum	0	$\frac{2}{3}(S_T - 190)$	$\frac{1}{3}(220 - S_T)$	0

$K_0 = 220$, i.e. $\lambda = \frac{2}{3}$. Table 2.3 shows the data of this example. Due to the last theorem it must hold:

$$\tfrac{2}{3}C_{K_1,T}(S_t, \tau) + \tfrac{1}{3}C_{K_0,T}(S_t, \tau) \geq C_{K_\lambda,T}(S_t, \tau) \tag{2.17}$$

Since this condition is obviously violated an arbitrage opportunity exists, and with the following portfolio a riskless gain can be realized:

1. A long position in $\lambda = \frac{2}{3}$ calls with delivery price K_1.
2. A long position in $1 - \lambda = \frac{1}{3}$ calls with delivery price K_0.
3. A short position in 1 call with delivery price $K_\lambda \stackrel{\text{def}}{=} \frac{2}{3}K_1 + \frac{1}{3}K_0$.

By setting up this portfolio at the current time t we realize an immediate profit of $+0.80$ EUR. The portfolio value at options' maturity T is given by Table 2.4 from which we can extract that we realize further profits for stock prices S_T between 190 and 220 of at most $\frac{20}{3}$ EUR.

We already said that option prices are monotonous functions of the delivery price. The following theorem for European options is more precise on this subject.

Table 2.5 Difference in the values of portfolios B and A at time T (Theorem 2.5)

Position	Value at time T		
	$S_T \leq K_1$	$K_1 < S_T < K_2$	$K_2 \leq S_T$
B 1	0	0	$S_T - K_2$
B 2	$K_2 - K_1$	$K_2 - K_1$	$K_2 - K_1$
$-A$	0	$-(S_T - K_1)$	$-(S_T - K_1)$
Sum	$K_2 - K_1$	$K_2 - S_T$	0

Theorem 2.5 *For two European calls (puts) with the same maturity date T and delivery prices $K_1 \leq K_2$ it holds at time $t \leq T$:*

$$0 \leq C_{K_1,T}(S_t, \tau) - C_{K_2,T}(S_t, \tau) \leq (K_2 - K_1)e^{-r\tau} \tag{2.18}$$

or

$$0 \leq P_{K_2,T}(S_t, \tau) - P_{K_1,T}(S_t, \tau) \leq (K_2 - K_1)e^{-r\tau} \tag{2.19}$$

with $\tau = T - t$ and r denoting the time to maturity and the interest rate, respectively. If call (put) option prices are differentiable as a function of the delivery price, then by taking the limit $K_2 - K_1 \to 0$ it follows

$$-1 \leq -e^{-r\tau} \leq \frac{\partial C}{\partial K} \leq 0 \quad \text{resp.} \quad 0 \leq \frac{\partial P}{\partial K} \leq e^{-r\tau} \leq 1 \tag{2.20}$$

Proof We proof the theorem for calls since for puts the reasoning is analogous. For this we consider a portfolio A containing one call with delivery price K_1 which we compare to a duplicating portfolio B. At time t the latter portfolio consists of the following two positions:

1. A long position in one call with delivery price K_2.
2. A long position in one zero bond with face value $(K_2 - K_1)$ maturing in T.

The difference of the value of portfolios B and A at time T is shown in Table 2.5. At time T portfolio B is clearly as valuable as portfolio A which given the no-arbitrage assumption must hold at time t as well. We conclude:

$$C_{K_2,T}(S_t, \tau) + (K_2 - K_1)e^{-r\tau} \geq C_{K_1,T}(S_t, \tau). \qquad \square$$

2.2 Portfolio Insurance

A major purpose of options is hedging, i.e. the protection of investments against market risk caused by random price movements. An example for active hedging with options is the portfolio insurance. That is to strike deals in order to change at

a certain point of time the risk structure of a portfolio such that at a future point of time

- the positive profits are reduced by a small amount (which can be interpreted as an insurance premium) and in that way
- the portfolio value does not drop below a certain *floor*.

The portfolio insurance creates a risk structure of the portfolio which prevents extreme losses. For illustration purposes we consider at first a simple example.

Example 2.4 An investor has a capital of 10,500 EUR at his disposal to buy stocks whose current price is 100 EUR. Furthermore, put options on the same stock with a delivery price of $K = 100$ and a time to maturity of 1 year are quoted at a market price of 5 EUR per contract. We consider two investment alternatives.

Portfolio A: Buying 105 stocks.
Portfolio B: Buying 100 stocks for 10,000 EUR and buying 100 put options for 500 EUR.

The price of the put options can be interpreted as the premium to insure the stocks against falling below a level of 10,000 EUR. Denoting the stock price in 1 year by S_T the value of the non-insured portfolio is $105 \cdot S_T$. This portfolio bears the full market risk that the stock price drops significantly below 100 EUR. The insured portfolio, however, is at least as worth as 10,000 EUR since if $S_T < 100$ the holder exercises the put options and sells the 100 stocks for 100 EUR each.

Should the stock price increase above 100 EUR the investor does not exercise the put which thus expires worthless. By buying the put some of the capital of portfolio B is sacrificed to insure against high losses. But, while the probabilities of high profits slightly decrease, the probabilities of high losses decrease to zero. Investing in portfolio B the investor looses at most 500 EUR which he paid for the put. Table 2.6 shows the impact of the stock price S_T in 1 year on both the insured and the non-insured portfolio values and returns.

The numerous conceivable strategies to insure portfolios can be classified by the frequency with which the positions in the portfolio have to be rebalanced. Two approaches can be distinguished:

- Static strategies rebalance the portfolio positions at most once before expiration of the investment horizon.
- Dynamic strategies rebalance the portfolio positions very frequently, ideally continuously, according to certain rules.

The static strategy sketched in the previous example can be modified. Instead of hedging by means of put options the investor can chose between the following two static strategies:

Strategy 1:   The investor buys an equal number of *stocks and puts*.
Strategy 2:   The investor buys *bonds* with a face value equal to the floor he is aiming at and with the remaining money buys calls on the stock.

Table 2.6 The effect of portfolio insurance on portfolio value and return

Stock price S_T (EUR)	Non-insured portfolio		Insured portfolio		Insured portfolio in % of the non-insured portfolio
	Value (EUR)	Return % p.a.	Value (EUR)	Return % p.a.	
50	5,250	−50	10,000	−4.8	190
60	6,300	−40	10,000	−4.8	159
70	7,350	−30	10,000	−4.8	136
80	8,400	−20	10,000	−4.8	119
90	9,450	−10	10,000	−4.8	106
100	10,500	0	10,000	−4.8	95
110	11,550	+10	11,000	+4.8	95
120	12,600	+20	12,000	+14.3	95
130	13,650	+30	13,000	+23.8	95
140	14,700	+40	14,000	+33.3	95

All approaches commonly practiced rely on modifications of the above basic strategies. While following the first strategy it is the put which guarantees that the invested capital does not drop below the floor, applying the second strategy it is the bond which insures the investor against falling prices. The stocks, followed by the calls make up for the profits in the event of rising prices. The similarity of both strategies follows from the put–call parity, see Theorem 2.3.

Before deciding on what kind of portfolio insurance will be used, some points need to be clarified:

1. Which financial instruments are provided by the market, and what are their characteristics (coupons, volatilities, correlation with the market etc.)?
2. What knowledge does the investor have about:

   – the composition of the portfolio (which financial instruments),
   – the amount of capital to invest,
   – the investment horizon,
   – the floor (lower bound of the portfolio value) or rather the minimum return he is aiming at the end of the investment. Given the floor F and the capital invested V the possibly negative minimum return of a 1-year investment is given by $\rho = \frac{F-V}{V}$.

Strategies 1 and 2 (described above) are illustrated in another example.

Example 2.5 Starting with the hypothesis that the investor has decided to invest in stock, depending on the type of return of the object, we can distinguish two cases (for negative returns, as storage costs of real values for example, the approach can be applied analogously):

(i) continuous dividend yield d
(ii) ex ante known discrete yields with a time 0 discounted total value of D_0.

Table 2.7 Data of Example 2.5

Data of Example 2.5	
Current point of time t	0
Available capital V	100,000 EUR
Target floor F	95,000 EUR
Investment horizon T	2 years
Current stock price S_0	100 EUR
Continuously compounded annual interest rate r	0.10
Annual stock volatility σ	0.30
Dividends during time to maturity	
Case (i): continuous dividends d	0.02
Case (ii): dividends with present value D_0	5 EUR

Data relating to the example is shown in Table 2.7. The volatility can be interpreted as a measure of variability of the stock price. The notion of volatility is an essential part of option pricing and will be looked at extensively later. Placing our considerations at the beginning, $t = 0$ of the investment, the time to maturity is $\tau = T - t = T$. For both strategies the goal is to determine the number n of stocks and/or (European) options and their delivery price K.

Case (i) The stock pays a continuous dividend at rate $d = 2\%$ p.a. which the investor reinvests immediately. At maturity T, the position in the stock has grown from n stocks to ne^{dT}. τ denotes the time to maturity, here at $t = 0\tau = T$. Thus, for strategy 1 the investor needs to buy in $t = 0$ the same number of put options. Since the amount he wants to invest in $t = 0$ is V it must hold

$$n \cdot S_0 + ne^{d\tau} \cdot P_{K,T}(S_0, \tau) = V. \tag{2.21}$$

The investor chooses the put options delivery price K such that his capital after 2 years does not drop below the floor he is aiming for F. That is, exercising the puts in time T (if $S_T \leq K$) must give the floor F which gives the second condition

$$ne^{d\tau} \cdot K = F \iff n = \frac{F}{K}e^{-d\tau}. \tag{2.22}$$

Substituting Eq. (2.22) into Eq. (2.21) gives

$$e^{-d\tau} S_0 + P_{K,T}(S_0, \tau) - \frac{V}{F} \cdot K = 0. \tag{2.23}$$

Due to the Black–Scholes pricing formula for European options that will be derived later in Sect. 6.2 the put price is expressed as a function of the parameter K. The delivery price which the investor is looking for can be computed by solving Eq. (2.23) numerically, for example by means of the Newton–Raphson method. In this case, K is equal to 99.56. To be sure that the capital value does not

Table 2.8 The effect of a portfolio insurance in case (i) on portfolio value and return ◓ SFEoptman

| Stock price S_T (EUR) | Non-insured portfolio | | Insured portfolio | | Insured portfolio in % |
	Value (EUR)	Return % p.a.	Value (EUR)	Return % p.a.	of the non-insured portfolio
70	72,857	−27	95,000	−5	130
80	83,265	−17	95,000	−5	114
90	93,673	−6	95,000	−5	101
100	104,081	+4	95,400	−5	92
110	114,489	+15	104,940	+5	92
120	124,897	+25	114,480	+14	92
130	135,305	+35	124,020	+24	92
140	145,714	+46	133,560	+34	92

drop below the floor $F = 95,000\,\text{EUR}$ the investor buys $n = \frac{F}{K}e^{-d\tau} = 916.6$ stocks and $n \cdot e^{d\tau} = 954$ puts with delivery price $K = 99.56$. The price of the put option given by the Black–Scholes formula is 8.72 EUR/put. ◓ SFEexerput

Following the corresponding strategy 2 the investor invests $Fe^{-r\tau} = 77,779.42\,\text{EUR}$ in bonds at time 0 which gives compounded to time T exactly the floor $F = 95,000\,\text{EUR}$. For the remaining capital of $V - Fe^{-r\tau} = 22,220.58\,\text{EUR}$ the investor buys 954 calls with delivery price $K = 99.56$ which has a price of 23.29 EUR/call according to the Black–Scholes formula. From the put–call parity follows the equivalence of both strategies, i.e. that both portfolios consisting of stocks and puts respectively zero bonds and calls have at each time t the same value:

$$n \cdot S_t + ne^{d\tau} P_{K,T}(S_t, \tau) = nKe^{-b\tau} + ne^{d\tau} C_{K,T}(S_0, \tau) \qquad (2.24)$$

where $\tau = T - t, b = r - d$ and $nKe^{-b\tau} = Fe^{-r\tau}$ due to Eq. (2.22). Table 2.8 shows the risk decreasing effect of the insurance.

Case (ii)   Until maturity the stock pays dividends with a time 0 discounted total value $D_0 = 5\,\text{EUR}$ which, after distribution, are immediately invested in bonds. At time T the dividend yield has a compounded value of $D_T = D_0 e^{r\tau} = 6.107\,\text{EUR}$ where $\tau = T$ denotes the time to maturity. Reasoning as in case (i) and taking the dividend D_T into account the investor buys n stocks respectively n puts, and obtains the following equations

$$n \cdot S_0 + n P_{K,T}(S_0 - D_0, \tau) = V \qquad (2.25)$$

and

$$nK + nD_T = F. \qquad (2.26)$$

Table 2.9 The effect of a portfolio insurance in case (ii) on portfolio value and return

Stock price S_T	Non-insured portfolio		Insured portfolio		Insured portfolio in %
	Value	Return	Value	Return	of the non-insured
(EUR)	(EUR)	% p.a.	(EUR)	% p.a.	portfolio
70	76,107	−24	94,996	−5	125
80	86,107	−14	94,996	−5	110
90	96,107	−4	94,996	−5	99
96.42	102,527	+3	94,996	−5	93
100	106,107	+6	98,313	−2	93
110	116,107	+16	107,579	+8	93
120	126,107	+26	116,844	+17	93
130	136,107	+36	126,110	+26	93
140	146,107	+46	135,375	+35	93

The subtraction of the cash dividend D_0 from the stock price S_0 in the option price formula cannot be justified until we introduce the binomial model (Chap. 7). Briefly, in a perfect market the stock price decreases instantaneously by the amount of the distributed dividend. Otherwise, an arbitrage opportunity arises. Substituting Eq. (2.26) into Eq. (2.25) gives:

$$S_0 + P_{K,T}(S_0 - D_0, \tau) - \frac{V}{F} \cdot (K + D_T) = 0 \qquad (2.27)$$

Solving the equations analogously as in case (i) the number n of stocks and puts and the delivery price K for strategy 1 are obtained:

$$K = 96.42 \quad \text{und} \quad n = \frac{F}{K + D_T} = 926.55$$

For strategy 2 the investor buys 926.55 calls at a price of 23.99 EUR/call with a delivery price $K = 96.42$. He invests $95{,}000e^{-r\tau} = 77{,}779.42$ in bonds. For case (ii) the effect of the portfolio insurance for both strategies is shown in Table 2.9 taking into account the time T compounded total dividend (Fig. 2.1).

The example shows how a portfolio insurance can be carried out by means of options in principle. In practice, the following problems frequently occur:

– The number n of stocks and options is not an integer. In a perfect financial market financial instruments are perfectly divisible, in reality, however, this is not the case. The error resulting from rounding up or down to the closest integer can be neglected only in large portfolios.
– Puts and calls traded on the market do not cover the whole range of delivery prices. Thus, options with the computed delivery price may not be available. Furthermore, options typically expire in less than 1 year which makes static strategies only viable for long-term investments.
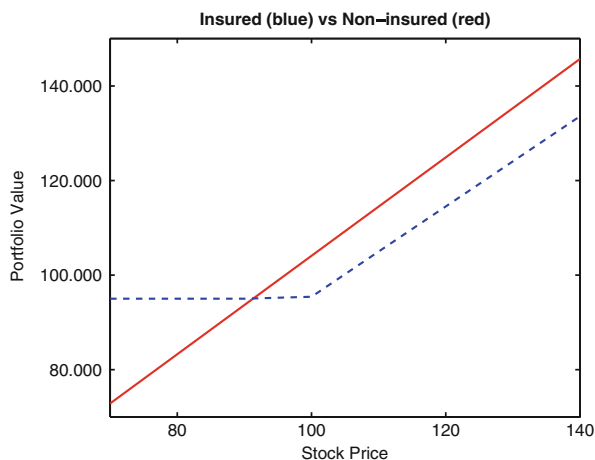
Fig. 2.1 The effect of a portfolio insurance: while the *dotted line* represents the value of the insured portfolio as a function of the stock price, the *straight line* represents the value of the non-insured portfolio as a function of the stock price   **Q** SFEoptman

– Finally, the market firstly provides American options, which are more expensive than European options, and which are sufficient to insure the portfolio. The additional exercise opportunities offered by American options are only of interest if the investor might need to close the portfolio early.

The fact that options are not traded at all delivery prices suggests that they should be reproduced by the delta hedge process described in Chap. 6. However, since a dynamic strategy is involved, transaction costs need to be taken into account and give rise to other problems. Finally, it is worth pointing out that when insuring large portfolios it is convenient to hedge by means of index options, i.e. puts and calls on the DAX for example. This is not only beneficial from a cost saving point of view but also because index options replace options on a single underlying transaction which are not traded on the market. To compute the exact effect of an index option hedge, the correlation of the portfolio with the index is needed. The latter correlation is obtained from the correlations of each individual stock contained in the portfolio with the index. Besides detailed model assumptions such as the Capital Asset Pricing Model (CAPM see Sect. 11.4.1) which among others concern the stock, returns are required.

2.3   Binary One-Period Model

The simplest of the option pricing formulae is the binomial option pricing formula. Here we take a look at a very simple model: *the binary one-period model*. The material in this section is only intended to be introductory. More details on the use of numerical procedures involving binomial trees are given in Chap. 7.

$$S_T = S^u$$
$$C_T = S_T - K$$

p

$$S_0$$
$$C_0$$

$1 - p$

$$S_T = S^d$$
$$C_T = 0$$

$$S^u \geq K \geq S^d$$
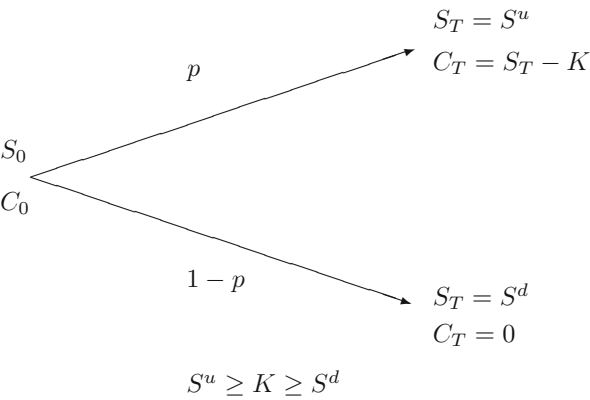
Fig. 2.2 Stock and option prices in the binary one-period model

Table 2.10 Cash flow of call and portfolio of stock and zerobond

		Flow at T	
Strategy	Flow at 0	S^u	S^d
Call	$-C_0$	$S_T - K$	0
Stock + zerobond	$-(xS^o + y)$	$xS^u + y$	$xS^d + y$

Consider a stock with a price of S_0 and a European call option on the stock with a strike price K where the current price is C_0. Assume that the call is being valued one period before expiration $(T = 1)$ and that the interest rate r is equal to 0 in the one-period model. We let the future stock price be one of only two values: the stock price can either increase from S_0 to S^u with probability p or decrease from S_0 to S^d with probability $(1 - p)$. If the stock price moves up to S^u, the payoff will be $S_T - K$; if the stock price moves down to S^d, the payoff will be 0, see Fig. 2.2.

Our goal is to determine the value C_0 of the call. The following different investment possibilities exist:

1. zerobond (with interest rate $r = 0$),
2. S_0 the current value of the stock,
3. C_0 (C_T) the price of European call at time 0 (T) with strike price K.

In order to value the call correctly, we examine two strategies. The first one is simply to buy the call. The second strategy is to choose a certain number of stocks x and a decisive amount of a zerobond y in a way that ensures the same payoff as the call at time T. Table 2.10 shows the cash flows for both strategies.

In order to duplicate the payoff of the "buy-a-call" strategy, both cash flows must match whether the stock price goes up or down:

$$xS^u + y = C^u = S_T - K \tag{2.28}$$

$$xS^d + y = C^d = 0 \tag{2.29}$$

Since the desired payoff will be identical for both strategies, the second portfolio strategy is called *hedge strategy*. Solving (2.28)–(2.29) we obtain:

$$x = \frac{C^u - C^d}{S^u - S^d} \tag{2.30}$$

$$y = \frac{S^u C^d - S^d C^u}{S^u - S^d} \tag{2.31}$$

Now we have the situation that one gets the same financial product (payment of r^u or r^d) for the price C_0 (option price) or for the price

$$xS_0 + y = S_0 \frac{C^u - C^d}{S^u - S^d} + \frac{S^u C^d - S^d C^u}{S^u - S^d} \tag{2.32}$$

$$= \frac{(S^u - S_0)C^d + (S_0 - S^d)C^u}{S^u - S^d} \tag{2.33}$$

Because the return of the option equals the return of the hedge strategy, both strategies must have the *same* initial cost:

$$xS_0 + y = C_0 \tag{2.34}$$

In fact, this yields the fair price C_0 of the call option:

$$C_0 = \frac{(S^u - S_0)C^d + (S_0 - S^d)C^u}{S^u - S^d} \tag{2.35}$$

It is worth noting that the price C_0 does not depend on p (probability of rising stock). This means clearly that we do not need to regard the probabilities of future up and down movements, because they are already incorporated in the current stock prices. In a two-period model the call price ultimately depends on the current stock price S_0, the possible value of the stock at time T (S^u and S^d) and the payoffs of the call at expiration (C^u and C^d). Simplifying the last equation we obtain:

$$C_0 = \frac{(S^u - S_0)C^d + (S_0 - S^d)C^u}{S^u - S^d}$$

$$= \frac{S^u - S_0}{S^u - S^d}C^d + \frac{S_0 - S^d}{S^u - S^d}C^u$$

$$= (1 - q)C^d + qC^u \tag{2.36}$$

$$S_T = S^u$$

$$C_T = S_T - K$$

$$q$$

$$S_0$$

$$C_0$$

$$1 - q$$

$$S_T = S^d$$

$$C_T = 0$$

$$S^u \geq K \geq S^d$$
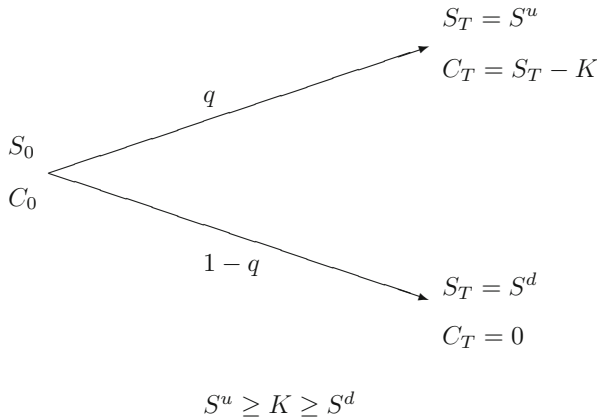
Fig. 2.3 Stock and option prices in the martingale model

with

$$q = \frac{S_0 - S^d}{S^u - S^d}. \tag{2.37}$$

Another way to look at (2.36) is to say that C_0 is the expected value of C_T with respect to the probability measure defined by q. This leads us to a second approach based on the *martingale model*. A martingale is a stochastic process where the expected value of all future values equals the current value. In the binary one-period model a martingale measure is simply given by a probability measure q such that the expected return of the share at time T is 0, i.e. the expected future stock price equals the current stock price S_0 (for interest rate $r = 0$):

$$S_0 = \mathsf{E}_q(S_T) = (1 - q)S^d + qS^u \tag{2.38}$$

where E_q indicates an expectation taken with respect to the martingale measure q. In terms of the one-period binomial tree, q is the probability of an "up" move and $(1 - q)$ the probability of a "down" move, see Fig. 2.3.

Solving (2.38), we obtain:

$$q = \frac{S_0 - S^d}{S^u - S^d}. \tag{2.39}$$

The martingale approach is now to determine the price C_0 as the expected payoff

$$\mathsf{E}_q \left[\max \left\{ S_T - K, 0 \right\} \right] = \mathsf{E}_q \left[C_T \right]. \tag{2.40}$$

Calculating this expected value of C_T, we derive the price of the call option:

$$C_0 = (1 - q)C^d + qC^u. \tag{2.41}$$

This is the same value we concluded from the former duplicating portfolio approach. In the case of interest rate $r > 0$ the option price changes to

$$C_0 = (1 + r)^{-1}\mathsf{E}_q\left[C_T\right]. \tag{2.42}$$

The martingale measure approach is analogous to Cox and Ross (1976) *risk neutral valuation* that one finds in for example Hull (2006).

Example 2.6   Consider the following binary one-period model: a stock with pricel $S_0 = 270$, a European call option on a stock with strike $K = 270$, a zerobond with price $1 \cdot (1 + r)^{-1}$ with interest rate $r = 5\%$ (i.e. price of zerobond corresponds to $1/1.05$). The stock can either increase to 300 or decrease to 250:

$$270 \begin{array}{c} \nearrow 300 \\ \searrow 250 \end{array}$$
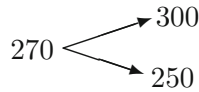
In the Hedge portfolio approach one calculates:

1. $S_T = 300, C_T = 30$ i.e. $300x + y = 30$
2. $S_T = 250, C_T = 0$ i.e. $250x + y = 0$

Solving this linear system we obtain $x = 0.6, y = -150$. Thus, selling 150 bonds and buying 0.6 stocks duplicates the payoff of the call. Therefore the price of the call is:

$$x S_0 + y = C_0$$

$$0.6 \cdot 270 + (-150) \cdot \frac{1}{1.05} = 19.14$$

and the corresponding probability is $q = \frac{270-250}{300-250} = 0.4$.

In the martingale measure approach S has to be a martingale under the measure q, i.e. the expected value of the discounted process $S_T/(1 + r)$ equals S_0:

$$S_0 = 270 = \mathsf{E}_q\left[\frac{S_T}{1 + r}\right] = q \cdot \frac{300}{1.05} + (1 - q) \cdot \frac{250}{1.05}$$

This leads to $q = 0.67$. The price of the call is therefore

$$C_0 = \mathsf{E}_q\left[\frac{C_T}{1 + r}\right]$$

$$= q \cdot \frac{30}{1.05} + (1 - q) \cdot \frac{0}{1.05}$$

$$= 0.67 \cdot \frac{30}{1.05}$$
$$= 19.14$$

We can see that both approaches provide the same results.

We see from the previous example that there are two different approaches of calculation for the option price:

- calculation of the hedge strategy, i.e. determine portfolio of stocks and bonds that duplicates the cash flow of the option;
- martingale measure approach, i.e. calculation of the expected option payment in the martingale model.

2.4 Recommended Literature

The fundamental arbitrage relations of financial derivatives can be found in every modern finance textbook, as for example Hull (2006). In principle each option pricing theory is based on theses relations, as it is the case for the model of Black and Scholes (1973) for example, see also the fundamental article of Merton (1973). The idea of portfolio assurance was introduced by Leland (1980). It is nowadays covered above all by practical risk management textbooks as Das (1997). The hedge portfolio approach and the martingale approach used in the binary one-period model for calculation of option prices are discussed in Elton, Gruber, Brown and Goztmann (2002) and Ross, Westerfield and Jaffe (2005). The martingale measure approach analogous to the classic work of Cox and Ross (1976) is presented in Yan (1999) and Hull (2006).

2.5 Exercises

Exercise 2.1 *Consider an American call option with a 40 USD strike price on a specific stock. Assume that the stock sells for 45 USD a share and pays no dividends at expiration. The option sells for 5 USD 1 year before expiration. Describe an arbitrage opportunity, assuming the annual interest rate is 10 %.*

Exercise 2.2 *Consider a European call option on a stock with current spot price $S_0 = 20$, exercise price $K = 18$ and time to maturity of 6 months. The annual risk-free rate is $r = 10$ %. What is the upper and lower bound (limit) of the price of the call option?*

Exercise 2.3 *Consider a European put option on a stock with the same market conditions as in Exercise 2.2. What is the upper and lower bound on the price of the put option?*

Exercise 2.4 *Consider again the European call option discussed in Exercise 2.3. What would be the bounds for this option if the stock underlying the option would pay out a dividend $D = 2$ USD?*

Exercise 2.5 *Consider again the European put option discussed in Exercise 2.4. What would be the bounds for this option if the stock underlying the option would pay out a dividend $D = 2$ USD?*

Exercise 2.6 (Convexity of the Price of a Put Option) *It was stated that the price of an American or European option is a convex function of its exercise price. Prove the validity of this statement for put options.*

Exercise 2.7 (The Price of a Put option) *Consider two put options on the same underlying asset with the same maturity. The exercise prices and the prices of these two options are as follows:*

Exercise price	Option price
$K_1 = 80$ EUR	38.2 EUR
$K_2 = 50$ EUR	22.6 EUR

There is a third put option on the same underlying asset with the same maturity. The exercise price of this option is 60 EUR. What can be said about the price of this option?

Exercise 2.8 *The present price of a stock without dividends is 250 EUR. The market value of a European call with strike price 235 EUR and maturity 180 days is 21.88 EUR. The annual risk-free rate is 1 %.*

(a) *Assume that the market price for a European put with a strike price of 235 EUR is 5.25 EUR. Show that this is inconsistent with Put–call parity.*
(b) *Describe how you can take advantage of this situation by finding a combination of purchases and sales which provides an instant profit with no liability 180 days from now.*

Exercise 2.9 *Prove that the following relationship holds, using no-arbitrage theory.*

$$F_0(T_2) = F_0(T_1)(1 + r)^{T_2 - T_1} - FV(D)$$

where $F_0(T)$ is today's futures price for delivery at time T, $T_2 > T_1$, and FV(D) is the future value to which any dividends paid between T_1 and T_2 will grow if invested risklessly until time T_2.

Exercise 2.10 *Suppose currently the DAX index is at 3200, and the DAX index futures with maturity in June matures exactly 6 months from now, and are priced at 3220.*

(a) *If the current interest rate is 2.5 % semi-annually, and the dividend rate of the index is 1.5 % semi-annually, is there an arbitrage opportunity available? If there is, calculate the profits available on the strategy.*
(b) *Is there an arbitrage opportunity if the interest rate that can be earned on the proceeds of a short sale is only 2 % semi-annually?*

Exercise 2.11 *The present exchange rate between USD and EUR is 1.22 USD/EUR. The price of a domestic 180-day Treasury bill is 99.48 USD per 100 USD face value. The price of the analogous Euro instrument is 99.46 EUR per 100 EUR face value.*

(a) *What is the theoretical 180-day forward exchange rate?*
(b) *Suppose the 180-day forward exchange rate available in the marketplace is 1.21 USD/EUR. This is less than the theoretical forward exchange rate, so an arbitrage is possible. Describe a risk-free strategy for making money in this market. How much does it gain, for a contract size of 100 EUR?*

Exercise 2.12 (Valuation of a Forward Contract) *What is the value of a forward contract with $K = 100$, $S_t = 95$, $r = 10\%$, $d = 5\%$ and $\tau = 0.5$?*

Exercise 2.13 *Suppose there is a 1-year future on a stock-index portfolio with the future price of 2,530 USD. The stock index currently is 2500, and an 2,500 USD investment in the index portfolio will pay a year-end dividend of 40 USD. Assume that the 1-year risk-free interest rate is 3 %.*

(a) *Is this future contract mispriced?*
(b) *If there is an arbitrage opportunity, how can an investor exploit it using a zero-net-investment arbitrage portfolio?*
(c) *If the proceeds from the short sale of the shares are kept by the broker (you do not receive interest income from the fund), does this arbitrage opportunity still exist?*
(d) *Given the short sale rules, how high and how low can the futures price be without arbitrage opportunities?*

Exercise 2.14 *Suppose there is a 1-year future on a stock-index portfolio with future price 2,530 EUR. The stock index currently is 2500, and an 2,500 EUR investment in the index portfolio will pay an year-end dividend 40 EUR. Assume that 1-year risk-free interest rate is 3 %, compounded continuously.*

(a) *Is this future contract mispriced?*
(b) *How can you construct a zero-net-investment arbitrage portfolio to exploit this arbitrage opportunity?*
(c) *If the proceeds from the short sale of the shares are kept by the broker (you do not receive interest income from the fund), does this arbitrage opportunity still exist?*

(d) Given the short sale rules, how high and how low can the futures price be without arbitrage opportunities?

Exercise 2.15 *The price of a stock is 50 USD at time $t = 0$. It is estimated that the price will be either 25 USD or 100 USD in $t = 1$ (it is assumed that no dividends are paid). A European call with an exercise price of 50 USD is worth C at time $t = 0$. This call will expire at time $t = 1$. The market interest rate is 25 %.*

(a) What return can the owner of the following hedge portfolio expect in $t = 1$ for following actions?

   (i) Sell 3 calls for C each
   (ii) Buy 2 stocks for 50 USD each
   (iii) Borrow 40 USD at the market interest rate

(b) Calculate the price C of a call.

Exercise 2.16 *The stock price of CASE Airlines is 100 USD. During each of the following two periods the price will either rise by $1/10$ or fall by $1/11$. The risk-free interest rate is 5 % in each period. Calculate the price of a European put option with a time to expiration of two periods and an exercise price of 102 USD.*

Chapter 3
Basic Concepts of Probability Theory

3.1 Real Valued Random Variables

Thanks to Newton's laws, dropping a stone from a height of $10\,\mathrm{m}$, the point of time of its impact on the ground is known before executing the experiment. Quantities in complex systems (such as stock prices at a certain date, daily maximum temperature at a certain place) are, however, not deterministically predictable, although it is known which values are more likely to occur than others. Contrary to the falling stone, data which cannot be described successfully by deterministic mechanism can be modelled by random variables.

Let X be such a random variable that (as a model for stock prices) takes values in real time. The appraisal of which values of X are more and which are less likely is expressed by the *probability* of events as $\{a < X < b\}$ or $\{X \le b\}$. The set of all probabilities

$$\mathrm{P}(a \le X \le b)\,, \quad -\infty < a \le b < \infty\,,$$

determines the *distribution* of X. In other words, the distribution is defined by the probabilities of all events which depend on X. In the following, we denote the probability distribution of X by $\mathcal{L}(X)$.

The probability distribution is uniquely defined by the *cumulative probability distribution*

$$F(x) = \mathrm{P}(X \le x)\,, \quad -\infty < x < \infty\,.$$

$F(x)$ monotonously increases and converges for $x \to -\infty$ to 0, and for $x \to \infty$ to 1. If there is a function p, such that the probabilities can be computed by means of an integral

$$\mathrm{P}(a < X < b) = \int_a^b p(x)dx\,,$$

p is called a *probability density*, or briefly density of X. Then the cumulative distribution function is a primitive of p:

$$F(x) = \int_{-\infty}^{x} p(y)dy.$$

For small h it holds:

$$P(x - h < X < x + h) \approx 2h \cdot p(x).$$

Thus $p(x)$ is a measure of the likelihood that X takes values close to x.

The most important family of distributions with densities is the *normal distribution* family. It is characterized by two parameters μ, σ^2. The densities are given by

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

$$\varphi(x) = \varphi_{0,1}(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}.$$

The distribution with density $\varphi(x)$ is called *standard normal distribution*. "X a normal random variable with parameters μ, σ^2" is commonly abbreviated by "X is N(μ, σ^2) distributed". The cumulative distribution function of a standard normal distribution is denoted by Φ and it holds:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)dy.$$

If X is N(μ, σ^2) distributed, then the centred and scaled random variable $(X - \mu)/\sigma$ is standard normal distributed. Therefore, its cumulative distribution function is given by:

$$F(x) = P(X \le x) = P\left(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

A distribution which is of importance in modelling stock prices is the *lognormal distribution*. Let X be a positive random variable whose natural logarithm $\log(X)$ is normally distributed with parameters μ, σ^2. We say that X is lognormally distributed with parameters μ, σ^2. Its cumulative distribution function follows directly from the above definition:

$$F(x) = P(X \le x) = P(\log X \le \log x) = \Phi\left(\frac{\log x - \mu}{\sigma}\right), \quad x > 0.$$

Deriving $F(x)$ once, we obtain its density function with parameters μ, σ^2:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{x}\exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\} = \frac{1}{\sigma x}\varphi\left(\frac{\log x - \mu}{\sigma}\right), \quad x > 0.$$

If X is a random variable that takes only finite or countably infinite values $x_1, \ldots, x_n$, then X is said to be a discrete random variable and its distribution is fully determined by the probabilities:

$$P(X = x_k), \quad k = 1, \ldots, n.$$

The simplest discrete random variables take only 2 or 3 values, for example ± 1 or $-1, 0, +1$. They constitute the basis of binomial or trinomial trees which can be used to construct discrete random processes in computers. Such tree methods are reasonable approximations to continuous processes which are used to model stock prices.

In this context, *binomially distributed* random variables appear. Let $Y_1, \ldots, Y_n$ be independent random variables taking two values, 0 or 1, with probabilities

$$p = P(Y_k = 1), \ 1 - p = P(Y_k = 0), \quad k = 1, \ldots, n.$$

We call such random variables *Bernoulli distributed* with parameter p. The number of ones appearing in the sample $Y_1, \ldots, Y_n$ equals the sum $X = \sum_{k=1}^{n} Y_k$ which is binomial distributed with parameters n, p :

$$X = \sum_{k=1}^{n} Y_k, \ P(X = m) = \binom{n}{m}p^m(1-p)^{n-m}, \quad m = 0, \ldots, n.$$

<div align="right">🔍 SFEBinomial</div>

Instead of saying X is binomial distributed with parameters n, p, we use the notation "X is $B(n, p)$ distributed". Hence, a Bernoulli distributed random variable is $B(1, p)$ distributed.

If n is large enough, a $B(n, p)$ distributed random variable can be approximated by a $N\{np, np(1 - p)\}$ distributed random variable Z, in the sense that

$$P(a < X < b) \approx P(a < Z < b). \tag{3.1}$$

The central limit theorem is more precise on that matter. In classical statistics it is used to avoid, for large n, the tedious calculation of binomial probabilities. Conversely, it is possible to approximate the normal distribution by an easy simulated binomial tree.   🔍 SFEclt

3.2 Expectation and Variance

The mathematical *expectation* or the *mean* $E[X]$ of a real random variable X is a measure for the location of the distribution of X. Adding to X a real constant c, it holds for the expectation: $E[X + c] = E[X] + c$, i.e. the location of the distribution is translated. If X has a density $p(x)$, its expectation is defined as:

$$E(X) = \int_{-\infty}^{\infty} xp(x)dx.$$

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

Let $X_1, \ldots, X_n$ be a sample of identically independently distributed (i.i.d.) random variables (see Sect. 3.4) having the same distribution as X, then $E[X]$ can be estimated by means of the sample mean:

$$\hat{\mu} = \frac{1}{n} \sum_{t=1}^{n} X_t.$$

A measure for the dispersion of a random variable X around its mean is given by the *variance* $Var(X)$:

$$Var(X) = E[(X - EX)^2]$$

$$\text{Variance} = \text{mean squared deviation of a random variable}$$

$$\text{around its expectation.}$$

If X has a density $p(x)$, its variance can be computed as follows:

$$Var(X) = \int_{-\infty}^{\infty} (x - EX)^2 p(x)dx.$$

The integral can be infinite. There are empirical studies giving rise to doubt that some random variables appearing in financial and actuarial mathematics and which model losses in highly risky businesses dispose of a finite variance.

As a quadratic quantity the variance's unit is different from that of X itself. It is better to use the standard deviation of X which is measured in the same unity as X:

$$\sigma(X) = \sqrt{Var(X)}.$$

Given a sample of i.i.d. variables $X_1, \ldots, X_n$ which have the same distribution as X, the sample variance can be estimated by:

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{t=1}^{n} (X_t - \hat{\mu})^2.$$

A $N(\mu, \sigma^2)$ distributed random variable X has mean μ and variance σ^2. The 2σ area around μ contains with probability of a little more than 95 % observations of X:

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \approx 0.95.$$

A lognormally distributed random variable X with parameters μ and σ^2 has mean and variance

$$E(X) = \exp\{\mu + \frac{1}{2}\sigma^2\}, \quad Var(X) = \exp\{2\mu + \sigma^2\}\{\exp(\sigma^2) - 1\}.$$

A $B(n, p)$ distributed variable X has mean np and variance $np(1 - p)$. The approximation (3.1) is chosen such that the binomial and normal distribution have identical mean and variance.

3.3 Skewness and Kurtosis

Definition 3.1 (Skewness) The skewness of a random variable X with mean μ and variance σ^2 is defined as

$$S(X) = \frac{E[(X - \mu)^3]}{\sigma^3}.$$

If the skewness is negative (positive) the distribution is skewed to the left (right). Normally distributed random variables have a skewness of zero since the distribution is symmetrical around the mean. Given a sample of i.i.d. variables $X_1, \ldots, X_n$, skewness can be estimated by (see Sect. 3.4)

$$\hat{S}(X) = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^3}{\hat{\sigma}^3}, \tag{3.2}$$

with $\hat{\mu}, \hat{\sigma}^2$ as defined in the previous section.

Definition 3.2 (Kurtosis) The kurtosis of a random variable X with mean μ and variance σ^2 is defined as

$$Kurt(X) = \frac{E[(X - \mu)^4]}{\sigma^4}.$$

Normally distributed random variables have a kurtosis of 3. Financial data often exhibits higher kurtosis values, indicating that values close to the mean and extreme positive and negative outliers appear more frequently than for normally distributed random variables. For i.i.d. sample kurtosis can be estimated by

$$\widehat{Kurt}(X) = \frac{\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{\mu})^4}{\hat{\sigma}^4}. \tag{3.3}$$

Example 3.1 The empirical standard deviation of monthly DAX data from 1979: 1 to 2000: 10 is $\hat{\sigma} = 0.056$, which corresponds to a yearly volatility of $\hat{\sigma} \cdot \sqrt{12} = 0.195$. Later in Sect. 6.4.5, we will explain the factor $\sqrt{12}$ in detail. The kurtosis of the data is equal to 5.428, i.e. greater than 3 which suggests a non-normality of the DAX returns.   ◘ SFEsumm

3.4 Random Vectors, Dependence, Correlation

A *random vector* $(X_1, \ldots, X_n)$ from $\mathbb{R}^n$ can be useful in describing the mutual dependencies of several random variables $X_1, \ldots, X_n$, for example several underlying stocks. The joint distribution of the random variables $X_1, \ldots, X_n$ is as in the univariate case, uniquely determined by the probabilities

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n), \quad -\infty < a_i \leq b_i < \infty, \, i = 1, \ldots, n.$$

If the random vector $(X_1, \ldots, X_n)$ has a density $p(x_1, \ldots, x_n)$, the probabilities can be computed by means of the following integrals:

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_n \leq X_n \leq b_n) = \int_{a_n}^{b_n} \ldots \int_{a_1}^{b_1} p(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

The univariate or marginal distribution of X_j can be computed from the joint density by integrating out the variables not of interest.

$$P(a_j \leq X_j \leq b_j) = \int_{-\infty}^{\infty} \ldots \int_{a_j}^{b_j} \ldots \int_{-\infty}^{\infty} p(x_1, \ldots, x_n) dx_1 \ldots dx_n.$$

The intuitive notion of *independence* of two random variables X_1, X_2 is formalized by requiring:

$$P(a_1 \leq X_1 \leq b_1, a_2 \leq X_2 \leq b_2) = P(a_1 \leq X_1 \leq b_1) \cdot P(a_2 \leq X_2 \leq b_2),$$

i.e. the joint probability of two events depending on the random vector (X_1, X_2) can be factorized. It is sufficient to consider the univariate distributions of X_1 and X_2 exclusively. If the random vector (X_1, X_2) has a density $p(x_1, x_2)$, then X_1 and X_2 have densities $p_1(x)$ and $p_2(x)$ as well. In this case, independence of both random variables is equivalent to a joint density which can be factorized:

$$p(x_1, x_2) = p_1(x_1) p_2(x_2).$$

Dependence of two random variables X_1, X_2 can be very complicated. If X_1, X_2 are jointly normally distributed, their dependency structure can be rather easily quantified by their covariance:

$$\mathsf{Cov}(X_1, X_2) = \mathsf{E}[(X_1 - \mathsf{E}[X_1])(X_2 - \mathsf{E}[X_2])],$$

as well as by their correlation:

$$\mathsf{Corr}(X_1, X_2) = \frac{\mathsf{Cov}(X_1, X_2)}{\sigma(X_1) \cdot \sigma(X_2)} \, .$$

The correlation has the advantage of taking values between -1 and $+1$, the correlation is also scale invariant. For jointly normally distributed random variables, independence is equivalent to zero correlation, while complete dependence is equivalent to either a correlation of $+1$ (X_1 is large when X_2 is large) or a correlation of -1 (X_1 is large when X_2 is small).

In general, it holds for *independent* random variables $X_1, \ldots, X_n$

$$\mathsf{Cov}(X_i, X_j) = 0 \qquad \text{for } i \neq j \, .$$

This implies a useful computation rule:

$$\mathsf{Var}\left(\sum_{j=1}^{n} X_j\right) = \sum_{j=1}^{n} \mathsf{Var}(X_j) \, .$$

If $X_1, \ldots, X_n$ are independent and have all the same distribution:

$$\mathrm{P}(a \leq X_i \leq b) = \mathrm{P}(a \leq X_j \leq b) \qquad \text{for all } i, j \, ,$$

we call them *independently and identically distributed (i.i.d.)*.

3.5   Conditional Probabilities and Expectations

The *conditional probability* that a random variable Y takes values between a and b conditioned on the event that a random variable X takes values between x and $x + \Delta_x$, is defined as

$$\mathrm{P}(a \leq Y \leq b | x \leq X \leq x + \Delta_x) = \frac{\mathrm{P}(a \leq Y \leq b \, , \, x \leq X \leq x + \Delta_x)}{\mathrm{P}(x \leq X \leq x + \Delta_x)} \, , \qquad (3.4)$$

provided the denominator is different from zero. The conditional probability of events of the kind $a \leq Y \leq b$ reflects our opinion of which values are more plausible than others, given that another random variable X has taken a certain value. If Y is independent of X, the probabilities of Y are not influenced by prior knowledge about X. It holds:

$$\mathrm{P}(a \leq Y \leq b | x \leq X \leq x + \Delta x) = \mathrm{P}(a \leq Y \leq b) \, .$$

As Δx goes to 0 in Eq. (3.4), the left side of Eq. (3.4) converges heuristically to $P(a \leq Y \leq b | X = x)$. In the case of a continuous random variable X having a density p_X, the left side of Eq. (3.4) is not defined since $P(X = x) = 0$ for all x. But, it is possible to give a sound mathematical definition of the conditional distribution of Y given $X = x$. If the random variables Y and X have a joint distribution $p(x, y)$, then the conditional distribution has the density

$$p_{Y|X}(y|x) = \frac{p(x, y)}{p_X(x)} \quad \text{for} \quad p_X(x) \neq 0$$

and $p_{Y|X}(y|x) = 0$ otherwise. Consequently, it holds:

$$P(a \leq Y \leq b | X = x) = \int_a^b p_{Y|X}(y|x)dy.$$

The expectation with respect to the conditional distribution can be computed by:

$$\mathsf{E}(Y|X = x) = \int_{-\infty}^{\infty} y \, p_{Y|X}(y|x)dy \overset{\text{def}}{=} \eta(x).$$

The function $\eta(x) = \mathsf{E}(Y|X = x)$ is called the *conditional expectation of Y given* $X = x$. Intuitively, it is the expectation of the random variable Y knowing that X has taken the value x.

Considering $\eta(x)$ as a function of the random variable X the conditional expectation of Y given X is obtained:

$$\mathsf{E}(Y|X) = \eta(X).$$

$\mathsf{E}(Y|X)$ is a random variable, which can be regarded as a function having the same expectation as Y. The conditional expectation has some useful properties, which we summarize in the following theorem.

Theorem 3.1 *Let X, Y, Z be real valued continuous random variables having a joint density.*

(a) If X, Y are independent, then $\mathsf{E}(Y|X = x) = \mathsf{E}(Y)$

(b) If $Y = g(X)$ is a function of X, then

$$\mathsf{E}\{Y|X = x\} = \mathsf{E}\{g(X)|X = x\} = g(x).$$

In general, it holds for random variables of the kind $Y = Zg(X)$:

$$\mathsf{E}\{Y|X = x\} = \mathsf{E}\{Zg(X)|X = x\} = g(x)\mathsf{E}\{Z|X = x\}.$$

(c) The conditional expectation is linear, i.e. for any real numbers a, b it holds:

$$\mathsf{E}(aY + bZ|X = x) = a\mathsf{E}(Y|X = x) + b\mathsf{E}(Z|X = x).$$

(d) The law of iterated expectations: $\mathsf{E}\{\mathsf{E}(Y|X)\} = \mathsf{E}(Y).$

The concept of the conditional expectation can be generalized analogously for multivariate random vectors Y and X. Let $S_t, t = 0, 1, 2, \ldots$ be a sequence of chronologically ordered random variables, for instance as a model of daily stock prices, let $Y = S_{t+1}$ and $X = (S_t, \ldots, S_{t-p+1})^\top$, then the conditional expectation

$$\mathsf{E}(Y|X = x) = \mathsf{E}(S_{t+1}|S_t = x_1, \ldots, S_{t-p+1} = x_p)$$

represents the expected stock price of the following day $t + 1$ given the stock prices $x = (x_1, \ldots, x_p)^\top$ of the previous p days. Since the information available at time t (relevant for the future evolution of the stock price) can consist of more than only a few past stock prices, we make frequent use of the notation $\mathsf{E}(Y|\mathcal{F}_t)$ for the expectation of Y given the information available up to time t. For all t, $\mathcal{F}_t$ denotes a family of events (having the structure of a σ-algebra, i.e. certain combinations of events of $\mathcal{F}_t$ are again elements of $\mathcal{F}_t$) representing the information available up to time t. $\mathcal{F}_t$ consists of events of which it is known whether they occur up to time t or not. Since more information unveils as time evolves, we must have $\mathcal{F}_s \subset \mathcal{F}_t$ for $s < t$, see Definition 5.1. Leaving out the exact definition of $\mathsf{E}(Y|\mathcal{F}_t)$ we confine to emphasize that the computation rules given in Theorem 3.1, appropriately reformulated, can be applied to the general conditional expectation.

3.6 Recommended Literature

Durrett (1991), Ross (1994), Pitman (1997), Krengel (2000) and Jacod and Protter (2000), among others, give an introduction to probability theory. An introduction to martingale theory which is imperative for the understanding of advanced mathematical finance is given by Williams (1991).

3.7 Exercises

Exercise 3.1 *Check that the random variable X with $P(X = 1) = 1/2$, $P(X = -4) = 1/3$, $P(X = 5) = 1/6$ has skewness 0 but is not distributed symmetrically.*

Exercise 3.2 *Show that if $\mathsf{Cov}(X, Y) = 0$ it does not imply that X and Y are independent.*
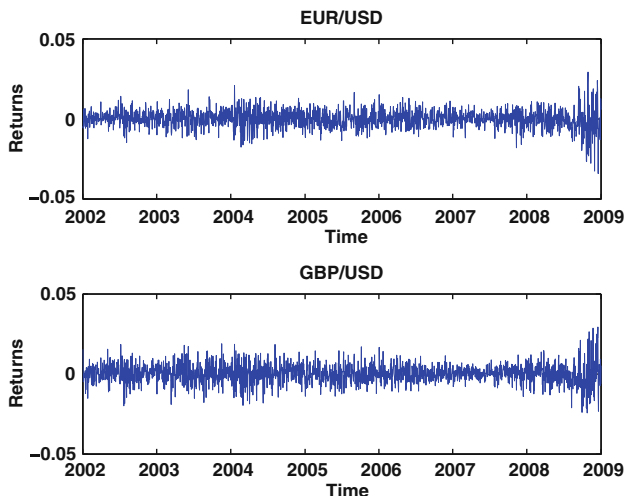
Fig. 3.1 Exchange rate returns   ⧉ SFSmvol01

Exercise 3.3 *Show that the correlation is invariant w.r.t. linear transformations.*

Exercise 3.4 *We consider a bivariate exchange rates example, two European currencies, EUR and GBP, with respect to the USD. The sample period is $01/01/2002$ to $01/01/2009$ with altogether $n = 1{,}828$ observations. Figure 3.1 shows the time series of returns on both exchange rates.*

Compute the correlation of the two exchange rate time series and comment on the sign of the correlation.

Exercise 3.5 *Compute the conditional moments $\mathsf{E}(X_2 \mid x_1)$ and $\mathsf{E}(X_1 \mid x_2)$ for the pdf of*

$$f(x_1, x_2) = \begin{cases} \frac{1}{2}x_1 + \frac{3}{2}x_2 & 0 \le x_1, x_2 \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Exercise 3.6 *Show that the function*

$$f_Y(y_1, y_2) = \begin{cases} \frac{1}{2}y_1 - \frac{1}{4}y_2 & 0 \le y_1 \le 2, \ |y_2| \le 1 - |1 - y_1| \\ 0 & \text{otherwise} \end{cases}$$

is a probability density function.

Exercise 3.7 *Prove that $\mathsf{E}X_2 = \mathsf{E}\{\mathsf{E}(X_2|X_1)\}$, where $\mathsf{E}(X_2|X_1)$ is the conditional expectation of X_2 given X_1.*

Exercise 3.8 *The conditional variance is defined as $\mathsf{Var}(Y|X) = \mathsf{E}[\{Y - \mathsf{E}(Y|X)\}^2|X]$. Show that $\mathsf{Var}(Y) = \mathsf{E}\{\mathsf{Var}(Y|X)\} + \mathsf{Var}\{\mathsf{E}(Y|X)\}$.*

Exercise 3.9 *A European car company has tested a new model and reports the consumption of gasoline (X_1) and oil (X_2). The expected consumption of gasoline is 8 L per 100 km (μ_1) and the expected consumption of oil is 1 L per 10,000 km (μ_2). The measured consumption of gasoline is 8.1 L per 100 km ($\overline{x}_1$) and the measured consumption of oil is 1.1 L per 10,000 km ($\overline{x}_2$). The asymptotic distribution of*

$$\sqrt{n}\left\{\begin{pmatrix}\overline{x}_1\\\overline{x}_2\end{pmatrix} - \begin{pmatrix}\mu_1\\\mu_2\end{pmatrix}\right\} \text{ is } N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0.1 & 0.05\\0.05 & 0.1\end{pmatrix}\right).$$

For the American market the basic measuring units are miles (1 mile $\approx$ 1.6 km) and gallons (1 gallon $\approx$ 3.8 L). The consumptions of gasoline (Y_1) and oil (Y_2) are usually reported in miles per gallon. Can you express $\overline{y}_1$ and $\overline{y}_2$ in terms of $\overline{x}_1$ and $\overline{x}_2$? Recompute the asymptotic distribution for the American market!

Exercise 3.10 *Let $X \sim N(\mu, \sigma^2)$ be a random variable, c is a constant, please calculate $\mathsf{E}[\max(X - c, 0)]$.*

Chapter 4
Stochastic Processes in Discrete Time

A *stochastic process* or random process consists of chronologically ordered random variables $\{X_t; \ t \geq 0\}$. For simplicity we assume that the process starts at time $t = 0$ in $X_0 = 0$. In this chapter, we consider exclusively processes in *discrete time*, i.e. processes which are observed at equally spaced points time $t = 0, 1, 2, \ldots$. Typical examples are daily, monthly or yearly observed economic data as stock prices, rates of unemployment or sales figures.

4.1 Binomial Processes

One of the simplest stochastic processes is an *ordinary random walk*, a process whose increments $Z_t = X_t - X_{t-1}$ from time $t - 1$ to time t take exclusively the values $+1$ or -1. Additionally, we assume the increments to be i.i.d. and independent of the starting value X_0. Hence, the ordinary random walk can be written as:

$$X_t = X_0 + \sum_{k=1}^{t} Z_k \qquad , \ t = 1, 2, \ldots \tag{4.1}$$

$X_0, Z_1, Z_2, \ldots$ independent and

$$\mathrm{P}(Z_k = 1) = p \quad , \quad \mathrm{P}(Z_k = -1) = 1 - p \qquad \text{for all } k \, .$$

Letting the process go up by u and go down by d, instead, we obtain a more general class of *binomial processes*:

$$\mathrm{P}(Z_k = u) = p \, , \quad \mathrm{P}(Z_k = -d) = 1 - p \quad \text{for all } k \, ,$$

where u and d are constant (u = up, d = down).

Linear interpolation of the points (t, X_t) reflects the time evolution of the process and is called a *path* of an ordinary random walk. Starting in $X_0 = a$, the process moves on the grid of points (t, b_t), $t = 0, 1, 2, \ldots$, $b_t = a - t, a - t + 1, \ldots, a + t$. Up to time t, X_t can grow at most up to $a + t$ (if $Z_1 = \ldots = Z_t = 1$) or can fall at least to $a - t$ (if $Z_1 = \ldots = Z_t = -1$). Three paths of an ordinary random walk are shown in Figs. 4.1 ($p = 0.5$), 4.2 ($p = 0.4$) and 4.3 ($p = 0.6$).

For generalized binomial processes the grid of possible paths is more complicated. The values which the process X_t starting in a can possibly take up to time t are given by

$$b_t = a + n \cdot u - m \cdot d, \text{ where } n, m \geq 0, \quad n + m = t.$$

If, from time 0 to time t, the process goes up n times and goes down m times then $X_t = a + n \cdot u - m \cdot d$. That is, n of t increments $Z_1, \ldots, Z_t$ takes the value u, and m increments take the value $-d$. The grid of possible paths is also called a binomial tree.

The mean of the *symmetric ordinary random walk* ($p = \frac{1}{2}$) starting in 0 ($X_0 = 0$) is for all times t equal to 0 :

$$\mathsf{E}[X_t] = 0 \qquad \text{for all } t .$$

Otherwise, the *random walk* has a *trend* or *drift*, for ($p > \frac{1}{2}$) it has a positive drift and for ($p < \frac{1}{2}$) it has a negative drift. The process grows or falls in average:
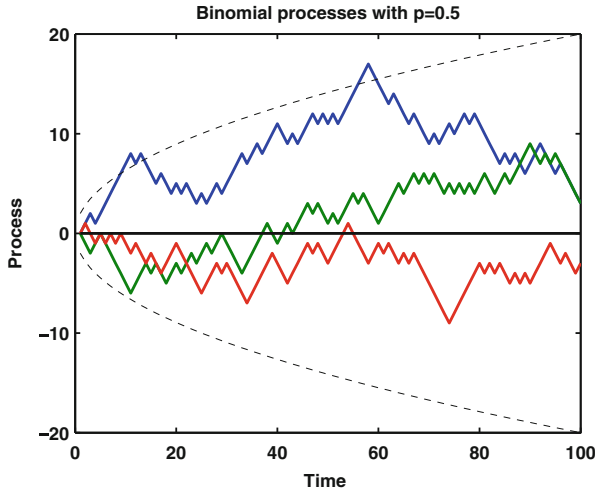
Fig. 4.1 Three paths of a symmetric ordinary random walk. (2σ)-Intervals around the drift (which is zero) are given as well   ⬛ SFEBinomp
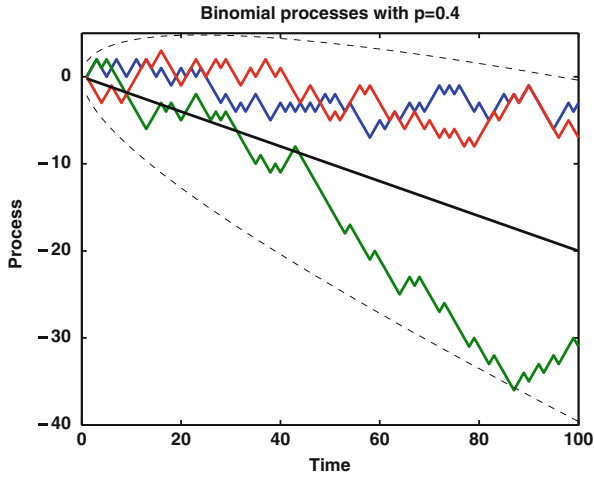
Fig. 4.2 Three paths of an ordinary random walk with $p = 0.4$. (2σ)-Intervals around the drift (which is the line with negative slope) are given as well ⬛ SFEBinomp
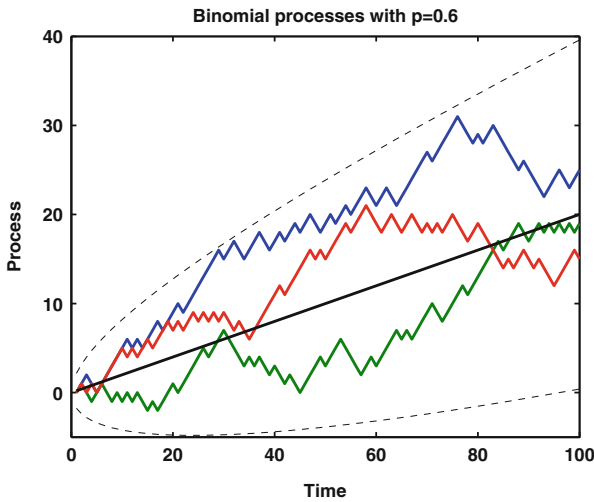
Fig. 4.3 Three paths of an ordinary random walk with $p = 0.6$. (2σ)-Intervals around the drift (which is the line with positive slope) are given as well ⬛ SFEBinomp

$$\mathsf{E}[X_t] = t \cdot (2p - 1),$$

since it holds for all increments $\mathsf{E}[Z_k] = 2p - 1$. Hence, the trend is linear in time. It is the upward sloping line in Fig. 4.3 ($p = 0.6$) and the downward sloping line in Fig. 4.2 ($p = 0.4$).

For the generalized *binomial process* with general starting value X_0 it holds analogously $\mathsf{E}[Z_k] = (u + d)p - d$ and thus:

$$\mathsf{E}[X_t] = \mathsf{E}[X_0] + t \cdot \{(u + d)p - d\}.$$

As time evolves the set of values X_t grows, and its variability increases. Since the summands in (4.1) are independent and $\mathsf{Var}(Z_k) = \mathsf{Var}(Z_1)$ for all k, the variance of X_t is given by (refer to Sect. 3.4):

$$\mathsf{Var}(X_t) = \mathsf{Var}(X_0) + t \cdot \mathsf{Var}(Z_1).$$

Hence, the variance of X_t grows linearly with time, as does the standard deviation. For the random walks depicted in Figs. 4.1 ($p = 0.5$), 4.2 ($p = 0.4$) and 4.3 ($p = 0.6$) the intervals $[\mathsf{E}[X_t] - 2\sigma(X_t); \mathsf{E}[X_t] + 2\sigma(X_t)]$ are shown as well. For large t, these intervals should contain 95 % of the realizations of processes.

The variance of the increments can be easily computed. We use the following result which leads to the binomial distribution. Define

$$Y_k = \frac{Z_k + d}{u + d} = \begin{cases} 1 & \text{if } Z_k = u \\ 0 & \text{if } Z_k = -d \end{cases}$$

or

$$Z_k = (u + d)\, Y_k - d \tag{4.2}$$

we obtain the following representation of the binomial process

$$X_t = X_0 + (u + d)\, B_t - t\, d \tag{4.3}$$

where

$$B_t = \sum_{k=1}^{t} Y_k \tag{4.4}$$

is a $B(t, p)$ distributed random variable.

Given the distribution of X_0, the distribution of X_t is specified for all t. It can be derived by means of a simple transformation of the binomial distribution $B(t, p)$. From Eqs. (4.2) to (4.4) we obtain for $X_0 = 0$:

$$\mathsf{Var}(X_t) = t(u + d)^2\, p(1 - p)$$

and for large t the distribution of X_t can be approximated by:

$$\mathcal{L}(X_t) \approx \mathsf{N}(t\{(u + d)p - d\},\, t(u + d)^2\, p(1 - p)).$$
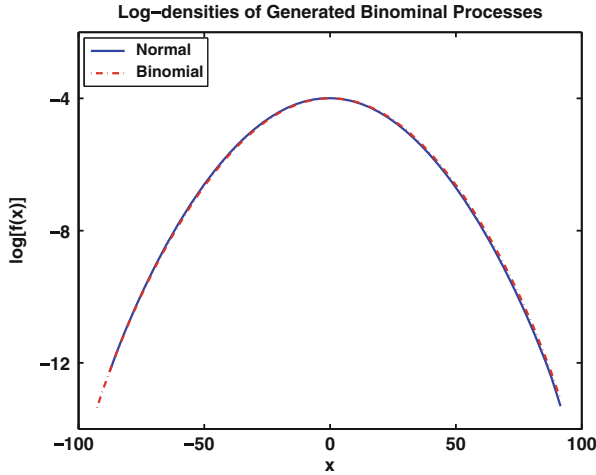
Fig. 4.4 The distribution of 100 paths of an ordinary symmetric random walk of length 100 and a kernel density estimation of 100 normally distributed random variables all shown on a log scale

SFEbinomv

For $p = \frac{1}{2}, u = d = \Delta x$, the following approximation holds for $\mathcal{L}(X_t)$:

$$N(0, t \cdot (\Delta x)^2).$$

Figure 4.4 shows the fit for $t = 100$.

4.2   Trinomial Processes

In contrast to binomial processes, a *trinomial process* allows a quantity to stay constant within a given period of time. In the latter case, the increments are described by:

$$P(Z_k = u) = p \,, \ P(Z_k = -d) = q \,, \ P(Z_k = 0) = r = 1 - p - q \,,$$

and the process X_t is again given by:

$$X_t = X_0 + \sum_{k=1}^{t} Z_k$$

where $X_0, Z_1, Z_2, \ldots$ are mutually independent. To solve the Black–Scholes equation, some algorithms use trinomial schemes with time and state dependent probabilities p, q and r. Figure 4.5 shows five simulated paths of a trinomial process with $u = d = 1$ and $p = q = 0.25$.
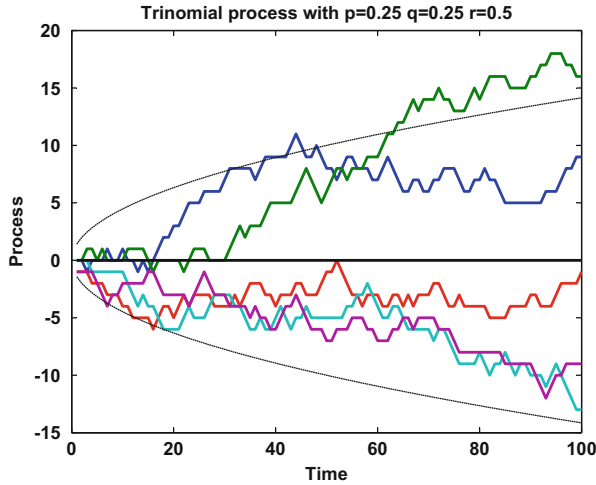
Fig. 4.5 Five paths of a trinomial process with $p = q = 0.25$. (2σ)-Intervals around the trend (which is zero) are given as well   Q SFETrinomp

The exact distribution of X_t cannot be derived from the binomial distribution but for the trinomial process a similar relations hold:

$$\mathsf{E}[X_t] = \mathsf{E}[X_0] + t \cdot \mathsf{E}[Z_1] = \mathsf{E}[X_0] + t \cdot (pu - qd)$$

$$\mathsf{Var}(X_t) = \mathsf{Var}(X_0) + t \cdot \mathsf{Var}(Z_1), \text{ where}$$

$$\mathsf{Var}(Z_1) = p(1 - p)u^2 + q(1 - q)d^2 + 2pq\, ud \; .$$

For large t, X_t is approximately $\mathsf{N}\{\mathsf{E}[X_t], \mathsf{Var}(X_t)\}$-distributed.

4.3 General Random Walks

Binomial and trinomial processes are simple examples for *general random walks*, i.e. stochastic processes $\{X_t; \, t \geq 0\}$ satisfying:

$$X_t = X_0 + \sum_{k=1}^{t} Z_k \, , \qquad t = 1, 2, \ldots$$

where X_0 is independent of $Z_1, Z_2, \ldots$ which are i.i.d. The increments have a distribution of a real valued random variable. Z_k can take a finite or countably infinite number of values; but it is also possible for Z_k to take values out of a continuous set.

As an example, consider a *Gaussian random walk* with $X_0 = 0$, where the finitely many $X_1, \ldots, X_t$ are jointly normally distributed. Such a random walk can be constructed by assuming identically, independently and normally distributed increments. By the properties of the normal distribution, it follows that X_t is $N(\mu t, \sigma^2 t)$-distributed for each t. If $X_0 = 0$ and $\mathsf{Var}(Z_1)$ is finite, it holds approximately for all random walks for t large enough:

$$\mathcal{L}(X_t) \approx N(t \cdot \mathsf{E}[Z_1], \ t \cdot \mathsf{Var}(Z_1)).$$

This result follows directly from the central limit theorem for i.i.d. random variables.

Random walks are processes with *independent increments*. That means, the increment Z_{t+1} of the process from time t to time $t + 1$ is independent of the past values $X_0, \ldots, X_t$ up to time t. In general, it holds for any $s > 0$ that the increment of the process from time t to time $t + s$

$$X_{t+s} - X_t = Z_{t+1} + \ldots + Z_{t+s}$$

is independent of $X_0, \ldots, X_t$. It follows that the best prediction, in terms of mean squared error, for X_{t+1} given $X_0, \ldots, X_t$ is just $X_t + \mathsf{E}[Z_{t+1}]$. As long as the price of only one stock is considered, this prediction rule works quite well. As already as 100 years ago, Bachelier postulated (assuming $\mathsf{E}[Z_k] = 0$ for all k): "The best prediction for the stock price of tomorrow is the price of today."

Processes with independent increments are also *Markov-processes*. In other words, the future evolution of the process in time t depends exclusively on X_t, and the value of X_t is independent of the past values $X_0, \ldots, X_{t-1}$. If the increments Z_k and the starting value X_0, and hence all X_t, can take a finite or countably infinite number of values, then the *Markov-property* is formally expressed by:

$$P(a_{t+1} < X_{t+1} < b_{t+1} | X_t = c, \ a_{t-1} < X_{t-1} < b_{t-1}, \ldots, a_0 < X_0 < b_0)$$

$$= P(a_{t+1} < X_{t+1} < b_{t+1} | X_t = c).$$

If $X_t = c$ is known, additional information about $X_0, \ldots, X_{t-1}$ does not influence the opinion about the range in which X_t will probably fall.

4.4 Geometric Random Walks

The essential idea underlying the random walk for real processes is the assumption of mutually independent increments of the order of magnitude for each point in time. However, economic time series in particular do not satisfy the latter assumption. Seasonal fluctuations of monthly sales figures for example are in *absolute terms* significantly greater if the yearly average sales figure is high. By contrast, the relative or percentage changes are stable over time and do not depend on the

current level of X_t. Analogously to the random walk with i.i.d. absolute increments $Z_t = X_t - X_{t-1}$, a *geometric random walk* $\{X_t; t \geq 0\}$ is assumed to have i.i.d. relative increments

$$R_t = \frac{X_t}{X_{t-1}}, \quad t = 1, 2, \dots .$$

For example, a geometric *binomial random walk* is given by

$$X_t = R_t \cdot X_{t-1} = X_0 \cdot \Pi_{k=1}^t R_k \tag{4.5}$$

where $X_0, R_1, R_2, \dots$ are mutually independent and for $u > 1$, $d < 1$:

$$P(R_k = u) = p, \ P(R_k = d) = 1 - p .$$

Given the independence assumption and $\mathsf{E}[R_k] = (u - d)p + d$ it follows from Eq. (4.5) that $\mathsf{E}[X_t]$ increases or decreases exponentially as the case may be $\mathsf{E}[R_k] > 1$ or $\mathsf{E}[R_k] < 1$:

$$\mathsf{E}[X_t] = \mathsf{E}[X_0] \cdot (\mathsf{E}[R_1])^t = \mathsf{E}[X_0] \cdot \{(u - d)p + d\}^t.$$

If $\mathsf{E}[R_k] = 1$ the process is on average stable, which is the case for

$$p = \frac{1 - d}{u - d} .$$

For a recombining process, i.e. $d = \frac{1}{u}$, this relationship simplifies to

$$p = \frac{1}{u + 1} .$$

Taking logarithms in Eq. (4.5) yields:

$$\log X_t = \log X_0 + \sum_{k=1}^t \log R_k .$$

The process $\tilde{X}_t = \log X_t$ is itself an ordinary binomial process with starting value $\log X_0$ and increments $Z_k = \log R_k$:

$$P(Z_k = \log u) = p, \quad P(Z_k = \log d) = 1 - p .$$

For t large, $\tilde{X}_t$ is approximately normally distributed, i.e. $X_t = \exp(\tilde{X}_t)$ is approximately lognormally distributed.

4.5   Binomial Models with State Dependent Increments

Binomial processes and more general random walks model the stock price at best locally. They proceed from the assumption that the distribution of the increments $Z_t = X_t - X_{t-1}$ is the same for each value X_t, regardless of whether the stock price is substantially greater or smaller than X_0. Absolute increments $X_t - X_{t-1} = (R_t - 1) X_{t-1}$ of a geometric random walk depend on the level of X_{t-1}. Thus, geometric random walks are processes which do not have independent absolute increments. Unfortunately, when modelling the stock price dynamics globally, the latter processes are too simple to explain the impact of the current price level on the future stock price evolution. A class of processes which take this effect into account are binomial processes with state dependent (and possibly time dependent) increments:

$$X_t = X_{t-1} + Z_t \,, \quad t = 1, 2, \dots \tag{4.6}$$

$$\mathrm{P}(Z_t = u) = p(X_{t-1}, t) \,, \;\; \mathrm{P}(Z_t = -d) = 1 - p(X_{t-1}, t) \,.$$

Since the distribution of Z_t depends on the state X_{t-1} and possibly on time t, increments are neither independent nor identically distributed. The deterministic functions $p(x, t)$ associate a probability to each possible value of the process at time t and to each t. Stochastic processes $\{X_t; \; t \geq 0\}$ which are constructed as in (4.6) are still *markovian* but without having independent increments.

Accordingly, geometric binomial processes with state dependent relative increments can be defined (for $u > 1, \; d < 1$):

$$X_t = R_t \cdot X_{t-1} \tag{4.7}$$

$$\mathrm{P}(R_t = u) = p(X_{t-1}, t) \,, \;\; \mathrm{P}(R_t = d) = 1 - p(X_{t-1}, t) \,.$$

Processes as defined in (4.6) and (4.7) are mainly of theoretic interest, since without further assumptions it is rather difficult to estimate the probabilities $p(x, t)$ from observed stock prices. But generalized binomial models (as well as the trinomial models) can be used to solve differential equations numerically, as the Black–Scholes equation for American options for example.

4.6   Recommended Literature

Numerous textbooks deal with stochastic processes in discrete time. We recommend a basic book by Brzezniak and Zastawniak (1999). The classic reference work is, of course, Gikhman and Skorokhod (1974). There are three volumes that are reprinted for today's applications.

4.7   Exercises

Exercise 4.1 *Consider an ordinary random walk $X_t = \sum_{k=1}^{t} Z_k$ for $t = 1, 2, \ldots$, $X_0 = 0$, where $Z_1, Z_2, \ldots$ are i.i.d. with $P(Z_i = 1) = p$ and $P(Z_i = -1) = 1 - p$. Calculate*

(a) $P(X_t > 0)$
(b) $P(X_t > 0)$
(c) $P(Z_2 = 1 | X_3 = 1)$

Exercise 4.2 *Consider an ordinary random walk $X_t = \sum_{k=1}^{t} Z_k$ for $t = 1, 2, \ldots$; $X_0 = 0$, where $Z_1, Z_2, \ldots$ are i.i.d. with $P(Z_i = 1) = p$ and $P(Z_i = -1) = 1 - p$. Let $\tau = \min\{t : |X_t| > 1\}$ be a random variable denoting the first time t when $|X_t| > 1$. Calculate $E[\tau]$.*

Exercise 4.3 *Consider an ordinary random walk $X_t = \sum_{k=1}^{t} Z_k$ for $t = 1, 2, \ldots$, $X_0 = 0$, where $Z_1, Z_2, \ldots$ are i.i.d. with $P(Z_i = 1) = p$ and $P(Z_i = -1) = 1 - p$. Consider also a process $M_t = \max_{s \leq t} X_s$. Calculate $P(X_3 = M_3)$ and $P(M_4 > M_3)$.*

Exercise 4.4 *Let $X_t = \sum_{k=1}^{t} Z_k$ be a general random walk for $t = 1, 2, \ldots$, $X_0 = 0$, and $Z_1, Z_2, \ldots$ are i.i.d. with $\mathsf{Var}(Z_i = 1)$. Calculate $\mathsf{Cor}(X_s, X_t)$.*

Exercise 4.5 *Let $X_t = \sum_{k=1}^{t} Z_k$ be a general random walk for $t = 1, 2, \ldots$, $X_0 = 0$, and $Z_1, Z_2, \ldots$ are i.i.d. and symmetric random variables. Show that*

$$P(\max_{i \leq t} |X_i| > a) \leq 2P(|X_t| > a).$$

Exercise 4.6 *It is a well known fact that the kernel density estimate $\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i)$ is biased. Therefore the comparison of a kernel density estimate with the analytical form of the true e.g. normal, density can be misleading. One has rather to compare the hypothetical density with the expected value $\mathsf{E} \hat{f}_h(x)$ density given as $g(x) = \int_{-\infty}^{+\infty} K_h(x - u) f(u) du$ where $f(u)$ is the true density. Illustrate this fact with a standard normal distribution. Plot the true density f, a kernel density estimate and bias corrected density g.*

Exercise 4.7 *Consider a binomial process $X_t = \sum_{k=1}^{t} Z_k$ for $t = 1, 2, \ldots$, $X_0 = 0$, with state dependent increments. Let $P(Z_t = 1) = p(X_{t-1}) = 1/(2^{|X_{t-1}|+1})$ if $X_{t-1} \geq\!= 0$ and $P(Z_t = 1) = p(X_{t-1}) = 1 - 1/(2^{|X_{t-1}|+1})$ otherwise. To complete the setting $P(Z_t = -1) = 1 - p(X_{t-1})$. Calculate the distribution of X_t for the first five steps.*

Chapter 5
Stochastic Integrals and Differential Equations

This chapter provides the tools needed for option pricing. The field of stochastic processes in continuous time, which are defined as solutions of stochastic differential equations, has an important role to play. To illustrate these notions we use repeated approximations by stochastic processes in discrete time and refer to the results from Chap. 4.

A stochastic process in continuous time $\{X_t; \ t \geq 0\}$ consists of chronologically ordered random variables, but here the variable t is continuous, i.e. t is a positive real number.

Stock prices are actually processes in discrete time. But to derive the Black–Scholes equation they are approximated by continuous time processes which are easier to handle analytically. However the simulation on a computer of such processes or the numerical computation of say American options is carried out by means of discrete time approximations. We therefore switch the time scale between discrete and continuous depending on what is more convenient for the actual computation.

5.1 Wiener Process

We begin with a simple symmetric random walk $\{X_n; \ n \geq 0\}$ starting in 0 $(X_0 = 0)$. The increments $Z_n = X_n - X_{n-1}$ are i.i.d. with:

$$\mathrm{P}(Z_n = 1) = \mathrm{P}(Z_n = -1) = \frac{1}{2}.$$

By shortening the period of time of two successive observations we accelerate the process. Simultaneously, the increments of the process become smaller during the shorter period of time. More precisely, we consider a stochastic process $\{X_t^{\Delta}; \ t \geq 0\}$ in continuous time which increases or decreases in a time step Δt with

probability $\frac{1}{2}$ by Δx. Between these jumps the process is constant (alternatively we could interpolate linearly). At time $t = n \cdot \Delta t$ the process is:

$$X_t^\Delta = \sum_{k=1}^{n} Z_k \cdot \Delta x = X_n \cdot \Delta x$$

where the increments $Z_1 \Delta x, Z_2 \Delta x, \ldots$ are mutually independent and take the values Δx or $-\Delta x$ with probability $\frac{1}{2}$, respectively. From Sect. 4.1 we know:

$$\mathsf{E}[X_t^\Delta] = 0 , \qquad \mathsf{Var}(X_t^\Delta) = (\Delta x)^2 \cdot \mathsf{Var}(X_n) = (\Delta x)^2 \cdot n = t \cdot \frac{(\Delta x)^2}{\Delta t} .$$

Now, we let Δt and Δx become smaller. For the process in the limit to exist in a reasonable sense, $\mathsf{Var}(X_t^\Delta)$ must be finite. On the other hand, $\mathsf{Var}(X_t^\Delta)$ should not converge to 0, since the process would no longer be random. Hence, we must choose:

$$\Delta t \to 0, \ \Delta x = c \cdot \sqrt{\Delta t} , \ \text{ such that } \ \mathsf{Var}(X_t^\Delta) \to c^2 t .$$

If Δt is small, then $n = t/\Delta t$ is large. Thus, the random variable X_n of the ordinary symmetric random walk is approximately $\mathrm{N}(0, n)$ distributed, and therefore for all t (not only for t such that $t = n \ \Delta t$):

$$\mathcal{L}(X_t^\Delta) \approx \mathrm{N}(0, \ n(\Delta x)^2) \approx \mathrm{N}(0, \ c^2 t) .$$

Thus the limiting process $\{X_t; \ t \geq 0\}$ which we obtain from $\{X_t^\Delta; \ t \geq 0\}$ for $\Delta t \to 0, \ \Delta x = c \ \sqrt{\Delta t}$ has the following properties:

(i) X_t is $\mathrm{N}(0, c^2 t)$ distributed for all $t \geq 0$.
(ii) $\{X_t; \ t \geq 0\}$ has *independent increments*, i.e. for $0 \leq s < t$, $X_t - X_s$ is independent of X_s (since the random walk $\{X_n; \ n \geq 0\}$ defining $\{X_t^\Delta; \ t \geq 0\}$ has independent increments).
(iii) For $0 \leq s < t$ the increment $(X_t - X_s)$ is $\mathrm{N}(0, \ c^2 \cdot (t - s))$ distributed, i.e. its distribution depends exclusively on the length $t - s$ of the time interval in which the increment is observed (this follows from (i) and (ii) and the properties of the normal distribution).

A stochastic process $\{X_t; \ t \geq 0\}$ in continuous time satisfying (i)–(iii) is called *Wiener process* or *Brownian motion* starting in 0 ($X_0 = 0$). The standard Wiener process resulting from $c = 1$ will be denoted by $\{W_t; \ t \geq 0\}$. For this process it holds for all $0 \leq s < t$

$$\mathsf{E}[W_t] = 0, \ \mathsf{Var}(W_t) = t$$

$$\text{Cov}(W_t, W_s) = \text{Cov}(W_t - W_s + W_s, W_s)$$
$$= \text{Cov}(W_t - W_s, W_s) + \text{Cov}(W_s, W_s)$$
$$= 0 + \text{Var}(W_s) = s$$

As for every stochastic process in continuous time, we can consider a path or realization of the Wiener process as a *randomly chosen function* of time. With some major mathematical instruments it is possible to show that the paths of a Wiener process are continuous with probability 1:

$$P(W_t \text{ is continuous as a function of } t) = 1.$$

That is to say, the Wiener process has no jumps. But W_t fluctuates heavily: the paths are continuous but highly erratic. In fact, it is possible to show that the paths are not differentiable with probability 1 (Fig. 5.1).

Being a process with independent increments the Wiener process is *markovian*. For $0 \leq s < t$ it holds $W_t = W_s + (W_t - W_s)$, i.e. W_t depends only on W_s and on the increment from time s to time t:

$$P(a < W_t < b \,|W_s = x \,, \text{ information about } W_{t'}, \ 0 \leq t' < s)$$
$$= P(a < W_t < b \,|W_s = x \,)$$

A graphical representation of the above equation is given in Fig. 5.2.
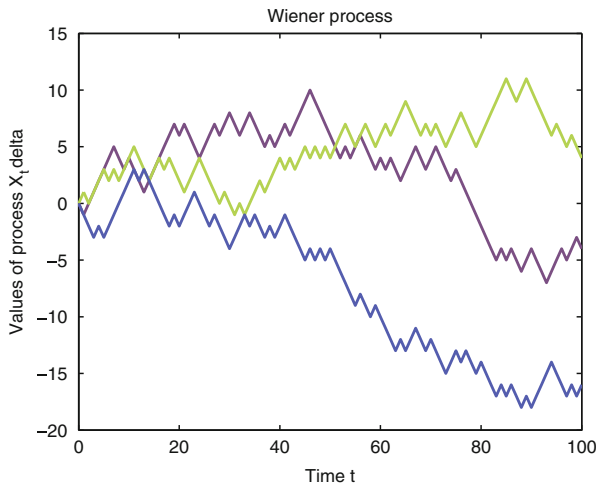
Fig. 5.1 Typical paths of a Wiener process   Q SFEWienerProcess
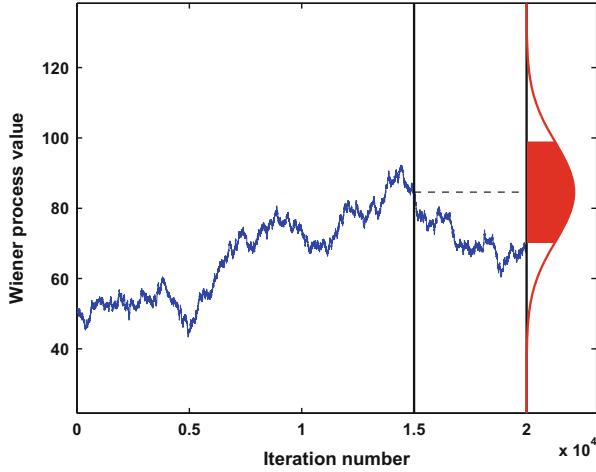
Fig. 5.2 Graphical representation of $P(a < W_t < b \,|W_s = x)$ $\boxed{\text{Q}}$ SFEwienerdens

Using properties (i)–(iii), the distribution of W_t given the outcome $W_s = x$ can be formulated explicitly. Since the increment $(W_t - W_s)$ is $N(0, t - s)$ distributed, W_t is $N(x, t - s)$ distributed given $W_s = x$:

$$P(a < W_t < b \,|W_s = x) = \int_a^b \frac{1}{\sqrt{t - s}} \varphi\Big(\frac{y - x}{\sqrt{t - s}}\Big) dy.$$

Proceeding from the assumption of a random walk $\{X_n; n \geq 0\}$ with drift $E[X_n] = n(2p - 1)$ instead of a symmetric random walk results in a process X_t^Δ which is no more zero on average, but

$$E[X_t^\Delta] = n \cdot (2p - 1) \cdot \Delta x = (2p - 1) \cdot t \, \frac{\Delta x}{\Delta t}$$

$$\mathsf{Var}(X_t^\Delta) = n \cdot 4p(1 - p) \cdot (\Delta x)^2 = 4p(1 - p) \cdot t \cdot \frac{(\Delta x)^2}{\Delta t}.$$

For $\Delta t \to 0$, $\Delta x = \sqrt{\Delta t}$, $p = \frac{1}{2}(1 + \mu\sqrt{\Delta t})$ we obtain for all t:

$$E[X_t^\Delta] \to \mu t \,, \;\; \mathsf{Var}(X_t^\Delta) \to t.$$

The limiting process is a Wiener process $\{X_t; \, t \geq 0\}$ with *drift* or *trend* μt. It results from the standard Wiener process:

$$X_t = \mu t + W_t.$$

Hence, it behaves in the same way as the standard Wiener process but it fluctuates on average around μt instead of 0. If $(\mu > 0)$ the process is increasing linearly on average, and if $(\mu < 0)$ it is decreasing linearly on average.

5.2 Stochastic Integration

In order to introduce a stochastic process as a solution of a stochastic differential equation we introduce the concept of the Itô-integral: a stochastic integral with respect to a Wiener process. Formally the construction of the Itô-integral is similar to the Stieltjes-integral. However, instead of integrating with respect to a deterministic function (Stieltjes-integral), the Itô-integral integrates with respect to a random function, more precisely, the path of a Wiener process. Since the integrant itself can be random, i.e. it can be a path of a stochastic process, one has to analyze the mutual dependencies of the integrant and the Wiener process.

Let $\{Y_t;\ t \geq 0\}$ be the process to integrate and let $\{W_t;\ t \geq 0\}$ be a standard Wiener process. The definition of a stochastic integral assumes that $\{Y_t;\ t \geq 0\}$ is non-anticipating. Intuitively, it means that the process up to time s does not contain any information about future increments $W_t - W_s$, $t > s$, of the Wiener process. In particular, Y_s is independent of $W_t - W_s$.

An integral of a function is usually defined as the limit of the sum of the suitably weighted function. Similarly, the *Itô integral* with respect to a Wiener process is defined as the limit of the sum of the (randomly) weighted (random) function $\{Y_t;\ t \geq 0\}$:

$$I_n = \sum_{k=1}^{n} Y_{(k-1)\Delta t} \cdot (W_{k\Delta t} - W_{(k-1)\Delta t}) , \quad \Delta t = \frac{t}{n}$$

$$\int_0^t Y_s\, dW_s = \lim_{n \to \infty} I_n , \tag{5.1}$$

where the limit is to be understood as the limit of a random variable in terms of mean squared error, i.e. it holds

$$\mathsf{E}\{\left[\int_0^t Y_s\, dW_s - I_n\right]^2\} \to 0, \quad n \to \infty.$$

It is important to note that each summand of I_n is a product of two independent random variables. More precisely, $Y_{(k-1)\Delta t}$, the process to integrate at the left border of the small interval $[(k-1)\Delta t, k\Delta t]$ is independent of the increment $W_{k\Delta t} - W_{(k-1)\Delta t}$ of the Wiener process in this interval.

It is not hard to be more precise on the non-anticipating property of $\{Y_t;\ t \geq 0\}$.

Definition 5.1 (Information Structure, Non-anticipating) For each $t \geq 0$, $\mathcal{F}_t$ denotes a family of events (having the structure of a σ-algebra, i.e. certain combinations of events contained in $\mathcal{F}_t$ are again in $\mathcal{F}_t$) which contain the available information up to time t. $\mathcal{F}_t$ consists of events from which is known up to time t whether they occurred or not. We assume:

$$\mathcal{F}_s \subset \mathcal{F}_t \qquad \text{for } s < t \quad \text{(information grows as time}$$
$$\text{evolves)}$$
$$\{a < Y_t < b\} \in \mathcal{F}_t \qquad \quad (Y_t \text{contains no information}$$
$$\text{about events occurring after}$$
$$\text{time } t)$$
$$\{a < W_t < b\} \in \mathcal{F}_t$$
$$W_t - W_s \text{ independent of } \mathcal{F}_s \quad \text{for } s < t \quad \text{(the Wiener process is adapted to}$$
$$\text{the evolution of information)}$$

Then, we call $\mathcal{F}_t$ the information structure at time t and the process $\{Y_t; \ t \geq 0\}$ non-anticipating with respect to the information structure $\mathcal{F}_t; \ t \geq 0$.

The process $\{Y_t\}$ is called non-anticipating since due to the second assumption it does not anticipate any future information. The evolving information structure $\mathcal{F}_t$ and the random variables Y_t, W_t are adapted to each other.

The integral depends crucially on the point of the interval $[(k-1)\Delta t, \ k\Delta t]$ at which the random variable Y_s is evaluated in (5.1). Consider the example $Y_t = W_t, \ t \geq 0$, i.e. we integrate the Wiener process with respect to itself. As a gedankenexperiment (thought experiment) we replace in (5.1) $(k-1)\Delta t$ by an arbitrary point $t(n,k)$ of the interval $[(k-1)\Delta t, \ k\Delta t]$. If we defined:

$$\int_0^t W_s \, dW_s = \lim_{n \to \infty} \sum_{k=1}^n W_{t(n,k)} (W_{k\Delta t} - W_{(k-1)\Delta t})$$

the expected values would converge as well. Hence by interchanging the sum with the covariance operator we get:

$$\mathsf{E}\left[\sum_{k=1}^n W_{t(n,k)} (W_{k\Delta t} - W_{(k-1)\Delta t})\right] = \sum_{k=1}^n \mathsf{Cov}(W_{t(n,k)}, \ W_{k\Delta t} - W_{(k-1)\Delta t})$$

$$= \sum_{k=1}^n \{t(n,k) - (k-1)\Delta t\} \to \mathsf{E}[\int_0^t W_s \, dW_s].$$

For $t(n,k) = (k-1)\Delta t$—which is the case for the Itô-integral—we obtain 0, for $t(n,k) = k\Delta t$ we obtain $n \cdot \Delta t = t$, and for suitably chosen sequences $t(n,k)$ we could obtain, for the expectation of the stochastic integral, any value between 0 and t. In order to assign a unique value to $\int_0^t W_s \, dW_s$, we have to agree on a certain sequence $t(n,k)$.

To illustrate how Itô-integrals are computed, and also that other than the usual computation rules have to be applied, we show that:

$$\int_0^t W_s \, dW_s = \frac{1}{2}(W_t^2 - W_0^2) - \frac{t}{2} = \frac{1}{2}(W_t^2 - t) \qquad (5.2)$$

Summing the differences $W_{k\Delta t}^2 - W_{(k-1)\Delta t}^2$, all terms but the first and the last cancel out and remembering that $n\Delta t = t$ we get

$$\frac{1}{2}(W_t^2 - W_0^2) = \frac{1}{2} \sum_{k=1}^n (W_{k\Delta t}^2 - W_{(k-1)\Delta t}^2)$$

$$= \frac{1}{2} \sum_{k=1}^n (W_{k\Delta t} - W_{(k-1)\Delta t})(W_{k\Delta t} + W_{(k-1)\Delta t})$$

$$= \frac{1}{2} \sum_{k=1}^n (W_{k\Delta t} - W_{(k-1)\Delta t})^2$$

$$+ \sum_{k=1}^n (W_{k\Delta t} - W_{(k-1)\Delta t}) \, W_{(k-1)\Delta t} \, .$$

While the second term converges to $\int_0^t W_s \, dW_s$, the first term is a sum of n independent identically distributed random variables and which is thus approximated due to the law of large numbers by its expected value

$$\frac{n}{2}\mathsf{E}[(W_{k\Delta t} - W_{(k-1)\Delta t})^2] = \frac{n}{2}\Delta t = \frac{t}{2} \, .$$

For smooth functions f_s, for example continuously differentiable functions, it holds $\int_0^t f_s \, df_s = \frac{1}{2}(f_t^2 - f_0^2)$. However, the stochastic integral (5.2) contains the additional term $-\frac{t}{2}$ since the local increment of the Wiener process over an interval of length Δt is of the size of its standard deviation—that is $\sqrt{\Delta t}$. The increment of a smooth function f_s is proportional to Δt, and therefore considerably smaller than the increment of the Wiener process for $\Delta t \to 0$.

5.3   Stochastic Differential Equations

Since the Wiener process fluctuates around its expectation 0 it can be approximated by means of symmetric random walks. As for random walks we are interested in stochastic processes in continuous time which grow on average, i.e. which have a

trend or *drift*. Proceeding from a Wiener process with arbitrary σ (see Sect. 5.1) we obtain the generalized Wiener process $\{X_t;\ t \geq 0\}$ with *drift rate* μ and variance σ^2:

$$X_t = \mu \cdot t + \sigma \cdot W_t \quad , \quad t \geq 0 . \tag{5.3}$$

The general Wiener process X_t is at time t, $N(\mu t, \sigma^2 t)$-distributed. For its increment in a small time interval Δt we obtain

$$X_{t+\Delta t} - X_t = \mu \cdot \Delta t + \sigma (W_{t+\Delta t} - W_t) .$$

For $\Delta t \to 0$ use the differential notation:

$$dX_t = \mu \cdot dt + \sigma \cdot d W_t \tag{5.4}$$

This is only a different expression for the relationship (5.3) which we can also write in integral form:

$$X_t = \int_0^t \mu \, ds + \int_0^t \sigma \, dW_s \tag{5.5}$$

Note, that from the definition of the stochastic integral it follows directly that $\int_0^t d W_s = W_t - W_0 = W_t$.

The differential notation (5.4) proceeds from the assumption that both the local drift rate given by μ and the local variance given by σ^2 are constant. A considerably larger class of stochastic processes which is more suited to model numerous economic and natural processes is obtained if μ and σ^2 in (5.4) are allowed to be time and state dependent. Such processes $\{X_t;\ t \geq 0\}$, which we call Itô-processes, are defined as solutions of *stochastic differential equations*:

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t \tag{5.6}$$

Intuitively, this means:

$$X_{t+\Delta t} - X_t = \mu(X_t, t)\Delta t + \sigma(X_t, t)(W_{t+\Delta t} - W_t),$$

i.e. the process' increment in a small interval of length Δt after time t is $\mu(X_t, t) \cdot \Delta t$ plus a random fluctuation which is $N(0,\ \sigma^2(X_t, t) \cdot \Delta t)$ distributed. A precise definition of a solution of (5.6) is a stochastic process fulfilling the integral equation

$$X_t - X_0 = \int_0^t \mu(X_s, s)ds + \int_0^t \sigma(X_s, s)dW_s \tag{5.7}$$

In this sense (5.6) is only an abbreviation of (5.7). For $0 \leq t' < t$, it follows immediately:

$$X_t = X_{t'} + \int_{t'}^{t} \mu(X_s, s)ds + \int_{t'}^{t} \sigma(X_s, s)dW_s \, .$$

Since the increment of the Wiener process between t' and t does not dependent on the events which occurred up to time t', it follows that an Itô-process is *Markovian*.

Discrete approximations of (5.6) and (5.7) which can be used to simulate Itô-processes are obtained by observing the process between 0 and t only at evenly spaced points in time $k\Delta t$, $k = 0, \ldots, n$, $n\Delta t = t$.

With $X_k = X_{k\Delta t}$ and $Z_k = (W_{k\Delta t} - W_{(k-1)\Delta t})/\sqrt{\Delta t}$ we get

$$X_{k+1} - X_k = \mu(X_k, k) \cdot \Delta t + \sigma(X_k, k) \cdot Z_{k+1} \cdot \sqrt{\Delta t}$$

or rather with the abbreviations $\mu_k(X) = \mu(X, k)\Delta t$, $\sigma_k(X) = \sigma(X, k)\sqrt{\Delta t}$:

$$X_n - X_0 = \sum_{k=1}^{n} \mu_{k-1}(X_{k-1}) + \sum_{k=1}^{n} \sigma_{k-1}(X_{k-1}) \cdot Z_k$$

with identical independently distributed $N(0, 1)$-random variables $Z_1, Z_2, \ldots$.

Example 5.1 The *Ornstein–Uhlenbeck (OU)* process, a *mean-reverting* process, represents a well-known example of a process with a constant variance $\sigma(X_t, t) = \sigma$ and nonconstant drift $\mu(X_t, t) = \alpha(\mu - X_t)$. The OU process has the following form:

$$dX_t = \alpha(\mu - X_t)dt + \sigma dW_t, \tag{5.8}$$

where α, μ and σ are positive constants. The drift term $\alpha(\mu - X_t)$ in (5.8) is negative when $X_t > \mu$ and it is positive when $X_t < \mu$; so, even if the process will never be free of random fluctuations, we can expect X_t to revert back to μ whenever it has drifted away. In addition, because of the constant local variance σ^2 (not proportional to the level as in the case of geometric Brownian motion), we expect X_t to fluctuate vigorously and to make many crossings of the μ-level. The OU process can be approximated on a discrete grid $t_i = i\Delta t$, $i = 0, 1, \ldots$ by

$$X_{t_i} - X_{t_{i-1}} = \alpha\mu\Delta t - \alpha X_{t_{i-1}}\Delta t + \sigma\sqrt{\Delta t}Z_i$$

where (Z_i) are independent standard normal random variables. Hence, the OU process is the limit of AR(1) processes $X_i = c + \phi X_{i-1} + \varepsilon_i$.

Example 5.2 The *Cox–Ingersoll–Ross (CIR)* process, also referred to as s *square-root diffusion* process, represents an example of a stochastic process where the variance as well as the drift term are not constant. The CIR Process is given by

$$dX_t = \alpha(\mu - X_t)dt + \sigma\sqrt{X_t}dW_t, \tag{5.9}$$
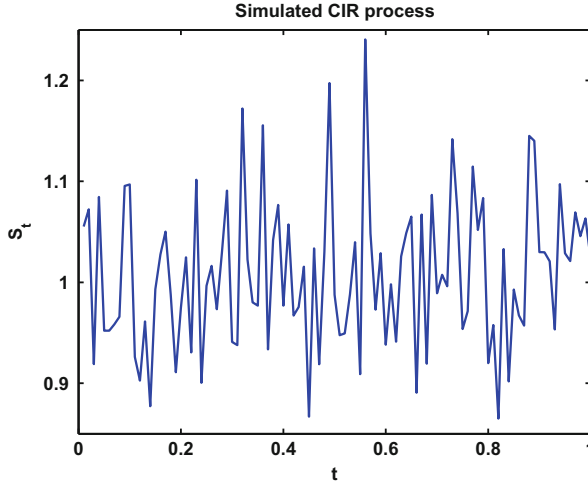
Fig. 5.3 Typical path of a CIR process ($\alpha = 5$, $\mu = 1$, $\sigma = 0.2$)   SFESimCIR

with a drift term $\mu(X_t, t) = \alpha(\mu - X_t)$ and a local variance $\sigma(X_t, t) = \sigma\sqrt{X_t}$; α, μ, σ are positive constants. Figure 5.3 shows a typical path of a CIR process.

5.4   The Stock Price as a Stochastic Process

Stock prices are stochastic processes in *discrete time* which only take *discrete values* due to the limited measurement scale. Nevertheless, stochastic processes in *continuous time* are used as models since they are analytically easier to handle than discrete models, e.g. the binomial or trinomial process. However, the latter is more intuitive and proves to be very useful in simulations.

Two features of the general Wiener process $dX_t = \mu dt + \sigma\, dW_t$ make it an unsuitable model for stock prices. First, it allows for negative stock prices, and second the local variability is higher for high stock prices. Hence, stock prices S_t are modeled by means of the more general Itô-process:

$$dS_t = \mu(S_t, t)dt + \sigma(S_t, t)dW_t \, .$$

This model, however, does depend on the unknown functions $\mu(X, t)$ and $\sigma(X, t)$. A useful and simpler variant utilizing only two unknown real model parameters μ and σ can be justified by the following reflection: The percentage return on the invested capital should on average not depend on the stock price at which the investment is made, and of course, should not depend on the *currency unit* (EUR, USD, . . .) in which the stock price is quoted. Furthermore, the average return should be proportional to the investment horizon, as it is the case for other investment

instruments. Putting things together, we request:

$$\frac{\mathsf{E}[dS_t]}{S_t} = \frac{\mathsf{E}[S_{t+dt} - S_t]}{S_t} = \mu \cdot dt \,.$$

Since $\mathsf{E}[dW_t] = 0$ this condition is satisfied if

$$\mu(S_t, t) = \mu \cdot S_t \,,$$

for given S_t. Additionally,

$$\sigma(S_t, t) = \sigma \cdot S_t$$

takes into consideration that the absolute size of the stock price fluctuation is proportional to the currency unit in which the stock price is quoted. In summary, we model the stock price S_t as a solution of the stochastic differential equation

$$d S_t = \mu \cdot S_t \, dt + \sigma \cdot S_t \cdot dW_t \,,$$

where μ is the *expected return* on the stock, and σ the *volatility*. Such a process is called *geometric Brownian motion* because

$$\frac{dS_t}{S_t} = \mu \, dt + \sigma \, dW_t \,.$$

We can interpret this equation as follows. For small time Δt the change in the stock can be approximated by

$$S_{t+\Delta t} - S_t = \mu S_t \Delta t + \sigma S_t (W_{t+\Delta t} - W_t),$$

where $W_{t+\Delta t} - W_t$ represents the change of the Brownian motion over the time interval Δt. As Δt becomes smaller the approximation becomes more accurate. A typical path of a geometric Brownian motion is plotted in Fig. 5.4. The rough nature of the path corresponds to the fact that no paths of a (geometric) Brownian motion are differentiable anywhere. We have chosen as parameters a mean return of 5 % and a volatility of 20 % which can be regarded as realistic values for stock price processes.

By applying Itôs lemma, which we introduce in Sect. 5.5, it can be shown that for a suitable Wiener process $\{Y_t; \ t \geq 0\}$ it holds

$$S_t = e^{Y_t} \quad \text{or} \quad Y_t = \log S_t \,.$$

Since Y_t is normally distributed, S_t is lognormally distributed. As random walks can be used to approximate the general Wiener process, geometric random walks can be used to approximate geometric Brownian motion and thus this simple model for the stock price (Fig. 5.5).
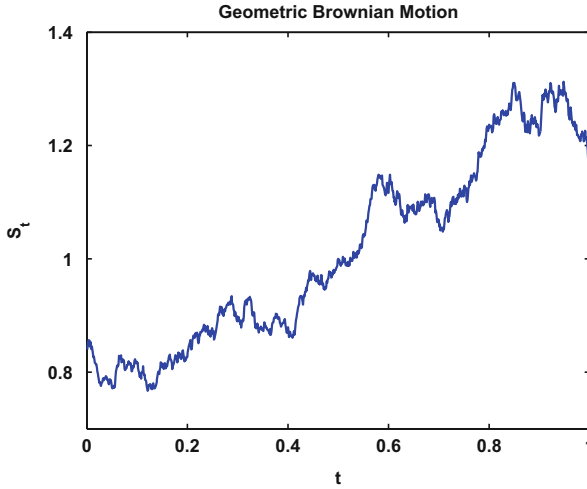
Fig. 5.4 Typical path of a geometric Brownian motion ($\mu = 0.05$, $\sigma = 0.2$)

🔍 SFEGBMProcess
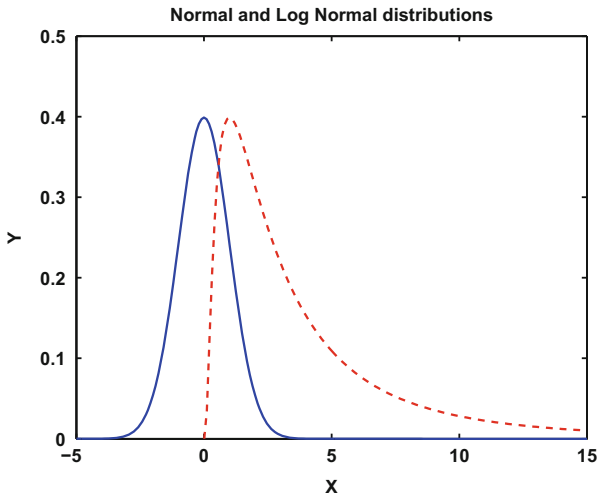
Fig. 5.5 Density comparison of lognormally and normally distributed random variables

🔍 SFELogNormal

5.5 Itô's Lemma

A crucial tool in dealing with stochastic differential equations is Itô's lemma. If
$\{X_t, \ t \geq 0\}$ is an Itô-process:

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)\, dW_t\,, \tag{5.10}$$

one is often interested in the dynamics of stochastic processes which are functions of X_t: $Y_t = g(X_t)$. Then $\{Y_t;\ t \geq 0\}$ can also be described by a solution of a stochastic differential equation from which interesting properties of Y_t can be derived, as for example the average growth in time t.

For a heuristic derivation of the equation for $\{Y_t;\ t \geq 0\}$ we assume that g is differentiable as many times as necessary. From a Taylor expansion it follows that:

$$
\begin{aligned}
Y_{t+dt} - Y_t &= g(X_{t+dt}) - g(X_t) \\
&= g(X_t + dX_t) - g(X_t) \\
&= \frac{dg}{dX}(X_t) \cdot dX_t + \frac{1}{2}\frac{d^2g}{dX^2}(X_t) \cdot (dX_t)^2 + \ldots
\end{aligned}
\tag{5.11}
$$

where the dots indicate the terms which can be neglected (for $dt \to 0$). Due to Eq. (5.10) the drift term $\mu(X_t, t)dt$ and the volatility term $\sigma(X_t, t)dW_t$ are the dominant terms since for $dt \to 0$ they are of size dt and $\sqrt{dt}$, respectively.

In doing this, we use the fact that $\mathsf{E}[(dW_t)^2] = dt$ and $dW_t = W_{t+dt} - W_t$ is of the size of its standard deviation, $\sqrt{dt}$. We neglect terms which are of a smaller size than dt. Thus, we can express $(dX_t)^2$ by a simpler term:

$$
\begin{aligned}
(dX_t)^2 &= (\mu(X_t, t)dt + \sigma(X_t, t)\,dW_t)^2 \\
&= \mu^2(X_t, t)(dt)^2 + 2\mu(X_t, t)\,\sigma(X_t, t)\,dt\,dW_t + \sigma^2(X_t, t)(dW_t)^2\,.
\end{aligned}
$$

We see that the first and the second terms are of size $(dt)^2$ and $dt \cdot \sqrt{dt}$, respectively. Therefore, both can be neglected. However, the third term is of size dt. More precisely, it can be shown that $dt \to 0$:

$$
(dW_t)^2 = dt.
$$

Thanks to this identity, calculus rules for stochastic integrals can be derived from the rules for deterministic functions (as Taylor expansions for example). Neglecting terms which are of smaller size than dt we obtain the following version of *Itôs lemma* from (5.11):

Lemma 5.1 (Itô's Lemma)

$$
\begin{aligned}
dY_t &= dg(X_t) \\
&= \left(\frac{dg}{dX}(X_t) \cdot \mu(X_t, t) + \frac{1}{2}\frac{d^2g}{dX^2}(X_t) \cdot \sigma^2(X_t, t) \right) dt \\
&\quad + \frac{dg}{dX}(X_t) \cdot \sigma(X_t, t)\,dW_t
\end{aligned}
$$

or—dropping the time index t and the argument X_t of the function g and its derivatives:

$$dg = \left(\frac{dg}{dX}\mu(X,t) + \frac{1}{2}\frac{d^2g}{dX^2}\sigma^2(X,t) \right)dt + \frac{dg}{dX}\sigma(X,t)dW_t .$$

Example 5.3 Consider $Y_t = \log S_t$ the logarithm of the geometric Brownian motion. For $g(X) = \log X$ we obtain $\frac{dg}{dX} = \frac{1}{X}$, $\frac{d^2g}{dX^2} = -\frac{1}{X^2}$. Applying Itô's lemma for the geometric Brownian motion with $\mu(X,t) = \mu X$, $\sigma(X,t) = \sigma X$ we get:

$$dY_t = \left(\frac{1}{S_t}\mu S_t - \frac{1}{2}\frac{1}{S_t^2}\sigma^2 S_t^2 \right)dt + \frac{1}{S_t}\cdot\sigma S_t dW_t$$

$$= (\mu - \frac{1}{2}\sigma^2)dt + \sigma dW_t$$

The *logarithm of the stock price* is a generalized Wiener process with drift rate $\mu^* = \mu - \frac{1}{2}\sigma^2$ and variance rate σ^2. Since Y_t is $N(\mu^*t, \sigma^2t)$-distributed S_t is itself *lognormally distributed* with parameters μ^*t and σ^2t .

A generalized version of Itô's lemma for functions $g(X,t)$ which are allowed to depend on time t is:

Lemma 5.2 (Itô's Lemma for Functions Depending Explicitly on Time)

$$dg = \left(\frac{\partial g}{\partial X}\cdot\mu(X,t) + \frac{1}{2}\frac{\partial^2 g}{\partial X^2}\sigma^2(X,t) + \frac{\partial g}{\partial t} \right)dt + \frac{\partial g}{\partial X}\sigma(X,t)dW_t \qquad (5.12)$$

$Y_t = g(X_t,t)$ is again an Itô process, but this time the drift rate is augmented by an additional term $\frac{\partial g}{\partial t}(X_t,t)$.

Example 5.4 Consider a forward contract on a non-dividend paying stock. Let S_0 denotes spot price at time 0 and T is the time to maturity of the forward contract. We assume that the risk-free interest rate r is constant for all maturities. The forward price at time 0 is then given by

$$F_0 = S_0e^{rT} \qquad (5.13)$$

and at a general time $t < T$ by

$$F_t = S_te^{r(T-t)}. \qquad (5.14)$$

Assuming that the process $\{S_t, t \geq 0\}$ is given by (5.10), we use Itô's lemma to determine the process for F_t. For $g(X,t) = Xe^{r(T-t)}$ we obtain:

$$\frac{dg}{dX} = e^{r(T-t)} \qquad \frac{d^2g}{dX^2} = 0 \qquad \frac{dg}{dt} = -rXe^{r(T-t)}.$$

Applying (5.12) to the geometric Brownian motion with $\mu(X, t) = \mu X$, $\sigma(X, t) = \sigma X$ we get

$$dF_t = \left\{ e^{r(T-t)} \mu S_t - r S_t e^{r(T-t)} \right\} dt + e^{r(T-t)} \sigma S_t dW_t.$$

Substituting $F_t = S_t e^{r(T-t)}$, this becomes

$$dF_t = (\mu - r) F_t dt + \sigma F_t dW_t.$$

Like S_t, the process $\{F_t, t \geq 0\}$ follows a geometric Brownian motion, however with an expected growth rate of $\mu - r$ rather than μ.

5.6   Recommended Literature

This chapter briefly summarizes results which belong to the main topics of stochastic analysis. Numerous textbooks, of diverse levels, introduce the calculus of stochastic integrals and stochastic differential equations, see for example, von Weizsäcker and Winkler (1990), Mikosch (1998) or Karatzas and Shreve (1999).

5.7   Exercises

Exercise 5.1 *Let W_t be a standard Wiener process. Show that the following processes are also standard Wiener processes:*

(a) $X_t = c^{-\frac{1}{2}} W_{ct}$ *for* $c > 0$
(b) $Y_t = W_{T+t} - W_T$ *for* $T > 0$
(c) $V_t = \begin{cases} W_t & \text{if } t \leq T \\ 2W_T - W_t & \text{if } t > T \end{cases}$
(d) $Z_t = t W_{\frac{1}{t}}$ *for* $t > 0$ *and* $Z_t = 0$.

Exercise 5.2 *Calculate* $\mathsf{Cov}(W_t, 3W_s - 4W_t)$ *and* $\mathsf{Cov}(W_s, 3W_s - 4W_t)$ *for* $0 \leq s \leq t$.

Exercise 5.3 *Let W_t be a standard Wiener process. The process $U_t = W_t - t W_1$ for $t \in [0, 1]$ is called Brownian bridge. Calculate its covariance function. What is the distribution of U_t?*

Exercise 5.4 *Calculate* $\mathsf{E}\left(\int_0^{2\pi} W_s dW_s \right)$

Exercise 5.5 *Consider the process $dS_t = \mu dt + \sigma dW_t$. Find the dynamics of the process $Y_t = g(S_t)$, where $g(S_t, t) = 2 + t + e^{S_t}$.*

Exercise 5.6 *Derive $\int_0^t W_s^2 dW_s$, where W_t is a Brownian motion.*

Exercise 5.7 *Let W_t be a standard Wiener process. Compute then $E[W_t^4]$*

Exercise 5.8 ***Product Rule/Partial Integration*** *If X_t and Y_t are two one-dimensional Itô processes with:*

$$X_t = X_0 + \int_0^t K_s ds + \int_0^t H_s dW_s$$

$$Y_t = Y_0 + \int_0^t \mu_s ds + \int_0^t \sigma_s dW_s$$

Then we have:

$$X_t \cdot Y_t = X_0 \cdot Y_0 + \int_0^t X_s dY_s + \int_0^t Y_s dX_s + \int_0^t d <X, Y>_s$$

$$= X_0 \cdot Y_0 + \int_0^t (X_s \mu_s + Y_s K_s + H_s \sigma_s) ds + \int_s^t (H_s \sigma_s + Y_s H_s) dW_s$$

Exercise 5.9 ***The Ornstein–Uhlenbeck Process*** *Prove that the following process*

$$S_t = e^{-\mu t} S_0 + e^{-\mu t} \int_0^t e^{\mu s} \sigma dW_s$$

is a solution of the Ornstein–Uhlenbeck SDE.

$$dS_t = -\mu S_t dt + \sigma dW_t$$

μ, σ are constant and W_t is a Standard Wiener process.

Exercise 5.10 *Apply the result from Exercise 5.9 for solving the following SDE:*

$$dX_t = a X_t (\theta - \log X_t) dt + \sigma X_t dW_t$$

Chapter 6
Black–Scholes Option Pricing Model

6.1 Black–Scholes Differential Equation

Simple, generally accepted economic assumptions are insufficient to develop a rational option pricing theory. Assuming a perfect financial market (Sect. 2.1) leads to elementary arbitrage relations which the options have to fulfil. While these relations can be used as a verification tool for sophisticated mathematical models, they do not provide an explicit option pricing function depending on parameters such as time, stock price and the options underlying parameters K, T. To obtain such a pricing function the value of the underlying financial instrument (stock, currency, ...) has to be modelled. In general, the underlying instrument is assumed to follow a stochastic process either in discrete or in continuous time. While the latter is analytically easier to handle, the former, which we will consider as an approximation of a continuous time process for the time being, is particularly useful for numerical computations. In the second part of this text, the discrete time version will be discussed as a financial time series model.

A model for stock prices which is frequently used and is also the basis of the classical Black–Scholes approach, is the so-called geometric Brownian motion. In this model the stock price S_t is a solution of the stochastic differential equation

$$dS_t = \mu S_t dt + \sigma S_t dW_t. \tag{6.1}$$

Equivalently, the process of stock price returns can be assumed to follow a standard Brownian motion, i.e.

$$\frac{dS_t}{S_t} = \mu dt + \sigma dW_t. \tag{6.2}$$

The drift μ is the expected return on the stock in the time interval dt. The volatility σ is a measure of the return variability around its expectation μ. Both parameters μ

and σ are dependent on each other and are important factors in the investors' risk preferences involved in the investment decision: The higher the expected return μ, the higher, in general, the risk quantified by σ.

Modelling the underlying as geometric Brownian motion provides a useful approximation to stock prices accepted by practitioners for short and medium maturity. In practice, numerous model departures are known: in some situations the volatility function $\sigma(x, t)$ of the general model (5.10) is different from the linear specification $\sigma \cdot x$ of geometric Brownian motion. The Black–Scholes' approach is still used to approximate option prices, and its basic idea to derive option prices can be applied to more general stock price models.

The Black–Scholes' approach relies on the idea introduced in Chap. 2, i.e. duplicating the portfolio consisting of the option by means of a second portfolio consisting exclusively of financial instruments whose values are known. The duplicating portfolio is chosen in such a way that both portfolios have equal values at the option's maturity T. It then follows from the assumption of a perfect financial market, and in particular of the lack of arbitrage opportunities, that both portfolios must have equal values at any time prior to time T. The duplicating portfolio can be created in two equivalent ways which we illustrate with an example of a call option on a stock with price S_t:

1. Consider a portfolio consisting of one call of which the price is to be computed. The duplicating portfolio is composed of stocks and risk-less zero bonds of which the quantity adapts continuously to changes in the stock price. Without loss of generality, the zero bond's nominal value can be set equal to one since the number of zero bonds in the duplicating portfolio can be chosen arbitrarily. At time t the two portfolios consist of:

   Portfolio A: One call option (long position) with delivery price K and maturity date T.

   Portfolio B: $n_t = n(S_t, t)$ stocks and $m_t = m(S_t, t)$ zero bonds with nominal value $B_T = 1$ and maturity date T.

2. Consider a perfect hedge-portfolio, which consists of stocks and written calls (by means of short selling). Due to a dynamic hedge-strategy the portfolio bears no risk at any time, i.e. profits from calls are neutralized by losses from stocks. Correspondingly, the duplicating portfolio is also risk free and consists exclusively of zero bonds. Again, the positions are adjusted continuously to changes in the stock price. At time t the two portfolios are composed of:

   Portfolio A: One stock and $n_t = n(S_t, t)$ (by means of short selling) written call options on the stock with delivery price K and maturity date T.

   Portfolio B: $m_t = m(S_t, t)$ zero bonds with face value $B_T = 1$ and maturity date T.

Let $T^* = T$ be the time when the call option expires as worthless, and otherwise let T^* be the time at which the option is exercised. While for a European call option it holds $T^* = T$ at any time, an American option can be exercised prior to maturity.

We will see that both in 1. the call value is equal to the value of the duplicating portfolio, and in 2. the hedge-portfolio's value equals the value of the risk free zero bond portfolio at any time $t \leq T^*$, and thus the same partial differential equation for the call value results; this is called *Black–Scholes equation*.

The Black–Scholes approach can be applied to any financial instrument $\mathcal{U}$ contingent on an underlying with price S_t if the latter price follows a geometric Brownian motion, and if the derivatives price F_t is a function only of the price S_t and time: $F_t = F(S_t, t)$. Then, according to the theorem below, a portfolio duplicating the financial instrument exists, and the approach illustrated in 1. can be applied to price the instrument. Pricing an arbitrary derivative the duplicating portfolio must not only have the same value as the derivative at exercising time T^*, but also the same cash flow pattern, i.e. the duplicating portfolio has to generate equal amounts of withdrawal profits or contributing costs as the derivative. The existence of a perfect hedge-portfolio of approach 2. can be shown analogously.

Theorem 6.1 *Let the value S_t of an object be a geometric Brownian motion (6.1). Let $\mathcal{U}$ be a derivative contingent on the object and maturing in T. Let $T^* \leq T$ be the time at which the derivative is exercised, or $T^* = T$ if it is not. Let the derivative's value at any time $t \leq T^*$ be given by a function $F(S_t, t)$ of the object's price and time.*

a) *A portfolio exists consisting of the underlying object and risk free bonds which duplicates the derivative in the sense that it generates up to time T^* the same cash flow pattern as $\mathcal{U}$, and that it has the same value as $\mathcal{U}$ at time T^*.*
b) *The derivative's value function $F(S, t)$ satisfies Black–Scholes partial differential equation*

$$\frac{\partial F(S, t)}{\partial t} - rF(S, t) + bS\frac{\partial F(S, t)}{\partial S} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 F(S, t)}{\partial S^2} = 0, \quad t \leq T^*. \quad (6.3)$$

Proof To simplify we proceed from the assumption that the object is a stock paying a continuous dividend yield d, and thus involving costs of carry $b = r - d$ with r the continuous compounded risk free interest rate. Furthermore, we consider only the case where $\mathcal{U}$ is a derivative on the stock, and that $\mathcal{U}$ does not generate any payoff before time T^*.

We construct a portfolio consisting of $n_t = n(S_t, t)$ shares of the stock and $m_t = m(S_t, t)$ zero bonds with maturity date T and a face value of $B_T = 1$. Let

$$B_t = B_T e^{-r(T-t)} = e^{-r(t-T)}$$

be the zero bond's value discounted to time t. We denote the time t portfolio value by

$$V_t \stackrel{\text{def}}{=} V(S_t, t) = n(S_t, t) \cdot S_t + m(S_t, t) \cdot B_t.$$

It can be shown that n_t and m_t can be chosen so that at exercise time and respectively at maturity of $\mathcal{U}$, the portfolio value is equal to the derivative's value, i.e. $V(S_{T^*}, T^*) = F(S_{T^*}, T^*)$. Furthermore, it can be shown that the portfolio does not generate any cash flow prior to T^*, i.e. it is neither allowed to withdraw nor to add any money before time T^*. All changes in the positions must be realized by buying or selling stocks or bonds, or by means of dividend yields.

Firstly, we investigate how the portfolio value V_t changes in a small period of time dt. By doing this, we use the notation $dV_t = V_{t+dt} - V_t$, $dn_t = n_{t+dt} - n_t$ etc.

$$dV_t = n_{t+dt} S_{t+dt} + m_{t+dt} B_{t+dt} - n_t S_t - m_t B_t$$
$$= dn_t S_{t+dt} + n_t dS_t + dm_t B_{t+dt} + m_t dB_t,$$

and thus

$$dV_t = dn_t (S_t + dS_t) + n_t dS_t + dm_t (B_t + dB_t) + m_t dB_t. \tag{6.4}$$

Since the stochastic process S_t is a geometric Brownian motion and therefore an Itô-process (5.10) with $\mu(x, t) = \mu x$ and $\sigma(x, t) = \sigma x$, it follows from the generalized Itô lemma (5.12) and Eq. (6.1)

$$dn_t = \frac{\partial n_t}{\partial t} dt + \frac{\partial n_t}{\partial S} dS_t + \frac{1}{2} \frac{\partial^2 n_t}{\partial S^2} \sigma^2 S_t^2 dt, \tag{6.5}$$

and an analogous relation for m_t. Using

$$(dS_t)^2 = (\mu S_t dt + \sigma S_t dW_t)^2 = \sigma^2 S_t^2 (dW_t)^2 + \mathcal{O}(dt) = \sigma^2 S_t^2 dt + \mathcal{O}(dt),$$

$$dB_t = rB_t dt, \ dS_t \cdot dt = \mathcal{O}(dt) \text{ and } dt^2 = \mathcal{O}(dt)$$

and neglecting terms of size smaller than dt it follows:

$$dn_t (S_t + dS_t) = \left(\frac{\partial n_t}{\partial t} dt + \frac{\partial n_t}{\partial S} dS_t + \frac{1}{2} \frac{\partial^2 n_t}{\partial S^2} \sigma^2 S_t^2 dt \right) S_t + \frac{\partial n_t}{\partial S} \sigma^2 S_t^2 dt, \tag{6.6}$$

$$dm_t (B_t + dB_t) = \left(\frac{\partial m_t}{\partial t} dt + \frac{\partial m_t}{\partial S} dS_t + \frac{1}{2} \frac{\partial^2 m_t}{\partial S^2} \sigma^2 S_t^2 dt \right) B_t. \tag{6.7}$$

The fact that neither the derivative nor the duplicating portfolio generates any cash flow before time T^* means that the terms $dn_t (S_t + dS_t)$ and $dm_t (B_t + dB_t)$ of dV_t in Eq. (6.4) which correspond to purchases and sales of stocks and bonds have to be financed by the dividend yields. Since one share of the stock pays in a small time interval dt a dividend amount of $d \cdot S_t \cdot dt$, we have

$$d \cdot n_t S_t \cdot dt = (r - b) \cdot n_t S_t \cdot dt = dn_t (S_t + dS_t) + dm_t (B_t + dB_t).$$

Substituting Eqs. (6.6) and (6.7) in the latter equation, it holds:

$$0 = (b - r)n_t S_t dt + \left(\frac{\partial m_t}{\partial t} dt + \frac{\partial m_t}{\partial S} dS_t + \frac{1}{2} \frac{\partial^2 m_t}{\partial S^2} \sigma^2 S_t^2 dt \right) B_t$$

$$+ \left(\frac{\partial n_t}{\partial t} dt + \frac{\partial n_t}{\partial S} dS_t + \frac{1}{2} \frac{\partial^2 n_t}{\partial S^2} \sigma^2 S_t^2 dt \right) S_t + \frac{\partial n_t}{\partial S} \sigma^2 S_t^2 dt.$$

Using Eq. (6.1) and summarizing the stochastic terms with differential dW_t as well as the deterministic terms with differential dt containing the drift parameter μ, and all other deterministic terms gives:

$$0 = \left(\frac{\partial n_t}{\partial S} S_t + \frac{\partial m_t}{\partial S} B_t \right) \mu S_t dt$$

$$+ \left\{ \left(\frac{\partial n_t}{\partial t} + \frac{1}{2} \frac{\partial^2 n_t}{\partial S^2} \sigma^2 S_t^2 \right) S_t + \frac{\partial n_t}{\partial S} \sigma^2 S_t^2 \right.$$

$$\left. + \left(\frac{\partial m_t}{\partial t} + \frac{1}{2} \frac{\partial^2 m_t}{\partial S^2} \sigma^2 S_t^2 \right) B_t + (b - r)n_t S_t \right\} dt$$

$$+ \left(\frac{\partial n_t}{\partial S} S_t + \frac{\partial m_t}{\partial S} B_t \right) \sigma S_t dW_t. \tag{6.8}$$

This is only possible if the stochastic terms disappear, i.e.

$$\frac{\partial n_t}{\partial S} S_t + \frac{\partial m_t}{\partial S} B_t = 0. \tag{6.9}$$

Thus the first term in (6.8) is neutralized as well. Hence the middle term must also be zero:

$$\left(\frac{\partial n_t}{\partial t} + \frac{1}{2} \frac{\partial^2 n_t}{\partial S^2} \sigma^2 S_t^2 \right) S_t + \frac{\partial n_t}{\partial S} \sigma^2 S_t^2$$

$$+ \left(\frac{\partial m_t}{\partial t} + \frac{1}{2} \frac{\partial^2 m_t}{\partial S^2} \sigma^2 S_t^2 \right) B_t + (b - r)n_t S_t = 0. \tag{6.10}$$

To further simplify we compute the partial derivative of Eq. (6.9) with respect to S:

$$\frac{\partial^2 n_t}{\partial S^2} S_t + \frac{\partial n_t}{\partial S} + \frac{\partial^2 m_t}{\partial S^2} B_t = 0 \tag{6.11}$$

and substitute this in Eq. (6.10). We then obtain

$$\frac{\partial n_t}{\partial t} S_t + \frac{\partial m_t}{\partial t} B_t + \frac{1}{2} \frac{\partial n_t}{\partial S} \sigma^2 S_t^2 + (b - r)n_t S_t = 0. \tag{6.12}$$

Since the stock price S_t does not depend explicitly on time, i.e. $\partial S_t / \partial t = 0$, the derivative of the portfolio value $V_t = n_t S_t + m_t B_t$ with respect to time gives:

$$\frac{\partial V_t}{\partial t} = \frac{\partial n_t}{\partial t} S_t + \frac{\partial m_t}{\partial t} B_t + m_t \frac{\partial B_t}{\partial t} = \frac{\partial n_t}{\partial t} S_t + \frac{\partial m_t}{\partial t} B_t + m_t r B_t.$$

This implies

$$\frac{\partial n_t}{\partial t} S_t + \frac{\partial m_t}{\partial t} B_t = \frac{\partial V_t}{\partial t} - r m_t B_t = \frac{\partial V_t}{\partial t} - r(V_t - n_t S_t).$$

Substituting this equation in Eq. (6.12) we eliminate m_t and obtain

$$\frac{1}{2} \sigma^2 S_t^2 \frac{\partial n}{\partial S} + \frac{\partial V_t}{\partial t} + b n_t S_t - r V_t = 0. \tag{6.13}$$

Since the zero bond value B_t is independent of the stock price S_t, i.e. $\partial B_t / \partial S = 0$, the derivative of the portfolio value $V_t = n_t S_t + m_t B_t$ with respect to the stock price gives (using Eq. (6.9))

$$\frac{\partial V_t}{\partial S} = \frac{\partial n_t}{\partial S} S_t + n_t + \frac{\partial m_t}{\partial S} B_t = n_t,$$

and thus

$$n_t = \frac{\partial V_t}{\partial S}. \tag{6.14}$$

That is, n_t is equal to the so-called *delta* or hedge-ratio of the portfolio (see Sect. 6.4.1). Since

$$m_t = \frac{V_t - n_t S_t}{B_t}$$

we can construct a duplicating portfolio if we know $V_t = V(S_t, t)$. We can obtain this function of stock price and time as a solution of the Black–Scholes differential equation

$$\frac{\partial V(S,t)}{\partial t} - r V(S,t) + b S \frac{\partial V(S,t)}{\partial S} + \frac{1}{2} \sigma^2 S^2 \frac{\partial^2 V(S,t)}{\partial S^2} = 0 \tag{6.15}$$

which results from substituting Eq. (6.14) in Eq. (6.13). To determine V we have to take into account a boundary condition which is obtained from the fact that the cash flows at exercising time, respective maturity, i.e. at time T^*, of the duplicating portfolio and the derivative are equal:

$$V(S_{T^*}, T^*) = F(S_{T^*}, T^*). \tag{6.16}$$

Since the derivative has, at any time, the same cash flow as the duplicating portfolio, $F(S,t)$ also satisfies the Black–Scholes differential equation, and at any time $t \leq T^*$ it holds $F_t = F(S_t, t) = V(S_t, t) = V_t$.   □

The Black–Scholes' differential equation fundamentally relies on the assumption that the stock price can be modelled by a geometric Brownian motion. This assumption is only justified, if the theory building on it reproduces the arbitrage relations derived in Chap. 2. For a particular example we verify this feature. Let $V(S_t, t)$ be the value of a future contract with delivery price K and maturity date T. The underlying object involves costs of carry at a continuous rate b. Since $V(S_t, t)$ depends only on the price of the underlying and time, it satisfies the conditions of Theorem 6.1. From Theorem 2.1 and substituting $\tau = T - t$ for the time to maturity it follows

$$V(S,t) = Se^{(r-b)(t-T)} - Ke^{r(t-T)}.$$

Substituting the above equation into Eq. (6.3) it can be easily seen that it is the unique solution of Black–Scholes' differential equation with boundary condition $V(S, T) = S - K$. Hence, Black–Scholes' approach gives the same price for the future contract as the model free noarbitrage approach.

Finally, we point out that modelling stock prices by geometric Brownian motion gives reasonable solutions for short and medium time spans. Applying the model to other underlyings such as currencies or bonds is more difficult. Bond options typically have significantly longer times to maturity than stock options. Their value does not only depend on the bond price but also on interest rates which have to be considered stochastic. Modelling interest rates reasonably involves other stochastic process, which we will discuss in latter chapters.

Generally, exchange rates cannot be modelled by a geometric Brownian motion. Empirical studies show that the performance of this model depends on the currency and on the time to maturity. Hence, applying the Black–Scholes' approach to currency options has to be verified in each case. If the model is used, the foreign currency has to be understood as the option underlying with a continuous foreign interest rate d corresponding to the continuous dividend yield of a stock. Thus, continuous costs of carry with rate $b = r - d$ equal the interest rate differential between the domestic and the foreign market. If the investor buys the foreign currency early, then he cannot invest his capital at home any more, and thus he loses the domestic interest rate r. However, he can invest his capital abroad and gain the foreign interest rate d. The value of the currency option results from solving Black–Scholes' differential equation (6.3) respecting the boundary condition implied by the option type.

6.2   Black–Scholes Formula for European Options

In this section we are going to use the Black–Scholes equation to compute the price of European options. We keep the notation introduced in the previous chapter. That is, we denote

$$C(S,t) = C_{K,T}(S,t), \qquad\qquad P(S,t) = P_{K,T}(S,t)$$

the value of a European call respectively put option with exercise price K and maturity date T at time $t \leq T$, where the underlying, for example a stock, at time t has a value of $S_t = S$. The value of a call option thus satisfies for all prices S with $0 < S < \infty$ the differential equation

$$rC(S,t) - bS\frac{\partial C(S,t)}{\partial S} - \frac{1}{2}\sigma^2 S^2\frac{\partial^2 C(S,t)}{\partial S^2} = \frac{\partial C(S,t)}{\partial t}, \; 0 \leq t \leq T, \; (6.17)$$

$$C(S,T) = \max\{0, S - K\}, \; 0 < S < \infty, \; (6.18)$$

$$C(0,t) = 0, \quad \lim_{S \to \infty} C(S,t) - S = 0, \; 0 \leq t \leq T. \; (6.19)$$

The first boundary condition (6.18) follows directly from the definition of a call option, which will only be exercised if $S_T > K$ thereby procuring the gain $S_T - K$. The definition of geometric Brownian motion implies that the process is absorbed by zero. In other words, if $S_t = 0$ for one $t < T$ it follows $S_T = 0$. That is the call will not be exercised, which is formulated in the first part of condition (6.19). Whereas the second part of (6.19) results from the reflection that the probability of the Brownian motion falling below K is fairly small once it has attained a level significantly above the exercise price. If $S_t \gg K$ for a $t < T$ then it holds with a high probability that $S_T \gg K$. The call will be, thus, exercised and yields the cash flow $S_T - K \approx S_T$.

The differential equation (6.17) subject to boundary conditions (6.18), (6.19) can be solved analytically. To achieve this, we transform it into a differential equation known from the literature. Firstly, we substitute the time variable t by the time to maturity $\tau = T - t$. By doing this, the problem with final condition (6.18) in $t = T$ changes to a problem subject to an initial condition in $\tau = 0$. Subsequently, we multiply (6.17) by $2/\sigma^2$ and replace the parameters r, b by

$$\alpha = \frac{2r}{\sigma^2}, \; \beta = \frac{2b}{\sigma^2},$$

as well as the variables τ, S by

$$v = \sigma^2(\beta - 1)^2\frac{\tau}{2}, \qquad\qquad u = (\beta - 1)\log\frac{S}{K} + v.$$

While for the original parameters $0 \le S < \infty, 0 \le t \le T$, for the new parameters
it holds that:

$$-\infty < u < \infty, \qquad\qquad 0 \le v \le \frac{1}{2}\sigma^2(\beta - 1)^2 T \stackrel{\text{def}}{=} v_T.$$

Finally, we set

$$g(u, v) = e^{r\tau} C(S, T - \tau)$$

and obtain the new differential equation

$$\frac{\partial^2 g(u, v)}{\partial u^2} = \frac{\partial g(u, v)}{\partial v}. \tag{6.20}$$

with the initial condition

$$g(u, 0) = K \max\{0, e^{\frac{u}{\beta-1}} - 1\} \stackrel{\text{def}}{=} g_0(u), \quad -\infty < u < \infty. \tag{6.21}$$

Problems with initial conditions of this kind are well known from the literature
on partial differential equations. They appear, for example, in modelling heat
conduction and diffusion processes. The solution is given by

$$g(u, v) = \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi v}} g_0(\xi) e^{-\frac{(\xi-u)^2}{4v}} d\xi.$$

The option price can be obtained by undoing the above variable and parameter
substitutions. In the following we denote, as in Chap. 2, by $C(S, \tau)$ the call option
price being a function of the time to maturity $\tau = T - t$ instead of time t. Then it
holds

$$C(S, \tau) = e^{-r\tau} g(u, v) = e^{-r\tau} \int_{-\infty}^{\infty} \frac{1}{2\sqrt{\pi v}} g_0(\xi) e^{-\frac{(\xi-u)^2}{4v}} d\xi.$$

Substituting $\xi = (\beta - 1) \log(x/K)$ we obtain the original terminal condition
$\max\{0, x - K\}$. Furthermore, replacing u and v by the variables S and τ we obtain

$$C(S, \tau) = e^{-r\tau} \int_{0}^{\infty} \max(0, x - K) \frac{1}{\sqrt{2\pi}\sigma\sqrt{\tau}x}$$

$$\exp\left\{-\frac{[\log x - \{\log S + (b - \frac{1}{2}\sigma^2)\tau\}]^2}{2\sigma^2\tau}\right\} dx. \tag{6.22}$$

In the case of geometric Brownian motion $S_T - S_t$ is lognormally distributed, i.e. $\log(S_T - S_t)$ is normally distributed with parameters $(b - \frac{1}{2}\sigma^2)\tau$ and $\sigma^2\tau$. The conditional distribution of S_T given $S_t = S$ is therefore lognormal as well but with parameters $\log S + (b - \frac{1}{2}\sigma^2)\tau$ and $\sigma^2\tau$. However, the integrant in Eq. (6.22) is except for the term $\max(0, x - K)$ the density of the latter distribution. Thus, we can interpret the price of a call as the discounted expected option payoff $\max(0, S_T - K)$, which is the terminal condition, given the current stock price S :

$$C(S, \tau) = e^{-r\tau} \mathsf{E}[\max(0, S_T - K) \mid S_t = S]. \tag{6.23}$$

This property is useful when deriving numerical methods to compute option prices. But before doing that, we exploit the fact that Eq. (6.22) contains an integral with respect to the density of the lognormal distribution to further simplify the equation. By means of a suitable substitution we transform the term into an integral with respect to the density of the normal distribution and we obtain

$$C(S, \tau) = e^{(b-r)\tau} S \Phi(y + \sigma\sqrt{\tau}) - e^{-r\tau} K \Phi(y), \tag{6.24}$$

where we use y as a abbreviation for

$$y = \frac{\log \frac{S}{K} + (b - \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}. \tag{6.25}$$

Φ denotes the standard normal distribution function

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{z^2}{2}} dz.$$

Equations (6.24) and (6.26) are called the *Black–Scholes Formulae*. Figure 6.1 represents the Black–Scholes prices for the European call option $C(S, \tau)$ for different values of times to maturity τ. When τ goes to 0, the price of the call is approaching to the payoff of the option. The economic reason behind this is that there is a little probability for fluctuations in the underlying price within such a small interval τ.

Figure 6.2 displays the Black–Scholes price $C(S, \tau)$ as a function of S_t, which is modelled as a geometric Brownian motion. The figure shows that the option price moves according to the price fluctuations of the underlying.

For the limit cases $S \gg K$ and $S = 0$ it holds:

1. If $S \gg K$ then $y \gg 0$ and thus $\Phi(y + \sigma\sqrt{\tau}) \approx \Phi(y) \approx 1$. It follows that the value of a call option on a non-dividend paying stock, $b = r$, approaches $S - e^{-r\tau} K$. That is, it can be approximated by the current stock price minus the discounted exercise price.
2. If $S = 0$ then $y = -\infty$ and therefore $\Phi(y) = 0$. Thus the option is worthless: $C(0, \tau) = 0$.
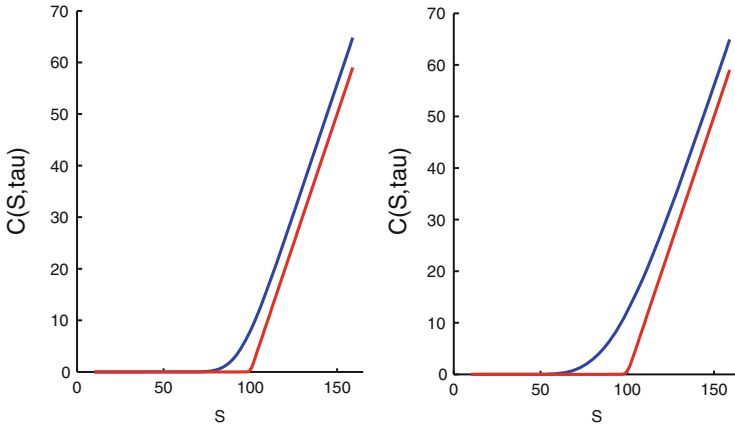
Fig. 6.1 Black–Scholes prices $C(S, \tau)$ for different times to maturity $\tau = 0.6$ and $r = 0.1$ and strike price $K = 100$. *Left figure* $\sigma = 0.15$, *right figure* $\sigma = 0.3$ Q SFEbsprices
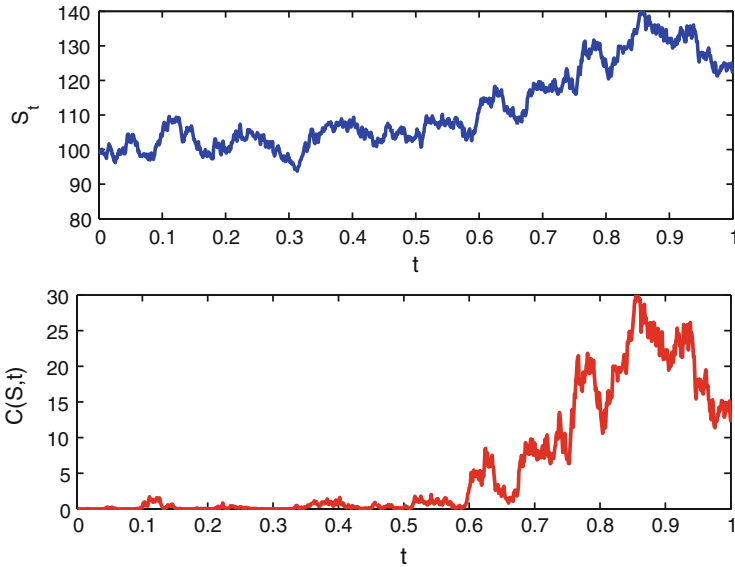
Fig. 6.2 *Upper panel*: sample path of the price process of the underlying S, *lower panel*: Black–Scholes prices $C(S, \tau)$ for strike $K = 100$, $r = 0.05$ and expiry at $T = 1$ where the initial value of the underlying is taken from the above sample path   Q SFEbsbm

The corresponding Black–Scholes Formula for the price $P(S, \tau)$ of a European put option can be derived by solving Black–Scholes differential equation subject to suitable boundary conditions. However, using the put–call parity (Theorem 2.3) is more convenient:

$$P(S, \tau) = C(S, \tau) - Se^{(b-r)\tau} + Ke^{-r\tau}.$$

From this and Eq. (6.24) we obtain

$$P(S, \tau) = e^{-r\tau} K \Phi(-y) - e^{(b-r)\tau} S \Phi(-y - \sigma\sqrt{\tau}). \tag{6.26}$$

As we see the value of European put and call options can be computed by explicit formulae. The terms in Eq. (6.24) for, say the value of a call option, can be interpreted in the following way. Restricting to the case of a non-dividend paying stock, $b = r$, the first term, $S \Phi(y + \sigma\sqrt{\tau})$, represents the value of the stock which the option holder obtains when he decides to exercise the option. The other term, $e^{-r\tau} K \Phi(y)$, represents the value of the exercise price. The quotient S/K influences both terms via the variable y.

Deriving Black–Scholes' differential equation we saw, in particular, that the value of a call option can be duplicated by means of bonds and stocks. The amount of money invested in stocks was $\frac{\partial C}{\partial S} S$ with $\frac{\partial C}{\partial S}$ being the hedge ratio. This ratio, also called delta, determines the relation of bonds and stocks necessary to hedge the option position. Computing the first derivative of Black–Scholes' formula in Eq. (6.24) with respect to S we obtain

$$\frac{\partial C(S, t)}{\partial S} = \Phi(y + \sigma\sqrt{\tau}).$$

Thus the first term in Eq. (6.24) reflects the amount of money of the duplicating portfolio invested in stocks, the second term the amount invested in bonds.

6.2.1   Numerical Approximation

Since the standard normal distribution can be evaluated only numerically, the implementation of Black–Scholes' formula depending on the standard normal distribution requires an approximation of the latter. This approximation can have an impact on the computed option value. To illustrate this, we consider several approximation formulae (e.g. Hastings, 1955).

(a) The normal distribution can be approximated in the following way:

$$\Phi(y) \approx 1 - (a_1 t + a_2 t^2 + a_3 t^3)e^{-y^2/2}, \text{ where } t = (1 + by)^{-1} \text{ and}$$

$$b = 0.332672527,$$
$$a_1 = 0.17401209,$$
$$a_2 = -0.04793922,$$
$$a_3 = 0.373927817.$$

The approximating error is independent of y of size 10^{-5}.

Q SFENormalApprox1

(b)

$$\Phi(y) \approx 1 - (a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5)e^{-y^2/2}, \text{ where } t = (1+by)^{-1} \text{ and}$$

$$b = 0.231641888,$$
$$a_1 = 0.127414796,$$
$$a_2 = -0.142248368,$$
$$a_3 = 0.71070687,$$
$$a_4 = -0.726576013,$$
$$a_5 = 0.530702714.$$

The error of this approximation is of size 10^{-7}.   ◘ SFENormalApprox2

(c) An approximation of the normal distribution, with error size 10^{-5} is given by:

$$\Phi(y) \approx \begin{cases} 0.5 - s, & \text{if } y < 0 \\ 0.5 + s, & \text{else} \end{cases}$$

where $t = |y|$,

$$s = \frac{1}{2} - \frac{1}{2(1 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5)^8}$$

and

$$a_1 = 0.099792714,$$
$$a_2 = 0.044320135,$$
$$a_3 = 0.009699203,$$
$$a_4 = -0.000098615,$$
$$a_5 = 0.00581551.$$

◘ SFENormalApprox3

(d) Finally we present the Taylor expansion:

$$\Phi(y) \approx \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\left(y - \frac{y^3}{1!2^13} + \frac{y^5}{2!2^25} - \frac{y^7}{3!2^37} + \cdots\right)$$

$$= \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\sum_{n=0}^{\infty}(-1)^n\frac{y^{2n+1}}{n!2^n(2n+1)}.$$

Table 6.1 Several approximations to the normal distribution

x	Norm-a	Norm-b	Norm-c	Norm-d	Iter
1.0000	0.8413517179	0.8413447362	0.8413516627	0.8413441191	6
1.1000	0.8643435425	0.8643338948	0.8643375717	0.8643341004	7
1.2000	0.8849409364	0.8849302650	0.8849298369	0.8849309179	7
1.3000	0.9032095757	0.9031994476	0.9031951398	0.9031993341	8
1.4000	0.9192515822	0.9192432862	0.9192361959	0.9192427095	8
1.5000	0.9331983332	0.9331927690	0.9331845052	0.9331930259	9
1.6000	0.9452030611	0.9452007087	0.9451929907	0.9452014728	9
1.7000	0.9554336171	0.9554345667	0.9554288709	0.9554342221	10
1.8000	0.9640657107	0.9640697332	0.9640670474	0.9640686479	10
1.9000	0.9712768696	0.9712835061	0.9712842148	0.9712839202	11
2.0000	0.9772412821	0.9772499371	0.9772538334	0.9772496294	12

Table 6.2 Prices of a European call option for different approximations of the normal distribution   **Q** SFEBSCopt1

Stock price S_t		230.00	EUR	
Exercise price K		210.00	EUR	
Time to maturity $\tau = T - t$		0.50000		
Continuous interest rate r		0.04545		
Volatility σ		0.25000		
No dividends				
	Norm-a	Norm-b	Norm-c	Norm-d
Option prices	30.74262	30.74158	30.74352	30.74157

By means of this series the normal distribution can be approximated arbitrarily close, depending on the number of terms used in the summation. Increasing the number of terms increases the number of arithmetic operations as well.

         Q SFENormalApprox4

Table 6.1 compares all four approximation formulae. The Taylor series was truncated after the first term where the absolute value was smaller than 10^{-5}. The last column shows the number of terms used.

Table 6.2 shows the price of a particular European call option computed by means of the four approximations presented above.

6.3   Simulation

Simulation techniques are essential tools of financial engineering. Their performance depends decisively on the quality of random numbers used, and today one cannot rely on random number generators in software packages to be satisfactory

in every respect. For applications, it should be checked that the generated random numbers reflect all the important features, depending on the particular application. Below, we briefly discuss how to simulate random variables.

6.3.1 Linear Congruential Generator

One of the most common pseudo random number generators is the linear congruential generator which uses a recurrence scheme to generate numbers:

$$N_i = (a N_{i-1} + b) \mod M \tag{6.27}$$

$$U_i = N_i / M \tag{6.28}$$

where N_i is the sequence of pseudo random numbers and (a, b, M) are generator-specific integer constants. mod is the modulo operation, a the multiplier and b the increment, $a, b, N_0 \in 0, 1, \ldots, M - 1$ with $a \neq 0$.

The linear congruential generator starts choosing an arbitrary seed N_0 and will always produce an identical sequence from that point on. The maximum amount of different numbers the formula can produce is the modulus M. The pseudo random variables N_i / M are uniformly distributed.

The period of a general linear congruential generator N_i is at most M, but in most cases it is less than that. The period should be large in order to ensure randomness, otherwise a small set of numbers can make the outcome easy to forecast. It may be convenient to set $M = 2^{32}$, since this makes the computation of $a N_{i-1} + b \mod M$ quite efficient.

In particular, $N_0 = 0$ must be ruled out in case $b = 0$, otherwise $N_i = 0$ would repeat. If $a = 1$, the sequence is easy to forecast and the generated sets are:

$$N_n = (N_0 + nb) \mod M$$

The linear congruential generator will have a full period if, Knuth (1997):

1. b and M are prime.
2. $a - 1$ is divisible by all prime factors of M.
3. $a - 1$ is a multiple of 4 if M is a multiple of 4.
4. $M > \max(a, b, N_0)$.
5. $a > 0, b > 0$.

Exactly, when the period is M, a grid point on a lattice over the interval [0,1] with size $\frac{1}{M}$ is occupied once.

Once the N_i numbers are generated, they can be arranged in m-tupels $(N_i, N_{i+1}, \ldots, N_{i+m-1})$ for $i \geq 1$. The corresponding points $(U_i, \ldots, U_{i+m-1}) \in [0, 1]^m$ will lie on $(m - 1)$ dimensional hyperplanes (parallel straight lines for

$m = 2$). Analysing the case for two planes or $m = 2$, where the distance between planes is large, the sequence is:

$$N_i = (aN_i + b) \mod M = aN_{i+1} + b - kM$$

for $kM \le aN_{i-1} + b < (k+1)M$.

For all integers z_0, z_1:

$$
\begin{aligned}
z_0 N_{i-1} + z_1 N_i &= z_0 N_{i-1} + z_1 (a N_{i-1} + b - kM) \\
&= N_{i-1}(z_0 + az_1) + z_1 b - z_1 kM \\
&= M(N_{i-1}\frac{z_0 + az_1}{M} - z_1 k) + z_1 b
\end{aligned}
$$

Hence

$$z_0 U_{i-1} + z_1 U_i = c + z_1 b M^{-1} \tag{6.29}$$

Therefore, for a given tupel (z_0, z_1) several parallel straight lines in the (U_{i-1}, U_i) plane are defined, one for each $c = c(i)$. Moreover, the distribution of the random numbers will be subject to the minimum number of hyperplanes or the maximum distance between them. That means, if a tupel exists (z_0, z_1) with only few of its straight lines cutting the square $[0, 1]^2$, then there will be areas of the square without random points, which will not satisfy the condition of a uniform distribution of the points in the unit square.

Notice that the variable c is an integer since only integer tupels are admitted and they require

$$z_0 + az_1 = 0 \mod M$$

The number of straight lines will be defined by solving c in Eq. (6.29). Applying $U_i \sim U[0, 1)$ an interval $c_{min} \le c \le c_{max}$ is obtained. For each c in that interval, its straight line cuts the square $[0, 1]^2$.

Example 6.1 Consider the sequence $N_i = 2N_{i-1} \mod 11$, with values $a = 2, b = 0, M = 11$. The solutions to Eq. (6.29) are $z_0 = -2$ and $z_1 = 1$. The family of straight lines in the (U_{i-1}, U_i) plane looks like:

$$-2U_{i-1} + U_i = c$$

For $U_i \sim U[0, 1]$, the variable c is defined over the interval $(-2, 1)$. Since c is an integer, only $c = -1$ and $c = 0$ hold and cut the interior of the square $[0, 1)^2$. Figure 6.3 shows that the points generated by the algorithm are lying on the straight lines with $c = -1$ and $c = 0$, whose points form a lattice.
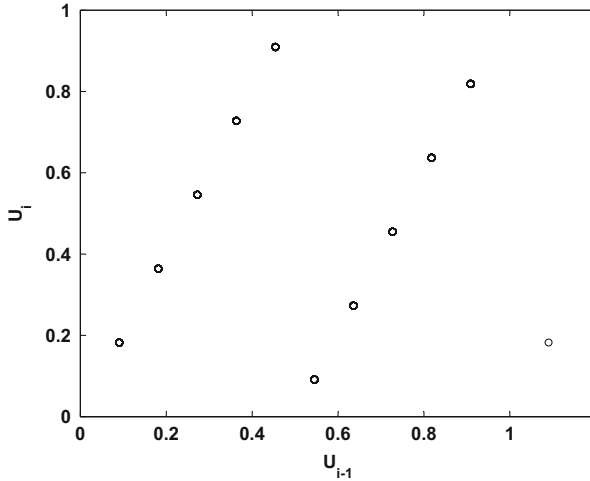
Fig. 6.3 The scatterplot of (U_{i-1}, U_i) Q SFErangen1

Example 6.2 Suppose $N_i = 1229 N_{i-1} \mod 2048$. The values of the tupel (z_0, z_1) are $z_0 = -1$ and $z_1 = 5$, because

$$-1 + 1229(5) = 6,144 = 3(2048)$$

The family of straight lines in the (U_{i-1}, U_i) plane is

$$-1 U_{i-1} + 5 U_i = c$$

and the distance between the straight lines over the vertical U_i axis is $\frac{1}{z_1} = \frac{1}{5}$. Figure 6.4 shows that all the points (U_{i-1}, U_i) generated are lying on six straight lines, with $c \in -1, 0, 1, 2, 3, 4$.

Pseudo random numbers produced by linear congruential generators are extremely sensitive to the choice of a, b and M. Historically, poor choices had ruined implementations of linear congruential generators. A famous example of this is RANDU, the official IBM $U[0, 1]$ generator for years. It is considered to be one of the most ill-conceived random number generators. The RANDU generator is defined as

$$N_i = a N_{i-1} \mod M$$

with $a = 2^{16} + 3$, $M = 2^{31}$ and N_0 odd. The defined values were chosen, because with a 32 bit integer word size the arithmetic of $\mod 2^{31}$ and $2^{16} + 3 = 65,539$ calculations could be done quickly, see Knuth (1997).
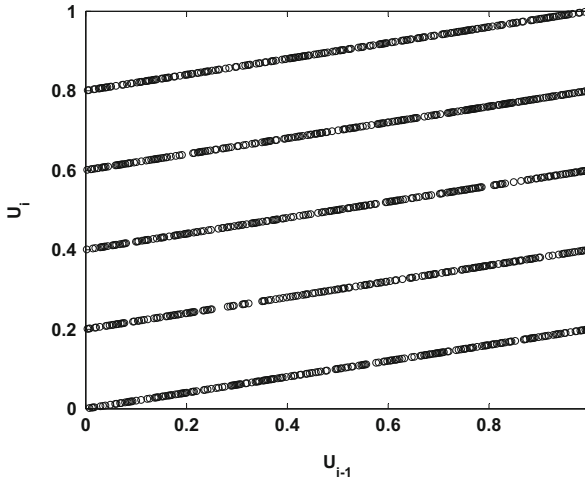
Fig. 6.4 The scatterplot of (U_{i-1}, U_i) ◻ SFErangen2

This generator fails the spectral test for dimensions greater than 2, which is a quality test for linear congruential generators. It is based on the fact that if a linear congruential generator is used to choose points in an n-dimensional space, triples of points will lie on, at most, $M^{\frac{1}{n}}$ hyperplanes. This is due to serial correlation between successive values of the sequence N_i. As a result of the wide use of RANDU in the early 1970s many simulation studies from that time are seen as suspicious (Press et al. (1992)).

To show the problem with the defined values consider the previous example, but this time every term should be taken with   mod 2^{31}. After recursion, the random points in the cube $[0, 1]^3$ lie on only 15 planes. Figure 6.5 shows the scatterplot (U_{i-2}, U_{i-1}, U_i) for the previous example using RANDU generator.

In practice, many common linear congruential generators fail statistical significance tests when exhibiting shorter than expected periods for some seed states, poor dimensional distribution, dependence between successive values, some bits being more random than others and a lack of uniformity. A disadvantage of the linear congruential generators is the restrictedness of the period M, which is limited to a 32 bit size. However, this can be corrected by *shuffling* the random numbers in a random way and the period will get close to infinity.
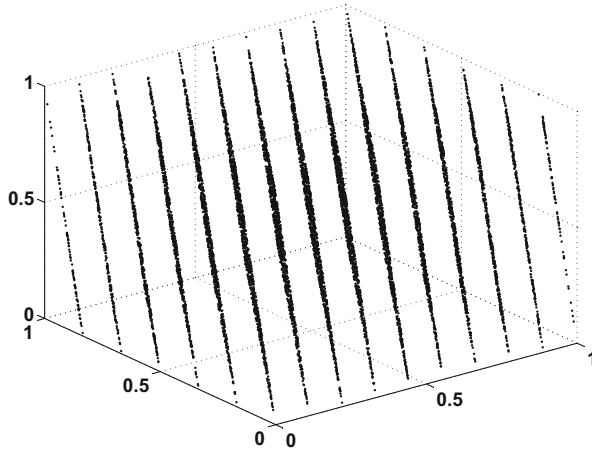
Fig. 6.5 The scatterplot of (U_{i-2}, U_{i-1}, U_i) ▣ SFErandu

6.3.2   Fibonacci Generators

Another example of pseudo random number generators is the Fibonacci generators, whose aim is to improve the standard linear congruential generator. These are based on a Fibonacci sequence:

$$N_{i+1} = N_i + N_{i-1} \mod M$$

This recursion formula is related to the Golden ratio. The ratio of consecutive Fibonacci numbers $\frac{F(n+1)}{F(n)}$ converges to the golden ratio γ as the limit, defined as one solution equal to $\frac{1+\sqrt{5}}{2} = 1.6180$ of the equation $x = 1 + \frac{1}{x}$.

The original formula is a three term recursion, which is not appropriate for generating random numbers. The modified approach, the lagged Fibonacci generator is defined as

$$N_{i+1} = N_{i-\nu} + N_{i-\mu} \mod M$$

for any $\nu, \mu \in \mathbb{N}$.

The quality of the outcome for this algorithm is sensitive to the choice of the initial values, ν and μ. Any maximum period of the lagged Fibonacci generator has a large number of different possible cycles. There are methods where a cycle can be chosen, but this might endanger the randomness of future outputs and statistical defects may appear.

Example 6.3 Let $\nu = 17$ and $\mu = 5$ be values in $\mathbb{N}$, the lagged Fibonacci generator is:

$$U_i = (U_{i-17} - U_{i-5})$$

if $U_i < 0$ then set $U_i = U_i + 1.0$.

This recursion produces random numbers $U_i \in [0, 1)$ and requires the generation of the initial 17 uniforms by a congruential generator. The algorithm for Fibonacci generators with seed N_0, equal to any number, is simplified as follows:
    Repeat:

$$\zeta = U_i - U_j$$

$$\text{if } \zeta < 0 : \zeta = \zeta + 1$$

$$U_i = \zeta$$

$$i = i - 1$$

$$j = j - 1$$

$$\text{if } i = 0 : i = 17$$

$$\text{if } j = 0 : j = 17$$

Figure 6.6 shows 10,000 random points generated with the Fibonacci Algorithm, where the initial uniform random variables $U_1 \ldots U_{17}$ were calculated with a linear
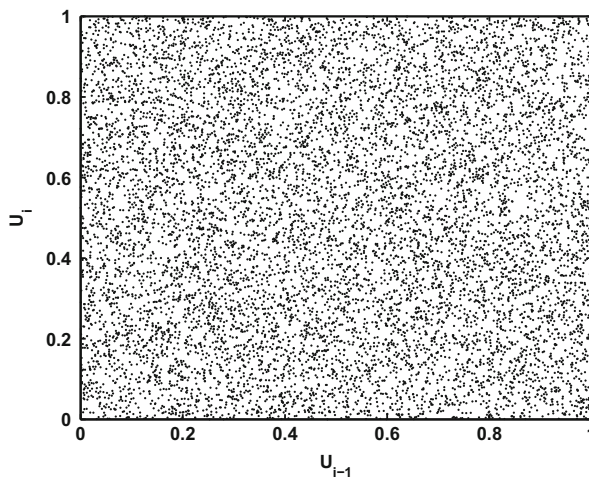
Fig. 6.6 The scatterplot of (U_{i-1}, U_i) ▣ SFEfibonacci

congruential generator with values M=714,025, a=1,366, b=150,889. Notice that the points show a random structure.

6.3.3 Inversion Method

Many programming languages can generate pseudo-random numbers which are distributed according to the standard uniform distribution and whose probability is the length $b - a$ of the interval $(a, b) \in (0, 1)$. *The inverse method* is a method of sampling a random number from any probability distribution, given its cumulative distribution function (cdf).

Suppose $U_i \sim U[0, 1]$ and $F(x)$ a strictly increasing continuous distribution then $X_i \sim F$, if $X_i = F^{-1}(U_i)$.

Proof

$$P(X_i \leq x) = P\left\{F^{-1}(U_i) \leq x\right\} = P\left\{U_i \leq F(x)\right\} = F(x)$$

Usually F^{-1} is often hard to calculate, but the problem can be solved using transformation methods. Suppose that X is a random variable with the density function $f(x)$ and the distribution function $F(x)$. Further assume h be strictly monotonous, then $Y = h(X)$ has the distribution function $F\left\{h^{-1}(y)\right\}$. If h^{-1} is continuous, then for all y the density of $h(X)$ is, Härdle and Simar (2012):

$$f_Y(y) = f_X\left\{h^{-1}(y)\right\}\left|\frac{dh^{-1}(y)}{dy}\right|$$

Example 6.4 Apply the transformation method in the exponential case. The density of an exponential function is $f_Y(y) = \lambda \exp\{-\lambda y\} I(y \geq 0)$ with $\lambda \geq 0$, and its inverse is equal to $h^{-1}(y) = \exp\{-\lambda y\}$ for $y \geq 0$. Define $y = h(x) = -\lambda^{-1} \log x$ with $x > 0$. We would like to know whether $X \sim U[0, 1]$ leads to an exponentially distributed random variable $Y \sim \exp(\lambda)$.

Using the definition of the transformation method, we have

$$f_Y(y) = f_X\left\{h^{-1}(y)\right\}\left|\frac{dh^{-1}(y)}{dy}\right| = |(-\lambda)\exp\{-\lambda y\}| = \lambda \exp\{-\lambda y\}$$

Hence $f_Y(y)$ is exponentially distributed.

6.3.4 Box–Muller Method

The Box–Muller method allows us to transform uniformly distributed random variables to a set of independent standard normally distributed random numbers. Let the unit square be $S = [0, 1]^2$ and $f_X(x) = 1$ the density of the uniform distribution. The most basic form of the transformation looks like:

$$y_1 = \sqrt{-2 \log x_1} \cos 2\pi x_2 = h_1(x_1, x_2)$$

$$y_2 = \sqrt{-2 \log x_2} \sin 2\pi x_2 = h_2(x_1, x_2)$$

where $h(x)$ is defined on S. Notice that when x_1 is very close to zero, the transformation can have numerical stability problems, especially during the generation of numbers in stochastic modelling.

In this method we start with two independent random numbers, x_1 and x_2, which are uniformly distributed over (0,1). Then, we apply the above transformations to get two new independent random numbers, which have a normal distribution with a zero mean and a standard deviation of one.

Following the transformation method approach, the inverse function $h(x)^{-1}$ is given by

$$h^{-1}(x) = \begin{cases} x_1 = \exp\left\{-\frac{1}{2}(y_1^2 + y_2^2)\right\} \\ x_2 = (2\pi)^{-1} \arctan \frac{y_2}{y_1} \end{cases}$$

The determinant of the Jacobian matrix is

$$|Jacobian| = \det \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{pmatrix}$$

with

$$\frac{\partial x_1}{\partial y_1} = \exp\left\{-\frac{1}{2}(y_1^2 + y_2^2)\right\} (-y_1)$$

$$\frac{\partial x_1}{\partial y_2} = \exp\left\{-\frac{1}{2}(y_1^2 + y_2^2)\right\} (-y_2)$$

$$\frac{\partial x_2}{\partial y_1} = \frac{1}{2\pi} \left(\frac{1}{1 + \frac{y_2^2}{y_1^2}}\right) \frac{y_2}{y_1^2}$$

$$\frac{\partial x_2}{\partial y_2} = \frac{1}{2\pi} \left(\frac{1}{1 + \frac{y_2^2}{y_1^2}} \right) \frac{1}{y_1}$$

Then,

$$|Jacobian| = \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(y_1^2 + y_2^2)\right\} \left(-y_1 \frac{1}{1 + \frac{y_2^2}{y_1^2}} \frac{1}{y_1} - y_2 \frac{1}{1 + \frac{y_2^2}{y_1^2}} \frac{y_2}{y_1^2} \right)$$

$$= -\frac{1}{2\pi} \exp\left\{-\frac{1}{2}(y_1^2 + y_2^2)\right\}$$

The last equation states that the determinant of the Jacobian is the density of the standard normal distribution in $\mathbb{R}^2$. The components of the vector Y are independent since this density is the product of two univariate densities. Hence $(Y_1, Y_2)\top \sim N_2(0, I_2)$.

The next Box–Muller algorithm simplifies the previous results.

1. $U_1 \sim U[0, 1], U_2 \sim U[0, 1]$
2. $\theta = 2\pi U_2, \rho = \sqrt{-2 \log U_1}$
3. $Z_1 = \rho \cos \theta$ is $N(0, 1)$
   $Z_2 = \rho \sin \theta$ is $N(0, 1)$

Summarizing, the Box–Muller algorithm provides two standard normal distributions, when the components of the vector X are $\sim U[0, 1]$. Figure 6.7 shows the results.

6.3.5 Marsaglia Method

The variant of Marsaglia method is the polar form of the Box–Muller transformation. It is faster and more robust since it avoids the calculation of the trigonometric functions in the Box–Muller method.

This method generates uniformly distributed random values V_1, V_2 over the interval [-1,1] by transforming $U[0, 1]$ variables. Two values (V_1, V_2) define a point in the (V_1, V_2) plane. The points will be accepted if $V_1^2 + V_2^2 < 1$, i.e. (V_1, V_2) are uniformly distributed on the unit circle with the density $f(V_1, V_2) = \pi^{-1}$.

The transformation that maps the coordinates of the unit circle into the unit square S is

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} V_1^2 + V_2^2 \\ 2\pi^{-1} \arctan(V_2/V_1) \end{pmatrix}$$

Hence $(X_1, X_2)^\top$ is uniformly distributed on S.
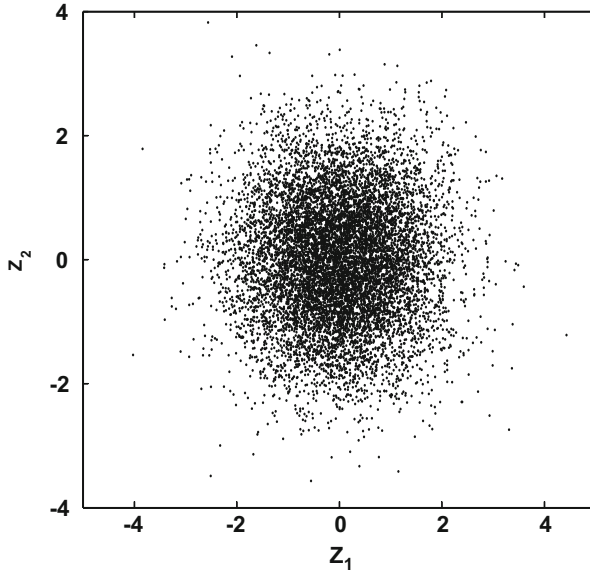
Fig. 6.7 The scatterplot of (Z_1, Z_2) 🇶 SFEbmuller

Using the previous results as inputs in the Box–Muller method, the following
relationships hold

$$\cos 2\pi X_2 = \frac{V_1}{\sqrt{V_1^2 + V_2^2}}$$

$$\sin 2\pi X_2 = \frac{V_2}{\sqrt{V_1^2 + V_2^2}}$$

and it is no longer necessary to evaluate trigonometric functions.

The Marsaglia (Polar method) algorithm is described by

1. $U_1, U_2 \sim U[0, 1]$; $V_i = 2U_i - 1$ with $W = V_1^2 + V_2^2 < 1$
2. $Z_1 = V_1 \sqrt{-2 \log(W)/W} \sim N(0, 1)$
   $Z_2 = V_2 \sqrt{-2 \log(W)/W} \sim N(0, 1)$

The Marsaglia Polar method, which is a type of rejection sampling, is the polar
equivalent of the Box–Muller method. Typically it is faster, because it uses only one
transcendental function instead of at least two, even though it throws away $1 - \frac{\pi}{4} \approx$
21.46 % of the total input uniformly distributed random number pairs per normal
pair variable generated, requiring $\frac{4}{\pi} \approx 1.2732$ input random numbers per output
random number.

6.4   Risk Management and Hedging

Trading options is particularly risky due to the possibly high random component. Advanced strategies to reduce and manage this risk can be derived from the Black–Scholes formula (6.24). To illustrate this issue we consider an example and some traditional strategies.

Example 6.5   A bank sells a European call option to buy 100,000 shares of a non-dividend paying stock for 600,000 EUR. The details of this option are given in Table 6.3.

   Applying Black–Scholes' formula (6.24) for a non-dividend paying stock, $b = r$, gives a theoretical value of 480,119 EUR, approximately 480,000 EUR, of the above option. That is, the bank sold the option about 120,000 EUR above its theoretical value. But it takes the risk to incur substantial losses.

   A strategy to manage the risk due to the option would be to do nothing, i.e. to take a *naked position*. Should the option be exercised at maturity the bank has to buy the shares for the stock price prevailing at maturity. Assume the stock trades at $S_T = 120$ EUR. Then an options' exercise costs the bank $100{,}000 \cdot (S_T - K) = 2{,}000{,}000$ EUR, which is a multiple of what the bank received for selling the derivative. However, if the stock trades below $K = 100$ EUR the option will not be exercised and the bank books a net gain of 600,000 EUR.   Q SFEBSCopt2

   In contrast to the naked position, it is possible to set up a *covered position* by buying 100,000 shares at $100{,}000 \cdot S_t = 9{,}800{,}000$ EUR at the same time the option is sold. In case $S_T > K$ the option will be exercised and the stocks will be delivered at a price of $100{,}000 \cdot K = 10{,}000{,}000$ EUR, which discounted to time t is about 9,800,000 EUR. Thus the bank's net gain is equal to 600,000 EUR, the price at which the option is sold. If the stock price decreases to $S_T = 80$ EUR then the option will not be exercised. However, the bank incurs a loss of 2,000,000 EUR due to the lower stock price, which is, as above, a multiple of the option price. Note that from the put–call parity for European options (Theorem 2.3) it follows that the risk due to a covered short call option position is identical to the risk due to naked long put option position.

   Both risk management strategies are unsatisfactory because the cost varies significantly between 0 and large values. According to Black–Scholes the option

Table 6.3   Data of the example

Current time t	6 weeks
Maturity T	26 weeks
Time to maturity $\tau = T - t$	20 weeks = 0.3846
Continuous annual interest rate r	0.05
Annualized stock volatility σ	0.20
Current stock price S_t	98 EUR
Exercise price K	100 EUR

costs on average around 480,000 EUR, and a perfect hedge eliminates the impact of random events such that the option costs exactly this amount.

An expensive hedging strategy, i.e. a strategy to decrease the risk associated with the sale of a call, is the so-called *stop-loss strategy*: The bank selling the option takes an uncovered position as long as the stock price is below the exercise price, $S_t < K$, and sets up a covered position as soon as the call is in-the-money, $S_t > K$.

The shares to be delivered in case of options exercise are bought as soon as the stock S_t trades above the exercise price K, and are sold as soon as S_t falls below the exercise price K.

Since all stocks are sold and bought at K after time 0 and at maturity T either the stock position is zero, $(S_t < K)$, or the stocks are sold at K to the option holder, $(S_t > K)$, this strategy bears no costs under perfect market assumptions.

Note that playing a stop-loss strategy bears a cost if $S_0 > K$, i.e. stocks are bought at S_0 and sold at K :

$$\text{costs of a stop-loss hedging strategy: } \max(S_0 - K,\ 0).$$

Because these costs are smaller than the Black–Scholes price $C(S_0, T)$ arbitrage would be possible by running a stop-loss strategy. However, this reasoning ignores some aspects:

- Buying and selling stocks bear transaction costs,
- Buying stocks before time T involves binding capital leading to renounce of interest rate revenue,
- practically it is not possible to buy or sell stocks exactly at K rather stocks are bought at $K + \delta$ if stocks are increasing and stocks are sold at $K - \delta$ if stocks are decreasing, for a $\delta > 0$.

In practice, purchases and sales take place only after Δt time units. The larger Δt, the greater δ in general, and the less transaction costs have to be paid. Hull (2006) investigated in a Monte Carlo study with $M = 1,000$ simulated stock price paths the stop-loss strategy's ability to reduce the risk associated with the sale of a call option. For each simulated path the costs $\Lambda_m, m = 1, \ldots, M$, caused by applying the stop-loss strategy are registered and their sample variance

$$\hat{v}_\Lambda^2 = \frac{1}{M} \sum_{m=1}^{M} (\Lambda_m - \frac{1}{M} \sum_{j=1}^{M} \Lambda_j)^2$$

is computed. Dividing the sample standard deviation by the call price measures the remaining risk of the stop-loss hedged short call position

$$L = \frac{\sqrt{\hat{v}_\Lambda^2}}{C(S_0, T)}.$$

Table 6.4 Performance of the stop-loss strategy

Δt (weeks)	5	4	2	1	$\frac{1}{2}$	$\frac{1}{4}$
L		1.02	0.93	0.82	0.77	0.76	0.76

Table 6.4 shows the results. A perfect hedge would reduce the risk to zero, i.e. $L = 0$.

In order to apply the concept of risk neutral valuation (see Cox and Ross (1976)) the probability measure has to be transformed such that the price process under this new measure is a martingale. By doing this the absence of arbitrage opportunities is guaranteed. In incomplete markets, however, a multitude of such transformations exist, see Harrison and Kreps (1979). In contrast to complete markets the trader cannot build up a self-financing portfolio reproducing the options payoff at maturity when the market is incomplete. Therefore hedging is no longer risk free, and option prices depend on risk preferences. In this context we want to point out that the lack of a perfect hedge is of great importance in practice.

6.4.1   Delta Hedging

In order to reduce the risk associated with option trading, more complex hedging strategies than those considered so far are applied. Let us look at the following example. Sell a call option on a stock, and try to make the value of this portfolio for small time intervals as insensitive as possible to small changes in the price of the underlying stock. This is what is called delta hedging. Later on, we consider further *Greeks* (gamma, theta, vega, etc.) to fine tune the hedged portfolio.

By the *delta* or the *hedge ratio* we understand the derivative of the option price with respect to the stock price. In a discrete time model we use the differential quotient of the change in the option price ΔC with respect to a change in the stock price ΔS :

$$\Delta = \frac{\partial C}{\partial S} \quad \text{or} \quad \Delta = \frac{\Delta C}{\Delta S}.$$

The delta of other financial instruments is defined accordingly. The stock itself has the value S. Consequently it holds that $\Delta = \partial S/\partial S = 1$. A futures contract on a non-dividend paying stock has a value of $V = S - K \cdot e^{-r\tau}$ (see Theorem 2.1) and thus its delta is $\Delta = \partial V/\partial S = 1$ as well. Stocks and future contracts can therefore be used equivalently in delta hedging strategies. If the latter are available, they are preferable due to lower transaction costs.

Example 6.6 A bank sells calls on 2,000 shares of a stock for a price of $C = 10$ EUR/share at a stock price of $S_0 = 100$ EUR/share. Let the call's delta be $\Delta = 0.4$.

To hedge the sold call options $\Delta \cdot 2000 = 800$ shares of the stock are added to the portfolio. Small changes in the option value will be offset by corresponding changes in the value of the portfolio's stock shares. Should the stock price increase by 1 EUR, i.e. the value of the stock position in the portfolio increases by 800 EUR, the value of one call on 1 share increases by $\Delta C = \Delta \cdot \Delta S = 0.4$ EUR and following the value of the portfolio's short call position decreases by 800 EUR. That is, gains and losses offset each other because the delta of the option position is neutralized by the delta of the stock position. The portfolio has a $\Delta = 0$, and the bank takes a *delta neutral position*.

Since the delta of an option depends on the stock price and time, among others, the position is only delta neutral for a short period of time. In practice, the portfolio has to be *re-balanced* frequently in order to adapt to the changing environment. Strategies to manage portfolio risk which involve frequent re-balancing are known as *dynamic hedging*. We point out that the Black–Scholes differential equation (6.3) can be derived by means of a dynamic hedge portfolio whose position is kept continuously delta neutral. This approach is analogous to reproducing the option by a duplicating portfolio.

Example 6.7 In continuation of Example 6.6, suppose that the underlying stock rises within a week to 110 EUR. Due to the time delay and the increased stock price, the option delta increases to $\Delta = 0.5$. In order to re-obtain a delta neutral position $(0.5 - 0.4) \cdot 2000 = 200$ shares of the stock have to be bought.

From the Black–Scholes formulae for the value of European call and put options on non-dividend paying stocks it follows for the delta that:

$$\Delta = \frac{\partial C}{\partial S} = \Phi(y + \sigma \sqrt{\tau}) \qquad (6.30)$$

$$\text{bzw. } \Delta = \frac{\partial P}{\partial S} = \Phi(y + \sigma \sqrt{\tau}) - 1,$$

with y being defined in Eq. (6.25).

Figure 6.8 displays the delta (6.30) as a function of time and stock price. For an increasing stock price delta converges to 1, for decreasing stock prices it converges to 0. Put differently, if the option is deep in-the-money (ITM) it will be exercised at maturity with a high probability. That is the reason why the seller of such an option should be long in the underlying to cover the exercise risk. On the other hand, if the option is far out-of-the-money it will probably not be exercised, and the seller can restrict themselves to holding a smaller part of the underlying.

Furthermore, the probability p that an out-of-the-money (OTM) option will be exercised and an ITM option will not be exercised at maturity is higher the longer the time to maturity. This explains why the delta for longer times to maturity becomes flatter (linear).

Fig. 6.8 Delta of European Call Option as a function of the stock price (*right axis*) and time to maturity (*left axis*)   **Q** SFEdelta

Table 6.5 Performance of the delta-hedging strategy

Δt (weeks)	5	4	2	1	$\frac{1}{2}$	$\frac{1}{4}$
L		0.43	0.39	0.26	0.19	0.14	0.09

Table 6.5 according to Hull (2006) shows (in the same spirit as Table 6.4) the performance of the delta hedging strategy contingent on the time increments Δt between re-balancing trades. If Δt is small enough the risk associated with a sold call option can be managed quite well. In the limit $\Delta t \to 0$ continuous re-balancing underlying the derivation of the Black–Scholes formula follows, and the risk is perfectly eliminated ($L = 0$). The linearity of the mathematical derivative implies for the delta Δ_p of a portfolio consisting of $w_1, \dots, w_m$ contracts of m financial derivatives $1, \dots, m$ with deltas $\Delta_1, \dots, \Delta_m$:

$$\Delta_p = \sum_{j=1}^{m} w_j \Delta_j.$$

Example 6.8 Consider a portfolio consisting of the following USD options

1. 200,000 bought calls (*long position*) with exercise price 1.70 EUR maturing in 4 months. The delta of an option on 1 USD is $\Delta_1 = 0.54$.

2. 100,000 written calls (*short position*) with exercise price 1.75 EUR maturing in 6 months and a delta of $\Delta_2 = 0.48$.
3. 100,000 written puts (short position) with exercise price 1.75 EUR maturing in 3 months with $\Delta_3 = -0.51$.

The portfolio's delta is (increases in values of written options have a negative impact on the portfolio value):

$$\Delta_p = 200\,000 \cdot \Delta_1 - 100\,000 \cdot \Delta_2 - 100\,000 \cdot \Delta_3$$
$$= 111\,000$$

The portfolio can be made delta neutral by selling 111,000 USD or by selling a corresponding future contract on USD (both have a delta of $\Delta = 1$).

6.4.2 Gamma and Theta

Using the delta to hedge an option position, the option price is locally approximated by a function which is linear in the stock price S. Should the time Δt elapsing until the next portfolio re-balancing not be very short, then this approximation would no longer be adequate (see Table 6.5). That is why a more accurate approximation, the Taylor expansion of C as a function of S and t, is considered:

$$\Delta C = C(S+\Delta S,\, t+\Delta t) - C(S,t) = \frac{\partial C}{\partial S} \cdot \Delta S + \frac{\partial C}{\partial t} \cdot \Delta t + \frac{1}{2}\frac{\partial^2 C}{\partial S^2}(\Delta S)^2 + \mathcal{o}(\Delta t),$$

where (as we have already seen in the demonstration of Theorem 6.1) ΔS is of size $\sqrt{\Delta t}$ and the terms summarized in $\mathcal{o}(\Delta t)$ are of size smaller than Δt. Neglecting all terms but the first, which is of size $\sqrt{\Delta t}$, the approximation used in delta hedging is obtained:

$$\Delta C \approx \Delta \cdot \Delta S.$$

Also taking the terms of size Δt into account it follows that:

$$\Delta C \approx \Delta \cdot \Delta S + \Theta \cdot \Delta t + \frac{1}{2}\Gamma(\Delta S)^2,$$

where $\Theta = \partial C / \partial t$ is the option's *theta* and $\Gamma = \partial^2 C / \partial S^2$ is the option's *gamma*. Θ is also called the option's *time decay*. For a call option on a non-dividend paying stock it follows from the Black–Scholes formula (6.24):

$$\Theta = -\frac{\sigma S}{2\sqrt{\tau}}\,\varphi(y + \sigma\sqrt{\tau}) - rKe^{-r\tau}\Phi(y) \tag{6.31}$$

$$\Gamma = \frac{1}{\sigma S \sqrt{\tau}} \, \varphi(y + \sigma \sqrt{\tau}), \tag{6.32}$$

where y is defined in Eq. (6.25).

Figures 6.9 and 6.10 display theta and gamma given by Eq. (6.31) respectively (6.32) as a function of stock price and time to maturity. Most sensitive to movements in stock prices are at-the-money options with a short time to maturity. Consequently, to hedge such options the portfolio has to be rebalanced frequently.

Assuming a delta neutral portfolio *gamma hedging* consists of buying or selling further derivatives to achieve a gamma neutral portfolio, i.e. $\Gamma = 0$, and thereby making the portfolio value even more insensitive to changes in the stock price. Note that on the one hand neither stocks nor future contracts can be used for gamma hedging strategies since both have a constant Δ and thus a zero gamma $\Gamma = 0$. On the other hand, however, those instruments can be used to make a gamma neutral portfolio delta neutral without affecting the portfolio's gamma neutrality. Consider an option position with a gamma of Γ. Using w contracts of an option traded on a stock exchange with a gamma of Γ_B, the portfolio's gamma is $\Gamma + w\Gamma_B$. By setting $w = -\Gamma/\Gamma_B$ the resulting gamma for the portfolio is 0.

Example 6.9 Let a portfolio of USD options and US-Dollars be delta neutral with a gamma of $\Gamma = -150{,}000$. A USD-call trade on the exchange with $\Delta_B = 0.52$ and $\Gamma_B = 1.20$. By adding $-\Gamma/\Gamma_B = 125{,}000$ contracts of this option the portfolio becomes gamma neutral. Unfortunately, its delta will be $125{,}000 \cdot \Delta_B = 65{,}000$. The delta neutrality can be achieved by selling 65,000 USD without changing the gamma.

Contrary to the evolution of the stock price the expiry of time is deterministic, and time does not involve any risk increasing the randomness. If both Δ and Γ are 0 then the option value changes (approximately risk free) at a rate $\Theta = \Delta C / \Delta t$. The parameter Θ is for most options negative, i.e. the option value decreases as the maturity date approaches.

From Black–Scholes's formula (6.24) it follows for a delta neutral portfolio consisting of stock options

$$rV = \Theta + \frac{1}{2}\sigma^2 S^2 \Gamma,$$

with V denoting the portfolio value. Θ and Γ depend on each other in a straightforward way. Consequently, Θ can be used instead of Γ to gamma hedge a delta neutral portfolio.
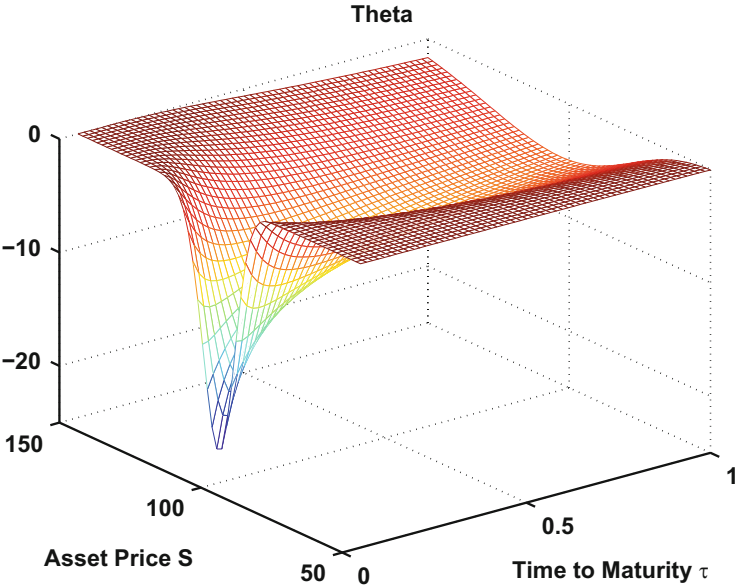
Fig. 6.9 Theta of European Call Option as a function of stock price (*right axis*) and time to maturity (*left axis*)   **Q** SFEtheta
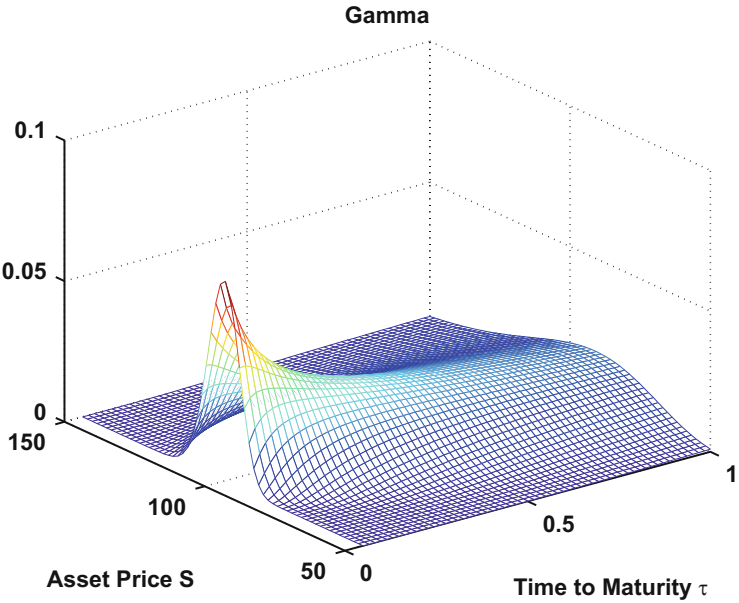
Fig. 6.10 Gamma of European Call Option as a function of stock price (*right axis*) and time to maturity (*left axis*)   **Q** SFEgamma

6.4.3 Rho and Vega

The Black–Scholes' approach proceeds from the assumption of a constant volatility σ. The appearance of smiles indicates that this assumption does not hold in practice. Therefore, it can be useful to make the portfolio value insensitive to changes in volatility. By doing this, the *vega* of a portfolio (in literature sometimes also called lambda or kappa) is used, which for a call option is defined by $\mathcal{V} = \frac{\partial C}{\partial \sigma}$.

For stocks and future contracts it holds $\mathcal{V} = 0$. Thus, in order to set up a vega hedge one has to make use of traded options. Since a vega neutral portfolio is not necessarily delta neutral two distinct options have to be involved to achieve simultaneously $\mathcal{V} = 0$ and $\Delta = 0$.

From Black–Scholes' formula (6.24) and the variable y defined in Eq. (6.25) it follows that the vega of a call option on a non-dividend paying stock is given by:

$$\mathcal{V} = S \sqrt{\tau} \varphi(y + \sigma \sqrt{\tau}). \tag{6.33}$$

Since the Black–Scholes formula was derived under the assumption of a constant volatility it is actually not justified to compute the derivative of (6.24) with respect to σ. However, the above formula for $\mathcal{V}$ is quite similar to an equation for $\mathcal{V}$ following on from a more general stochastic volatility model. For that reason, Eq. (6.33) can be used as an approximation to the real vega.

Figure 6.11 displays the vega given by Eq. (6.33) as a function of stock price and time to maturity. When the option is in-the-money, vega is low. If the stock price approaches the strike price, vega increases and reaches its peak when the option becomes at-the-money. Further, if the option becomes out-of-the-money, the sensitivity of the option with respect to volatility is low again. Thus, at-the-money options with a long time to maturity are most sensitive to changes in volatility.

In order to protect the portfolio against implied volatility fluctuations, investors can use a *vega hedging* strategy. If the vega of a portfolio and of an option are denoted by $\mathcal{V}_{Port}$ and $\mathcal{V}_{Opt}$ respectively, one can take a $-\mathcal{V}_{Port}/\mathcal{V}_{Opt}$ position in the option to make a portfolio vega neutral. However, a vega neutral position is not stable and thus, even small changes in stock prices might cost investors a lot of money.

Finally, the call option's risk associated with movements in interest rates can be reduced by using *rho* to hedge the position:

$$\rho = \frac{\partial C}{\partial r}.$$

For a call on a non-dividend paying stock it follows from Eq. (6.24)

$$\rho = K \, \tau \, e^{-r\tau} \Phi(y).$$
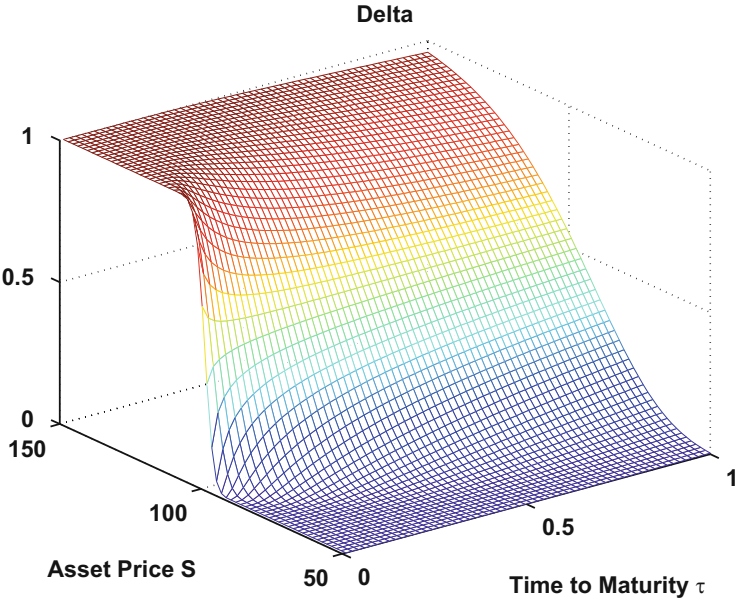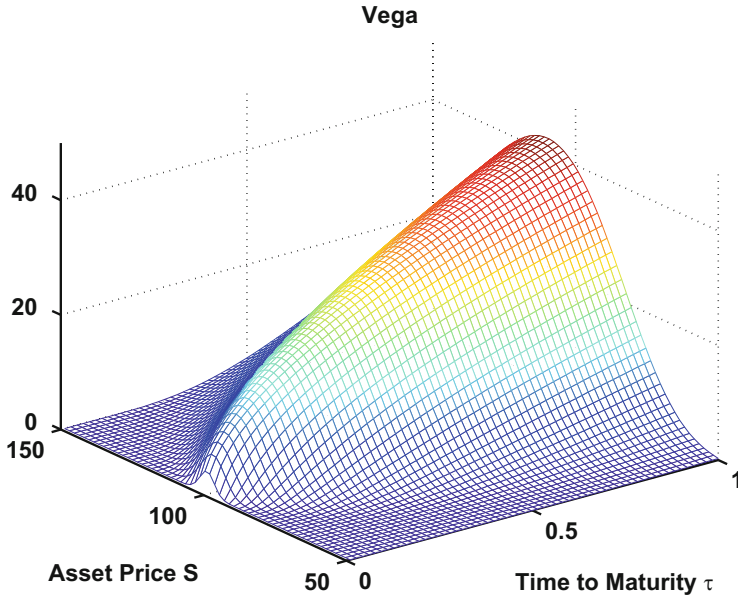
Fig. 6.11 Vega of European Call Option as a function of stock price (*right axis*) and time to maturity (*left axis*)   ◩ SFEvega

When hedging currency options domestic as well as foreign interest rates have to be taken into account. Consequently, rho hedging strategies need to consider two distinct values ρ_1 and ρ_2.

6.4.4 Volga and Vanna

Volga, also known as volgamma or vomma, is the sensitivity of vega to the change in implied volatility. Hence, it can also be defined as the second derivative of the option price with respect to volatility:

$$Volga = \frac{\partial \mathcal{V}}{\partial \sigma} = \frac{\partial^2 C}{\partial^2 \sigma}.$$

Calculating volga from the Black–Scholes formula, for a call option we obtain:

$$Volga = S \cdot \sqrt{\tau} \cdot \frac{y \cdot \left(y + \sigma \sqrt{\tau}\right)}{\sigma} \cdot \varphi \left(y + \sigma \sqrt{\tau}\right) \tag{6.34}$$

Figure 6.12 displays the volga given by Eq. (6.34) as a function of stock price and time to maturity.
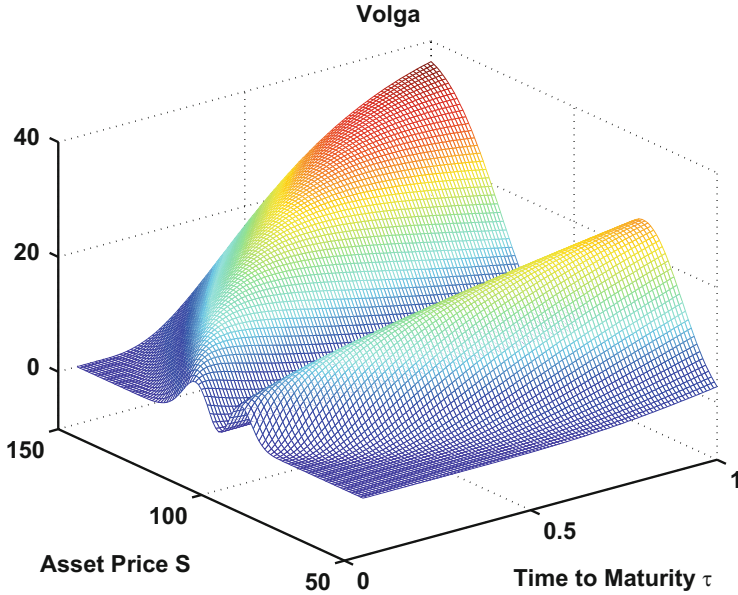
Fig. 6.12 Volga of European Call Option as a function of stock price (*right axis*) and time to maturity (*left axis*) [Q] SFEvolga

If the option approaches at-the-money, volga becomes small, i.e. vega changes slowly. Consequently, the adjustments to keep a portfolio vega neutral need to be made relatively infrequently. On the contrary, if the option approaches in-the-money or out-of-the-money, volga becomes high, i.e. if stock price approaches strike price, the behaviour of vega is unstable.

The sensitivity of vega with respect to the stock price is given by *vanna*:

$$Vanna = \frac{\partial \mathcal{V}}{\partial S} = \frac{\partial^2 C}{\partial \sigma \partial S}.$$

Vanna is derived from the Black–Scholes formula for a call option and takes the following form:

$$Vanna = -\frac{y}{\sigma} \cdot \varphi \left(y + \sigma \sqrt{\tau} \right) \tag{6.35}$$

Figure 6.13 displays the vanna given by Eq. (6.35) as a function of stock price and time to maturity.
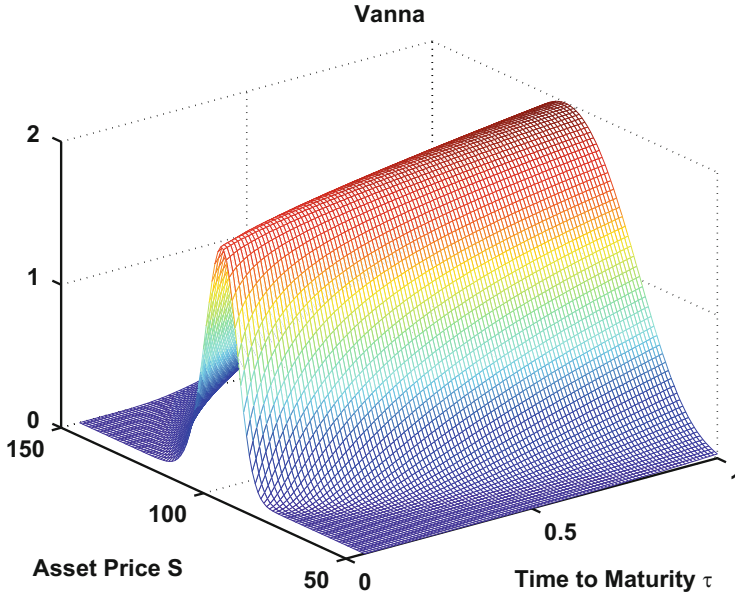
Fig. 6.13 Vanna of European Call Option as a function of stock price (*right axis*) and time to maturity (*left axis*)   Q SFEvanna

6.4.5   Historical and Implied Volatility

A property of the Black–Scholes formulae (6.24), (6.26) is that all option parameters are known, except the volatility parameter σ. In practical applications σ is estimated from available stock price observations or from prices of similar products traded on an exchange.

Historical volatility is an estimator for σ based on the variability of the underlying stock in the past. Let $S_0, \ldots, S_n$ be the stock prices at times $0, \Delta t, 2\Delta t, \ldots, n\Delta t$. If the stock price S_t is modelled as Brownian motion, the logarithmic relative increments

$$R_t = \log \frac{S_t}{S_{t-1}}, \quad t = 1, \ldots, n$$

are independent and identical normally distributed random variables. R_t is the increment $Y_t - Y_{t-1}$ of the logarithmic stock price $Y_t = \log S_t$ which as we saw in Sect. 5.4 is in a small time interval of length Δt a Wiener process with variance σ^2. Consequently the variance of R_t is given by

$$v = \mathsf{Var}(R_t) = \sigma^2 \cdot \Delta t.$$

A good estimator for $\text{Var}(R_t)$ is the sample variance

$$\hat{v} = \frac{1}{n-1} \sum_{t=1}^{n}(R_t - \bar{R}_n)^2$$

with $\bar{R}_n = \frac{1}{n} \sum_{t=1}^{n} R_t$ being the sample average. $\hat{v}$ is unbiased, i.e. $\text{E}[\hat{v}] = v$, and the random variable

$$(n-1)\frac{\hat{v}}{v}$$

is χ^2_{n-1} distributed (chi-square distribution with $n-1$ degrees of freedom). In particular this implies that the mean squared relative estimation error of $\hat{v}$ is given by

$$\text{E}\left(\frac{\hat{v}-v}{v}\right)^2 = \frac{1}{(n-1)^2} \text{Var}\left\{(n-1)\frac{\hat{v}}{v}\right\} = \frac{2}{n-1}.$$

Since it holds $v = \sigma^2 \Delta t$ an estimator for the volatility σ based on historical stock prices is

$$\hat{\sigma} = \sqrt{\hat{v}/\Delta t}.$$

By means of a Taylor expansion of the square root function and by means of the known quantities $\text{E}[\hat{v}]$ and $\text{Var}(\hat{v}/v)$ it follows that $\hat{\sigma}$ is unbiased neglecting terms of size n^{-1} :

$$\text{E}\hat{\sigma} = \sigma + \mathcal{O}\left(\frac{1}{n}\right),$$

and that the mean squared relative estimation error of $\hat{\sigma}$ is given by

$$\text{E}\left(\frac{\hat{\sigma}-\sigma}{\sigma}\right)^2 = \frac{1}{2(n-1)} + \mathcal{O}\left(\frac{1}{n}\right),$$

again neglecting terms of size smaller than n^{-1}. Thanks to this relation the reliability of the estimator $\hat{\sigma}$ can be evaluated.

Sample Parameter Selection

(a) As data daily settlement prices $S_0, \ldots, S_n$ are often used. Since σ is in general expressed as an annualized volatility Δt corresponds to 1 day on a yearly basis. Working with calender day count convention $\Delta t = 1/365$. Unfortunately, no data is available for weekends and holidays. The following empirical argument favours ignoring weekends and holidays: If the stock dynamics behaved on Saturdays and Sundays as they do on trading days, even if the dynamics were not

observed, then standard deviation of the change in the stock price from Friday
to Monday would be three times as large as the standard deviation between
two trading days, say Monday to Tuesday. This follows on from the fact that
for the Wiener process $Y_t = \log S_t$ the standard deviation of the increment
$Y_{t+\delta} - Y_t$ is $\sigma \cdot \delta$. Empirical studies of stock markets show, however, that both
standard deviations are proportional, with a constant of around 1 but in any case
significantly smaller than 3. Put in other words, the volatility decreases at the
weekend. A conclusion is that trading increases volatility, and that the stock
variability is not solely driven by external economic influences. Estimating
volatility should therefore be done by exclusively considering trading days.
Usually a year is supposed to have 252 trading days, i.e. $\Delta t = 1/252$.

Concerning monthly data, $\Delta t = 1/12$ is applied. In Sect. 3.3 we calculated
an annual volatility of 19 % based on the monthly DAX data.   ■ SFEsumm

(b) Theoretically, the larger n the more reliable $\hat{\sigma}$. However, empirically the
volatility is not constant over longer periods of time. That is to say that stock
prices from the recent past contain more information about the current σ as
do stock prices from longer ago. As a compromise the closing prices of the
last 90 days or 180 days are used respectively. Some authors advise the use of
historical data of a period which has the same length as the period in the future
to which the estimated volatility will be applied. In other words, if you want
to compute the value of a call expiring in 9 months you should use the closing
prices of the past 9 months.

The *implied volatility* of an option is computed from its market price observed
on an exchange and not from the prices of the underlying as is case for the historical
volatility. Consider a European call on a non-dividend paying stock ($d = 0$, $b = r$),
which has a quoted market price of C_B, then its implied volatility σ_I is given by
solving

$$S \ \Phi(y + \sigma_I \sqrt{\tau}) - e^{-r\tau} K \ \Phi(y) = C_B$$

with $\qquad y = \dfrac{1}{\sigma_I \sqrt{\tau}} \{\log \dfrac{S}{K} + (r - \dfrac{1}{2}\sigma_I^2)\tau\}.$

σ_I is the value of the volatility which, if substituted into the Black–Scholes
formula (6.24), would give a price equal to the observed market price C_B. σ_I is
implicitly defined as a solution to the above equation and has to be computed
numerically due to the fact that the Black–Scholes formula cannot be inverted.

The implied volatility can be used to get an idea of the market view of the stock
volatility. It is possible to construct an estimator using implied volatilities of options
on the same stock but which are different in time to maturity τ and exercise price
K. A weighting scheme takes the option price dependence on the volatility into
account.   ■ SFEVolSurfPlot

Example 6.10 Consider two traded options on the same underlying. One is at-the-money (ATM) and the other is deep ITM with volatilities of $\sigma_{I1} = 0.25$ respectively $\sigma_{I2} = 0.21$. At-the-money, the dependence of option price and volatility is particularly strong. That is, the price of the first option contains more information about the stock volatility and σ_{I1} can be considered a more reliable volatility estimate. Thus the estimator combining both implied volatilities should attribute a higher weight to σ_{I1}, as for example

$$\tilde{\sigma} = 0.8 \cdot \sigma_{I1} + 0.2 \cdot \sigma_{I2}.$$

Some authors suggest setting $\tilde{\sigma} = \sigma_{Im}$ with σ_{Im} being the volatility of the option which is most sensitive to changes in σ, i.e. the option with the highest vega $\partial C / \partial \sigma$ in absolute terms.

6.4.6   Realised Volatility

Realized volatility is constructed from high-frequency intra-day returns. Most stochastic volatility models do not include high-frequency intra-day data for the calculation of daily volatility and are inadequate in reflecting reality. Realized volatility, on the contrary, takes into consideration high-frequency intraday data for forecasting daily and lower frequency volatility and the distribution of returns. Based on the theory of quadratic variation, realized volatility is given by the sum of all intra-period high-frequency squared returns, period by period.

Denote as above $Y_t = \log S_t$ the logarithmic stock price and

$$R_t = Y(t\Delta) - Y\{(t-1)\Delta\}, \quad t = 1, 2, \ldots, n$$

the returns over an interval of length Δ, e.g. 1 day. Let $\sigma^2(t)$ denote the variance of R_t. We split the time interval $[(t-1)\Delta, t\Delta]$ into M small sub-intervals of length Δ/M each, corresponding to intra-day observations during the day t. Then, the realized volatility $\{Y\}_2^{1/2}$ is modelled via the sum of squared intra-day changes over a day:

$$\{Y\}_t = \sum_{j=1}^{M} \left[Y\left\{(t-1)\Delta + \frac{\Delta j}{M}\right\} - Y\left\{(t-1)\Delta + \frac{\Delta(j-1)}{M}\right\} \right]^2,$$

$\{Y\}_2^{1/2}$ is an estimate of the *actual volatility* $\sigma(t) = \{\mathsf{Var}(R_t)\}^{1/2}$. For example, for a 24-h market, daily realized volatility based on 5 min underlying returns is defined as the sum of 288 intra-day squared 5 min returns, taken day by day.

6.5   Recommended Literature

The classic papers of Black and Scholes (1973) and Merton (1973) which established modern derivatives pricing theory are worth reading. As does this book, Wilmott et al. (1995) present an extensive introduction to mathematics of financial derivatives without martingale theory. Two influential works contributing to modern financial mathematics but which apply more advanced results of the theory of stochastic processes are Harrison and Pliska (1981) and Delbaen and Schachermayer (1994). A discussion of the mathematical foundations of absence of arbitrage is given by Jensen and Nielsen (1996). Korn and Korn (1999) and Korn and Korn (2001) provide a compact introduction to modern financial mathematics. For the advanced mathematician Duffie (1996) and Baxter and Rennie (1996) represent good starts into derivative pricing using martingale theory. Korn (1999) puts the focus on problems arising in hedging and portfolio optimization. Crack (2004) gives a clear explanation of Black–Scholes option pricing theory, by discussing direct applications of the theory to trading, and the differences between the theoretical Black–Scholes world and the real world. Hull (2006) derives the Black–Scholes model for valuing the options on a non-dividend paying stock and shows how the model can be extended to deal with options on dividend paying stocks. The explanation of how volatility can be either estimated from the historical data or implied from option prices is also given by Hull (2006). Dynamics of the implied volatility and its relation to the risk perception has been investigated by Fengler et al. (2003) and Ait-Sahalia and Duarte (2003).

6.6   Exercises

Exercise 6.1

(a) *Simulate the stop-loss strategy hedging of a European call for a given stock price S and for a random stock price movement.*

Default values	
Strike price X	*50*
Time to maturity T	*20 weeks*
Riskfree rate r p.a.	*0.05*
Number of shares	*100,000*

$S = (49, 49.75, 52, 50.1, 48.375, 48.25, 48.75, 49.625, 48.25, 48.25,$
$51.125, 51.5, 49.875, 49.875, 48.75, 47.5, 48, 46.25, 48.12, 46.62, 48.12)$

(b) *Calculate the BS price of the call and the cost of hedging. Create the following table:*

Time	Stock price	Hedge strategy	Shares purchased	Cost of shares	Cumulative cost	int. cost
0	49	0	0	0	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
T						

(c) Modify your quantlet for $S_0 = 51$, i.e. the call is in the money.

Time	Stock price	Hedge strategy	Shares purchased	Cost of shares (10^3)	Cumulative cost(10^3)	int. cost
0	49	0	0	0	0	0
1	49.75	0	0	0	0	0
2	52	1	100,000	5,200	5,200	5
⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	48.12	0	0	0	514.2	

Exercise 6.2 *Check the performance measure of stop-loss strategy for an increasing hedging frequency based on 500 sample paths for the stock price. Create the following table:*

Time between hedge rebalancing (weeks)	5	4	2	1	0.5	0.25
Performance measure						

$$\frac{\text{standard deviation of cost of writing and hedging the option}}{\text{theoretical price of option}}$$

Time between hedge rebalancing	5	4	2	1	0.5	0.25
Performance measure	1.041	0.928	0.902	0.818	0.777	0.769

Q SFSexercise3

Exercise 6.3 *Calculate the value of* $\Delta = \frac{\partial C}{\partial S}$ *(i.e.* Δ *is the ratio of change of the option price with respect to the underlying stock price) for a European call option with strike price K and maturity T.*

Exercise 6.4 *Simulate* Δ *hedging similar to Exercise 6.1.*

Time	Stock price	Δ	Shares purchased	Cost of shares (10^3)	Cumulative cost (10^3)	int. cost
0	49	0.522	52,200	2,557.8	2,557.8	2.5
1	48.125	0.458	(6,400)	(308.0)	2,252.3	2.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮
9	53	0.787	11,300	598.9	4,000.5	3.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	57.25	1.0000	0	0.0	5,263.4	

Q SFSexercise5

Exercise 6.5 (Delta Hedging)

A bank sold 3,000 Calls and 2,500 Puts with the same maturity and exercise price on the same underlying stock. Δ Call is 0.42 and Δ Put is -0.38. How many stocks should the company sell or buy in order to have a delta neutral position?

Exercise 6.6 (Delta and Gamma Hedging)

Mrs. Ying Chen already has a portfolio with $\Delta = -300$ and $\Gamma = 250$. She wants to make her portfolio Δ and Γ neutral. In the market, she can buy/sell stocks and call options to achieve this. She calculated the Δ and Γ of the call option as 0.55 and 1.25, respectively. What should Mrs. Chen do in order to realize her goal?

Exercise 6.7 *A European call option on a stock when there are ex-dividend dates in 3 months and 6 months. The dividend on each ex-dividend date is expected to be 1, the current price is 80, the exercise price is 80, the volatility is 25 % per annum, the annually risk free rate is 7 %, the time to maturity is 1 year, calculate the option price using the Black–Scholes model.*

Exercise 6.8 *An investor longs a 3 months maturity call option with K = 220, and shorts a half year call with K = 220. Currently,* $S_0 = 220$*, risk free rate r = 0.06 (continuously compounded) and* $\sigma = 0.25$*. The stock pays no dividends.*

1. Calculate the Δ of the portfolio.
2. Calculate the Γ of the portfolio.
3. Calculate the Θ of the portfolio.

Exercise 6.9 *Mr Wang constructed a "collar" portfolio, which was established by buying a share of stock for 15 JPY, buying a 1-year put option with exercise price 12.5 JPY, and short selling a 1-year call option with exercise price 17.5 JPY. If the BS model holds, based on the volatility of the stock, Mr Wang calculated that for*

a strike price of 12.5 JPY and maturity of 1 year, $\Phi(y + \sigma \sqrt{\tau}) = 0.63$, *whereas for the exercise price of 17.5 JPY,* $\Phi(y + \sigma \sqrt{\tau}) = 0.32$ *where we use y as a abbreviation for*

$$y = \{\log \frac{S}{K} + (b - \frac{1}{2}\sigma^2)\tau\}/\sigma \sqrt{\tau}. \tag{6.36}$$

1. Draw a payoff graphic of this "collar" at the option expiration date.
2. If the stock price increase by 1 JPY, what will Mr Wang gain or lose from this portfolio?
3. What happens to the delta of the portfolio if the stock price becomes very high or very low?

Exercise 6.10 *A financial institution has the following portfolio of OTC options on a particular stock (Table 6.6):*

A traded call option is available which has a delta of 0.3 and a gamma of 1.8. What position in that traded option and underlying stock would make the portfolio both gamma neutral and delta neutral?

Exercise 6.11 *Knowing that the current price of oil is 100 EUR per barrel, a petrochemical firm PetroCC plans to buy a call option on oil with strike price = 100 EUR. The volatility of oil prices is 10 % per month, and the risk-free rate is 3 % per month.*

1. What is the value of the 4-month call option?
2. Suppose that instead of buying 4-month call options on 100,000 barrels of oil, the firm will synthetically replicate a call position by buying oil directly now, and delta-matching the hypothetical call option. How many barrels of oil should it buy?
3. If oil prices increase by 1 % after the first day of trading, how many barrels of oil should it buy or sell?

Exercise 6.12 *There are two calls on the same stock with the same time to maturity (1 year) but different strike price. Option A has a strike price 10 USD, while option B has a strike price 9.5 USD. The current stock price is $10. The stock does not pay dividend. Risk free rate is 3 %. One applies the Black–Scholes equation as the correct option pricing model. Yet despite that fact that one is confident that the*

Table 6.6 Portfolio on a particular stock

Type	Position	Delta	Gamma
Call	−2000	0.5	2.5
Call	+1500	0.7	0.7
Put	−4000	−0.4	1.1

appropriate volatility of the stock is 0.15 p.a. one observes option A selling for 0.8 USD and option B selling for 1 USD.

1. *Is the implied volatility of option A more or less than 15 %? What about that of option B?*
2. *Determine a delta-neutral position in the two calls that will exploit their apparent mispricing. Use $\sigma = 0.15$ to compute Delta.*

Exercise 6.13 *E-Tech Corp. stock sells for 80 Euro and pays no dividends. A 6-month call option with exercise price 80 Euro is priced at 7.23 Euro, while a 6-month call option with exercise price 90 Euro is priced at 5.38. The risk-free interest rate is 8 % per year.*

1. *What is the implied volatility of these two options?*
2. *Based on the above information, construct a profitable trading strategy.*

Exercise 6.14 *The Black–Scholes price of a call option with strike price K, maturity T is defined as follows in $t \in [0, T)$ at a stock price x:*

$$v(x, T - t) = x\Phi(d_+(x, T - t)) - K\exp\{-r(T - t)\}\Phi(d_-(x, T - t))$$

$r \in \mathbb{R}$ denotes the riskless interest rate and $\sigma > 0$ the volatility of the stock. The functions d_+ and d_- are given by

$$d_\pm(x, r) = \{\log x/K + (r \pm \sigma^2/2)\tau\}/\sigma\sqrt{\tau}$$

Calculate the "Greeks"

1. *$\Delta = \frac{\partial}{\partial x}v(x, T - t)$*
2. *$\Gamma = \frac{\partial^2}{\partial x^2}v(x, T - t)$*
3. *$\Theta = \frac{\partial}{\partial t}v(x, T - t)$*

and verify that v solves the partial differential equation

$$\left(\frac{1}{2}\sigma^2 x^2 \frac{\partial^2}{\partial x^2} + rx\frac{\partial}{\partial x} - \frac{\partial}{\partial t}\right)v(x, t) = rv(x, t) \text{ on } (0, \infty) \times (0, \infty)$$

and satisfies the boundary condition

$$v(x, T - t) \to (x - K)^+ \text{ for } t \to T$$

Exercise 6.15

1. *Show that following holds for $x, y > 0$ with $x \neq t$*

$$|v(x, t) - v(y, t)| < |x - y|$$

but

$$\frac{v(x,t) - v(y,t)}{v(y,t)} > \frac{x-y}{y} \ \textit{for } x > y$$

and respectively

$$\frac{v(x,t) - v(y,t)}{v(y,t)} < \frac{x-y}{y} \ \textit{for } x < y$$

for the special cases
2. *Calculate "Vega" $= \frac{\partial}{\partial\sigma} v(x,t)$*

Show that "Vega" is always positive and calculate the value x where "Vega" is maximal

Exercise 6.16 *In the Black–Scholes model, the stock price path can be modelled by the following stochastic process in $\mathbb{R}^2$: $dS(t) = \mu S(t)dt + \sigma dW(t)$, where μ denotes the drift, $S(t)$ the stock price, σ the constant volatility and W_t standard Brownian motion.*

1. *The market price of risk is defined as excess return. How is this defined in the Black–Scholes framework?*
2. *Give the explicit form of the stochastic process $S(t)$*
3. *Suppose now that a class of parametrized class of equivalent probabilities Q are introduced via the Girsanov transformation:*

$$W^\theta(t) = W(t) - \int_0^t \theta(u)du$$

where θ is a real valued, bounded continuous function. By using the Girsanov Theorem an equivalent probability measure denoted Q^θ exists, so that $W^\theta(t)$ is a Brownian motion for t. Show the dynamics of $S(t)$ under Q^θ.

Chapter 7
Binomial Model for European Options

A large range of options exist for which the boundary conditions of the Black–Scholes differential equation are too complex to solve analytically; an example being the American option. One therefore has to rely on numerical price computation. The best known method for this is to approximate the stock price process by a discrete time stochastic process, or, as in the approach followed by Cox, Ross, Rubinstein, model the stock price process as a discrete time process from the start. By doing this, the options time to maturity T is decomposed into n equidistant time steps of length

$$\Delta t = \frac{T}{n}.$$

We consider therefore the discrete time points

$$t_j = j\Delta t, \ j = 0, \ldots, n.$$

By S_j we denote the stock price at time t_j. At the same time, we discretize the set of values the stock price can take, such that it takes on many finite values $S_j^k, k = 1, \ldots, m_j$, with j denoting the point of time and k representing the value. If the stock price is in time t_j equal to S_j^k, then it can jump in the next time step to one of m_{j+1} new states S_{j+1}^l, $l = 1, \ldots, m_{j+1}$. The probabilities associated to these movements are denoted by p_{kl}^j:

$$p_{kl}^j = P(S_{j+1} = S_{j+1}^l | S_j = S_j^k),$$

with

$$\sum_{l=1}^{m_{j+1}} p_{kl}^j = 1, \quad 0 \le p_{kl}^j \le 1.$$

If we know the stock price at the current time, we can build up a tree of possible prices up to a certain point in time, for example the maturity date $T = t_n$. Such a tree is also called *stock price tree*. Should the option price be known at the final point in time t_n of the stock price tree, for example by means of the options intrinsic value, the option value at time t_{n-1} can be computed (according to (6.24)) as the discounted conditional expectation of the corresponding option prices at time t_n given the stock price at time t_{n-1} :

$$V(S_{n-1}^k, t_{n-1}) = e^{-r\Delta t} \mathsf{E}\{V(S_n, t_n) \mid S_{n-1} = S_{n-1}^k\}$$

$$= e^{-r\Delta t} \sum_{l=1}^{m_n} p_{kl}^{n-1} V(S_n^l, t_n). \tag{7.1}$$

$V(S, t)$ again denotes the option value at time t if the underlying has a price of S. Repeating this step for the remaining time steps t_j, $j = n-2, n-3, \ldots, 0$, means that the option prices up to time $t = 0$ can be approximated.

7.1 Cox–Ross–Rubinstein Approach to Option Pricing

As the simplest example to price an option, we consider the approach by Cox, Ross and Rubinstein (CRR) which is based on the assumption of a *binomial model*, and which can be interpreted as a numerical method to solve the Black–Scholes equation. We will look at European options exclusively and assume, for the time being, that the underlying pays no dividends within the time to maturity. Again, we discretize time and solely consider the points in time $t_0 = 0, t_1 = \Delta t, t_2 = 2\Delta t, \ldots, t_n = n\Delta t = T$ with $\Delta t = \frac{T}{n}$. The binomial model proceeds from the assumption that the discrete time stock price process S_j follows a geometric random walk (see Chap. 4), which is the discrete analogue of the geometric Brownian motion. The binomial model has the special feature that at any point in time the stock price has only two possibilities to move:

- either the price moves at rate u and with probability p in one direction (for example it moves up)
- or the price moves at rate d and with probability $1 - p$ in another direction (for example it moves $d$own).

Using the notation introduced above, if the stock price in time t_j is equal to S_j^k then in time t_{j+1} it can take only the values $u \cdot S_j^k$ and $d \cdot S_j^k$. The probabilities p and q are independent of j. All other probabilities p_{kl}^j associated to $S_{j+1}^l \neq u \cdot S_j^k$ and $\neq d \cdot S_j^k$ are 0.

In order to approximate the Black–Scholes differential equation by means of the CRR approach, the probabilities p, q as well as the rates u, d have to be chosen such that in the limit $\Delta t \to 0$ the binomial model converges to a geometric Brownian motion. That is, arguing as in (6.22) the conditional distribution of $\log(S_{j+1})$

given S_j must be asymptotically a normal distribution with expectation parameter $\log(S_j) + (b - \frac{1}{2}\sigma^2)\Delta t$ and variance parameter $\sigma^2 \Delta t$. However, the conditional distribution of $\log(S_{j+1})$ given S_j implied by the binomial model is determined by $\log(u \cdot S_j), \log(d \cdot S_j)$ and their associated probabilities p and q. We set the parameters of the geometric random walk such that the conditional expectations and variances implied by the binomial model are equal to their asymptotic values for $\Delta t \to 0$. Taking into account that $p + q = 1$ we obtain three equations for the four unknown variables p, q, u and d:

$$p + q = 1,$$

$$E \stackrel{\text{def}}{=} p \log(u \cdot S_j) + q \log(d \cdot S_j) = \log(S_j) + \left(b - \frac{1}{2}\sigma^2\right)\Delta t,$$

$$p\{\log(u \cdot S_j) - E\}^2 + q\{\log(d \cdot S_j) - E\}^2 = \sigma^2 \Delta t.$$

Due to the first equation, the current stock price S_j disappears from the remaining equations. By substituting $q = 1 - p$ into the latter two equations, we obtain, after some rearrangements, two equations and three unknown variables:

$$p \log\left(\frac{u}{d}\right) + \log d = \left(b - \frac{1}{2}\sigma^2\right)\Delta t,$$

$$(1 - p)p \left\{\log\left(\frac{u}{d}\right)\right\}^2 = \sigma^2 \Delta t.$$

To solve this nonlinear system of equations we introduce a further condition

$$u \cdot d = 1,$$

i.e. if the stock price moves up and subsequently down, or down and subsequently up, then it takes its initial value two steps later. This recombining feature is more than only intuitively appealing. It simplifies the price tree significantly. At time t_j there are only $m_j = j + 1$ possible values the stock price S_j can assume. More precisely, given the starting value S_0 at time t_0 the set of possible prices at time t_j is

$$S_j^k = S_0 u^k d^{j-k}, \ k = 0, \dots, j,$$

because it holds $S_{j+1}^{k+1} = u \cdot S_j^k$ and $S_{j+1}^k = S_j^k / u$. In the general case there would be $m_j = 2^j$ possible states since then not only the number of up and down movements would determine the final state but also the order of the up and down movements.

Solving the system of three equations for p, u, d and neglecting terms being small compared to Δt it holds approximatively:

$$p = \frac{1}{2} + \frac{1}{2}(b - \frac{1}{2}\sigma^2)\frac{\sqrt{\Delta t}}{\sigma}, \quad u = e^{\sigma\sqrt{\Delta t}}, \quad d = \frac{1}{u}. \tag{7.2}$$

For the option price at time t_j and a stock price $S_j = S_j^k$ we use the abbreviation $V_j^k = V(S_j^k, t_j)$. As in Eq. (7.1) we obtain the option price at time t_j by discounting the conditional expectation of the option price at time t_{j+1} :

$$V_j^k = e^{-r\Delta t}\{pV_{j+1}^{k+1} + (1-p)V_{j+1}^k\}. \tag{7.3}$$

At maturity $T = t_n$ the option price is known. In case of a European option we have

$$V_n^k = \max\{0, S_n^k - K\}, \ k = 0, \ldots, n. \tag{7.4}$$

Beginning with Eq. (7.1) and applying Eq. (7.3) recursively all option values $V_j^k, \ k = 0, \ldots, j, \ j = n-1, n-2, \ldots, 0$ can be determined.

Example 7.1 An example of a call option is given in Table 7.1. First the tree of stock prices is computed. Since $\Delta t = \tau/n = 0.1$ it follows from Eq. (7.2) that $u = 1.0823$. Given the current stock price $S_0 = 230$ the stock can either increase to $S_1^1 = uS_0 = 248.92$ or decrease to $S_1^0 = S_0/u = 212.52$ after the first time step. After the second time step, proceeding from state $S_1 = S_1^1$ the stock price can take the values $S_2^2 = uS_1 = 269.40$ or $S_2^1 = S_1/u = 230$, proceeding from $S_1 = S_1^0$ it

Table 7.1 Evolution of option prices (no dividend paying underlying)

Q SFEbitreeCDiv

Current stock price S_t	230.00
Exercise price K	210.00
Time to maturity τ	0.50
Volatility σ	0.25
Risk free rate r	0.04545
Dividend	None
Time steps	5
Option type	European call

Stock prices	Option prices					
341.50558						131.506
315.54682					106.497	
291.56126				83.457		81.561
269.39890			62.237		60.349	
248.92117		44.328		40.818		38.921
230.00000	30.378		26.175		20.951	
212.51708		16.200		11.238		2.517
196.36309			6.010		1.275	
181.43700				0.646		0.000
167.64549					0.000	
154.90230						0.000
Time	0.00	0.10	0.20	0.30	0.40	0.50

Table 7.2 Convergence of the price implied by the binomial model to the price implied by the Black–Scholes formula

Time steps	5	10	20	50	100	150	Black–Scholes
Option value	30.378	30.817	30.724	30.751	30.769	30.740	30.741

can move to $S_2^1 = 230$ or $S_2^0 = 196.36$ and so on. At maturity, after 5 time steps, the stock price S_5 can take the following six values $S_5^5 = u^5 S_0 = 341.51$, $S_5^4 = u^3 S_0 = 291.56, \ldots, S_5^0 = S_0/u^5 = 154.90$.

Next, given the tree of stock prices, we compute the option price at maturity applying Eq. (7.4), for example $V_5^4 = V(S_5^4, t_5) = S_5^4 - K = 81.561$ or $V_5^1 = 0$, since $S_5^1 = 181.44 < K$. Equation (7.2) implies $p = 0.50898$, since the cost of carry b is equal to the risk free interest rate r when no dividends are paid. Proceeding from the option's intrinsic values at maturity we compute recursively the option values at preceding points of time by means of Eq. (7.3). With $V_5^4 = 81.561$, $V_5^3 = 38.921$ we obtain the option value $V_4^3 = 60.349$ at time $t_4 = 0.4$ corresponding to a stock price $S_4 = S_4^3 = 269.40$ by substituting the known values of $p, r, \Delta t$. Analogously we obtain the option value $V_0^0 = 30.378$ at time $t_0 = 0$ and current stock price $S_0 = 230$ by means of Eq. (7.3) and the time $t_1 = 0.1$ option values $V_1^1 = 44.328$, $V_1^0 = 16.200$.

Using only 5 time steps 30.378 is just a rough approximation to the theoretical call value. However, comparing prices implied by the Black–Scholes formula (6.24) to prices implied by the CRR approach for different time steps n the convergence of the numerical binomial model solution to the Black–Scholes solution for increasing n is evident (see Table 7.2).

The numerical procedure to price an option described above does not change if the underlying pays a continuous dividend at rate d. It is sufficient to set $b = r - d$ instead of $b = r$ for the cost of carry. Dividends paid at discrete points of time, however, require substantial modifications in the recursive option price computation; these will be discussed in the following section.

Example 7.2 We consider a call on the US-Dollar with a time to maturity of 4 months, i.e. $\tau = 1/3$ years, a current exchange rate of $S = 1.50$ EUR/USD and an exercise price $K = 1.50$ EUR/USD. The continuous dividend yield, which corresponds to the US interest rate, is assumed to be 1 %, and the domestic interest rate is 9 %. It follows that the cost of carry being the difference between the domestic and the foreign interest rate is equal to $b = r - d = 8$ %. Table 7.3 gives as in the previous example the option prices implied by the binomial model.

7.2   Discrete Dividends

If dividend payments are made at discrete points in time, the tree of stock prices changes. By changing the price tree we have to distinguish two different cases. In the first case, dividends are paid as a percentage of the stock price. In the second case,

Table 7.3 Evolution of option prices (with continuous dividends)
Q SFEbitreeCDiv

Current EUR/USD-price S_t	1.50
Exercise price K	1.50
Time to maturity τ	0.33
Volatility σ	0.20
Risk free interest rate r	0.09
Continuous dividend d	0.01
Time steps	6
Option type	European call

Price	Option prices						
1.99034							0.490
1.89869						0.405	
1.81127					0.324		0.311
1.72786				0.247		0.234	
1.64830			0.180		0.161		0.148
1.57240		0.127		0.105		0.079	
1.50000	0.087		0.067		0.042		0.000
1.43093		0.041		0.022		0.000	
1.36504			0.012		0.000		0.000
1.30219				0.000		0.000	
1.24223					0.000		0.000
1.18503						0.000	
1.13046							0.000
Time	0.00	0.06	0.11	0.17	0.22	0.28	0.33

dividends are paid as a fixed amount of money. We confine ourselves to the case that dividends are paid only once during the time to maturity, say, at time $t^*, 0 < t^* \leq T$. Dividends paid at several points in time can be dealt with analogously. We assume that the underlying is a stock.

Using no arbitrage arguments it can be shown that the stock price jumps down by the amount of the dividend at the time the dividend is paid. Let us consider the following argument to visualize this. At time $t^* - dt$, immediately before the dividend payment, we buy the stock, cash in the dividend and sell the stock at time $t^* + dt$. By doing this, we make a gain of $D + S_{t^*+dt} - S_{t^*-dt}$, which for $dt \rightarrow 0$ would be without risk and, therefore, has to be zero if arbitrage is excluded. This is guaranteed if S_t jumps down by D at time t^*.

Table 7.4 Evolution of option prices (dividends as a percentage of the stock price) SFEbitreePDiv

Current stock price S_t	230.00
Exercise price K	210.00
Time to maturity τ	0.50
Volatility σ	0.25
Risk free interest rate r	0.04545
Discrete dividend δ	0.01
Dividend date t^*	0.15
Time steps	5
Option type	European call

Stock prices	Option prices					
341.50558						128.091
315.54682					103.341	(338.09)
291.56126				80.542	(312.39)	78.646
269.39890			59.543	(288.65)	57.655	(288.65)
248.92117		41.942	(266.70)	38.329	(266.70)	36.432
230.00000	28.384	(248.92)	24.087	(246.43)	18.651	(246.43)
212.51708	(230.00)	14.592	(227.70)	9.547	(227.70)	0.392
196.36309		(212.52)	4.886	(210.39)	0.199	(210.39)
181.43700			(194.40)	0.101	(194.40)	0.000
167.64549				(179.62)	0.000	(179.62)
154.90230					(165.97)	0.000
						(153.35)
Time	0.00	0.10	0.20	0.30	0.40	0.50
Dividend	1.00	1.00	0.99	0.99	0.99	0.99

7.2.1 Dividends as a Percentage of the Stock Price

Suppose that t^* is contained, say, in the ith time interval, i.e. $t_{i-1} < t^* \leq t_i$. Let the dividend paid at time t_i be a percentage δ of the stock price, that is the dividend amount that is paid is equal to δS_i. It follows that the stock price at time t_i is smaller by the dividend amount than the stock price without the dividend payment. Accordingly, all stock prices in the tree after time t_i change in the same way: all prices S_j^k, $j \geq i$, are multiplied by the factor $(1 - \delta)$. Following this correction the option values can be determined recursively as in the no dividend case.

Example 7.3 We consider a call option on a stock paying a dividend of $\delta = 1\%$ of the stock price at time 0.15. All other parameters of this example are those already given in Table 7.1. The results are shown in Table 7.4. First we ignore the dividend and compute the stock price tree as shown in Table 7.1. Following, all stock prices from the dividend date on, i.e. from time $t_2 = 0.2$ on (note that we have divided the

time period into 5 time steps $0 \leq t \leq 0.5$), are multiplied by the factor $(1 - \delta)$. In Table 7.4 the values in parentheses correspond to the stock prices that are decreased by the dividend amount, i.e. S_j^k, $j < i = 2$, respectively $0.99 \cdot S_j^k$, $j \geq i = 2$. Thus, the option prices at maturity change due to Eq. (7.4), for example $V_5^4 = V(0.99 \cdot S_5^4, t_5) = 0.99 \cdot 291.56 - K = 78.646$. Having determined the option values at maturity the preceding option values are again computed by recursively applying Eq. (7.2). Note, V_j^k corresponds to the stock price $0.99 \cdot S_j^k$ rather than to S_j^k, for $j \geq 2$, i.e. $t_j \geq t^*$. However, the current time $t_0 = 0 < t^* = 0.15$ is not concerned, i.e. $V_0^0 = 28.384$ is still the option price corresponding to the current stock price $S_0 = 230$.

7.2.2 Dividends as a Fixed Amount of Money

We assume now that at an ex ante fixed point in time t^* a fixed amount of money (for example 5.00 EUR) is paid. Now, the stock price jumps down by an amount which is independent of the stock price. It follows that the tree is not totally recombining anymore. The stock price tree splits up which can be visualized in a simple example. Suppose that at time t^*, $t_1 < t^* \leq t_2 < T$, a fixed dividend of D is paid. Figure 7.1 shows the stock price tree for this example. Before the dividend payment at time t_1 the nodes correspond to stock prices of the kind uS_0 and S_0/u. After the dividend payment, however, stock prices at time t_2 are given by $u^2 S_0 - D$, $S_0 - D$ and $S_0/u^2 - D$. Proceeding from these three prices the tree consists of six possible prices in time t_3, at time t_4 it consists of nine and so on. The stock price tree gets very vast the more time steps are considered, and is less useful for practical computations. To overcome this problem, we use the fact that the dividend is independent of the stock price and therefore not random anymore. We decompose the stock price S_j in a random and a deterministic component:

$$S_j = \tilde{S}_j + D_j,$$
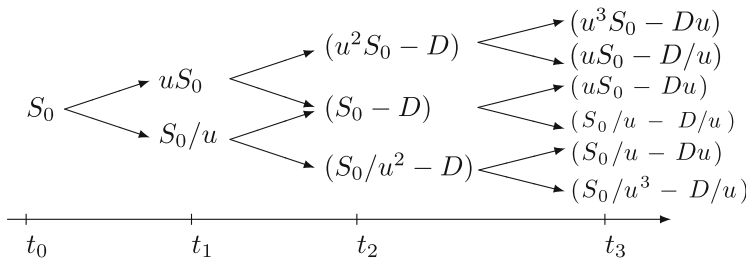
Fig. 7.1 Evolution of the stock price tree (dividends as a fixed amount of money)

Table 7.5 Evolution of option prices (discrete dividends as a fixed money amount)   🔍 SFEbitreeFDiv

Current stock price S_t	230.00
Exercise price K	210.00
Time to maturity τ	0.50
Volatility σ	0.25
Risk free interest rate r	0.04545
Discrete dividend δ	0.01
Dividend date t^*	0.15
Time steps	5
Option type	European put

Prices	Option prices					
339.51						129.507
314.55					104.506	(339.51)
289.56				81.475	(314.55)	79.561
268.40			60.264	(290.56)	58.358	(289.56)
246.92		42.478	(269.40)	38.836	(268.40)	36.921
229.00	28.768	(248.92)	24.435	(247.92)	18.960	(246.92)
210.52	(230)	14.823	(230.00)	9.734	(229.00)	0.516
195.36		(212.52)	4.997	(211.52)	0.262	(210.52)
179.44			(196.36)	0.133	(195.36)	0.000
166.65				(180.44)	0.000	(179.44)
152.90					(166.65)	0.000
						(152.90)
Time	0.00	0.10	0.20	0.30	0.40	0.50
Dividend	1.00	1.00	0.99	0.99	0.99	0.99

with D_j being the current present value of the dividend payment, i.e. before dividend payment, it is the time $t_j \le t^*$ discounted value of D, afterwards it is 0 :

$$D_j = \begin{cases} De^{-r(t^* - t_j)} & , \text{for } t_j \le t^*, \\ 0 & , \text{for } t^* < t_j. \end{cases} \tag{7.5}$$

In particular, at maturity it holds $D_n = 0$ and $S_n = \tilde{S}_n$.

In order to compute the option price we first construct a stock price tree of the random stock price component $\tilde{S}_n$ beginning in $\tilde{S}_0 = S_0 - D_0$. Starting at maturity $T = t_n$ we obtain:

$$\tilde{V}_{n-1} = e^{-r\Delta t} \mathsf{E}[\max(0, \tilde{S}_n - K)|\tilde{S}_{n-1}]$$

The other option prices are given as in the no dividend case by:

$$\tilde{V}_{j-1} = e^{-r\Delta t}\mathsf{E}[\tilde{V}_j|\tilde{S}_{j-1}].$$

The original option prices then correspond to $\tilde{V}_j^k$ given above. However, they do not correspond to the stock price $\tilde{S}_j^k$, rather than to the actual stock price

$$S_j^k = \tilde{S}_j^k + D_j.$$

Example 7.4 In this example, there are two dividend payments at time $t_1^* = 0.25$ and $t_2^* = 0.75$. Both dividends are $D^{(1)} = D^{(2)} = 1.00$ EUR. The parameters and results are given in Table 7.5. First, we compute the time t_j present value of all dividends with Eq. (7.5): $D_j = D^{(1)}e^{-r(t_1^*-t_j)} + D^{(2)}e^{-r(t_2^*-t_j)}$ for $t_j \leq t_1^*$, $D_j = D^{(2)}e^{-r(t_2^*-t_j)}$ for $t_1^* < t_j \leq t_2^*$ and $D_j = 0$ for $t_2^* < t_j$.

In particular, it holds that $D_j = 0$ for $t_j > t_2^*$. Below, we construct the stock price tree as in Table 7.1, but this time we start in $\tilde{S}_0 = S_0 - D_0 = 229.00$ rather than in $S_0 = 230$. Proceeding from the boundary values $\tilde{V}_5^k = K - \tilde{S}_n^k, k = 0, \ldots, 3, \tilde{V}_5^k = 0, k = 4, \ldots, 5$ we compute once again recursively the put prices at earlier points in time by means of Eq. (7.3). We have to take into account that, for example, the option price $\tilde{V}_3^2 = 38.836$ belongs to the stock price $S_3^2 = \tilde{S}_3^2 + D_3 = 247.92$ and not to $\tilde{S}_3^2 = 246.92$, which accounts for the dividend. It follows that the put option price at a current stock price $S_0 = 230$ is equal to $\tilde{V}_0^0 = 28.768$.

7.3 Recommended Literature

The starting point to price options by means of binomial processes is the classic work of Cox et al. (1979) who introduce this approach as an independent method rather than only as a numeric approximation to the Black–Scholes equations. Baxter and Rennie (1996) provide a detailed and modern description of option pricing with binomial trees. The numerical aspects are extensively discussed by Dewynne et al. (1993).

7.4 Exercises

Exercise 7.1 *Assume a call option with exercise price $K = 8$ at $T = 2$. We are now at $t = 0$. The current price of the stock is* 10. *For the first period, the stock market is expected to be very variable, and the underlying stock's price is expected to increase or decrease by* 20 %. *For the second year, a more stable market is expected and the stock price is expected to increase or decrease by* 10 % . *Assume that the risk-free*

rate is zero. What is the value of this call option? The intrinsic value at T is denoted as C_T.

Exercise 7.2 *Assume that a stock's daily returns exhibit a trinomial process. With equal probabilities (p=1/3), the stock's price either increases 3 or 2 % , or it decreases 4 % each day. What can be said about the price of this stock at the end of the year, assuming a $T = 260$ work days and an initial stock price of $X_0 = 100$?*

Exercise 7.3 *A European put option with a maturity of 1 year and a strike price of 120 EUR is written on a non-dividend-paying stock. Assume the current stock price S_0 is 120 EUR, the volatility σ is 25 % per year and the risk-free rate r is 6 % per year. Use a two-period binomial tree to value the option.*

1. *Construct an appropriate two-period pricing tree and check whether early exercise is optimal.*
2. *Describe the replicating portfolio at each node. Verify that the associated trading strategy is self-financing, and that it replicates the payoff.*

Exercise 7.4 *Show that the payoff of a call option cannot be replicated by stock and bond in a one period trinomial model. Assume zero interest rate for simplicity.*

Exercise 7.5 *Find the hedging strategy for a call option in a one period trinomial model such that the quadratic hedging error is minimal. The quadratic hedging error is understood as the square distance between the actual payoff of the option and the final value of the hedge portfolio. Assume zero interest rate for simplicity.*

Exercise 7.6 *Consider the one period trinomial model with the price $S_0 = 100$ at time $t = 0$. At time $t = 1$ the three possible stock prices are $S^u = 120$, $S^m = 100$ and $S^d = 80$. Find the risk neutral probabilities q^u, q^m, q^d of the up, middle and down movements such that the price of the call option with strike K is equal to the price of the hedging portfolio minimizing the quadratic hedging error (see Exercise 7.5) at time $t = 0$. Consider the cases $K_1 = 110$, $K_2 = 100$ and $K_3 = 90$. Assume zero interest rate for simplicity.*

Exercise 7.7 *Construct a three-step implied binomial tree for stock prices, transition probabilities and Arrow–Debreu prices using the Derman–Kani algorithm. Assume the current value of the underlying stock $S = 100$, with no dividend, time to maturity $T = 1$ year and the annually compounded riskless interest rate $r = 3$ % per year for all time expirations. In contrast to CRR binomial tree we use a nonconstant function for the implied volatility, let us assume following convex function of moneyness, defined as $\log(K/S_t)$:*

$$\hat{\sigma}(K, S_t) = -0.2/[\{\log(K/S_t)\}^2 + 1] + 0.3 .$$

Exercise 7.8 *Consider the call option from Exercise 7.7 and construct the IBTs using the method of Barle–Cakici (BC). Assume an exercise price $K = 100$ EUR/USD and compute the option price implied by the binomial tree. Make a plot of the Arrow–Debreu prices.*

Chapter 8
American Options

8.1   Arbitrage Relations for American Options

Pricing American options is a more complex task than for European options since they can be exercised any time up to expiry. The moment the holder chooses to exercise option depends on the spot price of the underlying asset S_t. In this sense the exercising time is a random variable itself. The Black–Scholes differential equation continues to hold as long as the options are not exercised. However, the boundary conditions are so complicated that an analytic solution is not possible. In this section we study American options in more detail while the numerical procedures of pricing will be discussed in the next section.

As shown in Sect. 2.1, the right to exercise early implies that the value of an American option can never drop below its intrinsic value. For example the value of an American put with exercise price K should not fall below $\max(K - S_t, 0)$. In general this condition does not hold for European options. As an immediate consequence, American puts would be exercised immediately if the value of the option drops below the intrinsic value before expiry.

To show this property, let us consider an American put on a stock with expiry date T. If the stock price S_{t^*} at time t^* is zero, then $S_t = 0$ holds for $t \geq t^*$ since the price process follows a geometric Brownian motion. It is then not worth waiting for a later exercise any more. If the put holder waits, he will lose the interest revenue on the value K that can be received from a bond investment for example. Thus, if $S_{t^*} = 0$, the value of the put at t^* is K which is the same as the intrinsic value. However, in this situation the respective European put cannot be exercised early, and we can only get K on expiry. If we discount this amount to time t^*, we only get $Ke^{-r\tau^*}$ with $\tau^* = T - t^*$, and that is the value of the European put at time t^*. Obviously, it is smaller than the value of an American put and, therefore, also smaller than its intrinsic value. Figure 8.1 shows the put value with a continuous cost of carry b.
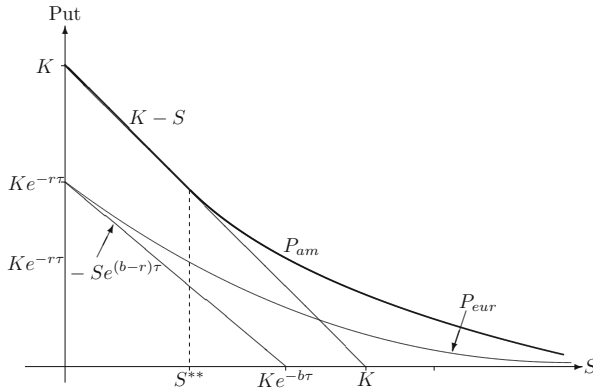
Fig. 8.1 European put and early exercise of an American put with a continuous costs of carry b

We remark that an early exercise of the put may be necessary even before the stock price drops to $S_t = 0$. Heuristically, for any given time t there is a critical stock price S^{**}, such that for $S_t \leq S^{**}$ the loss of interest on the intrinsic value of the option, which the holder can receive by exercising it immediately, is higher than the possible increase of the option value which may be expected up to expiry due to fluctuations of the stock price. Obviously, this critical price depends on time: $S^{**} = S^{**}(t)$, as it is determined by the interest revenue on K and by likely changes of the price of the underlying between time t and expiry T.

The derivation of the Black–Scholes differential equation holds as long as the option is not exercised. Taking into account that there are no transaction costs in perfect markets, the revenue, which can be realized from an early exercise, and, therefore, the value of the option coincides with the intrinsic value. One says that the option falls back to its intrinsic value by early exercising. These considerations imply that pricing American options leads to a free boundary value problem where the Black–Scholes differential equation is valid as long as the price S of the underlying is higher than the critical put-price $S^{**} = S^{**}(t)$ resp. lower than the critical call-price $S^* = S^*(t)$. The boundaries of this region determined by $S^{**}(t)$ resp. $S^*(t)$ are unknown.

Figure 8.2 shows the region where the option price $C = C(S, t)$ for an American call satisfies the Black–Scholes differential equation.

- In the interior $\{(S,t)|0 \leq S < S^*(t), t < T\}$ the Black–Scholes differential equation holds.
- At the boundaries $\{(S,t)|S = S^*(t), t < T\}$ and $\{(S,t)|0 \leq S, t = T\}$ the call falls back to the intrinsic value $\max(S - K, 0)$.
- $C(S, t)$ and $\dfrac{\partial C(S, t)}{\partial S}$ are continuous in the whole region including the boundaries.
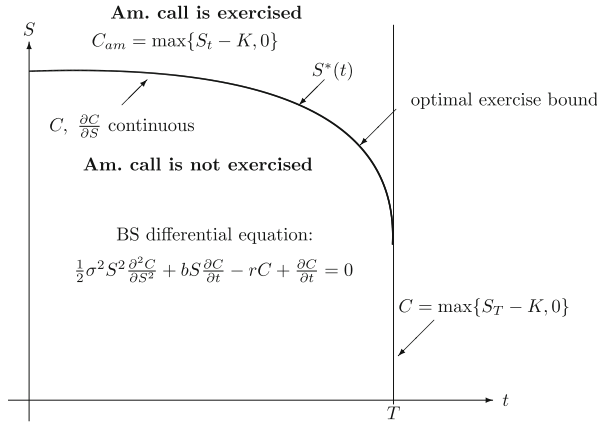
Fig. 8.2 The exercise boundary $S^*(t)$ for an American call

Numerical solutions for such free boundary problems are described in the next section. First, we derive some properties of American options which follow without considering a specific mathematical model for the price process S_t, using only the assumption of a perfect market, in particular the absence of arbitrage opportunities.

Theorem 8.1

1. *An American call on an asset that does not generate any positive income before maturity is not exercised early and, therefore, has the same value as the corresponding European call.*
2. *For an American call on an asset that generates positive income at discrete time points $t_1, \cdots, t_n$ the optimal exercise time can only lie just before one of those points. Consequently, in the case of continuous positive payments, any time point may be an optimal exercise time.*

Proof Let K denote the exercise price, T the expiry date, $\tau = T - t$ the time to maturity of the call and S_t the price of the underlying asset. $C_{am}(S, \tau)$ and $C_{eur}(S, \tau)$ denote the value of the respective American and European calls at time t with time to maturity $\tau = T - t$ and current spot price $S_t = S$. Using the put–call parity for European options we get:

1. In the case of discrete dividends D having the discounted value D_t at time t, they should be costs, i.e. $D_t \leq 0$, by the no gain assumption. Therefore, using the abbreviation $C = C_{eur}$, we get from Theorem 2.3:

$$C(S_t, \tau) = P(S_t, \tau) + S_t - D_t - Ke^{-r\tau} \geq S_t - Ke^{-r\tau} > S_t - K, \qquad (8.1)$$

where P denotes the respective put price. In case of continuous dividends with rate d, we again have by assumption $d \leq 0$. Then, it follows from $b - r = -d \geq 0$ that:

$$C(S_t, \tau) = P(S_t, \tau) + S_t e^{(b-r)\tau} - Ke^{-r\tau}$$

$$\geq S_t e^{(b-r)\tau} - Ke^{-r\tau} > S_t - K \tag{8.2}$$

for $\tau > 0$, i.e. $t < T$. In both cases we verify that $C(S_t, \tau) > S_t - K$ for European calls. Since $C_{am} \geq C$, we conclude that:

$$C_{am}(S_t, \tau) > S_t - K, \quad t < T,$$

i.e. the value of an American call is always higher than the intrinsic value until maturity. Therefore, early exercise is avoided.

2. Without loss of generality, we consider the case where t_1 is the next payment time. $\tilde{t} < t_1$ represents any earlier time. Let $\tilde{C}(S_t, \tilde{\tau})$ with $\tilde{\tau} = \tilde{t} - t$ be the value of a European call with the same exercise price K but with a different maturity at $\tilde{t}$. Since there are no payments before $\tilde{t}$ at all, it follows from part 1 that $\tilde{C}(S_t, \tilde{\tau}) > S_t - K$ for $t < \tilde{t}$. Due to the longer time to maturity and the possibility of early exercise of American calls, it follows that:

$$C_{am}(S_t, \tau) \geq \tilde{C}(S_t, \tilde{\tau}) > S_t - K \tag{8.3}$$

As in Part I, the value of an American call at any time $t < \tilde{t}$ lies strictly above the intrinsic value, which precludes an early exercise. Since $\tilde{t} < t_1$ can take any value, $C_{am}(S_t, \tau)$ can fall to the intrinsic value only at time t_1 (or, respectively, at a later time of payment t_j, $j > 1$). □

Figure 8.3 shows a graphical representation of the first part of the theorem.

- If $b \geq r$ which is equivalent to $d \leq 0$, then $C_{am} = C_{eur}$.
- If $b < r$ which is equivalent to $d > 0$, then $C_{am} > C_{eur}$.

It is also possible to derive a formula similar to the put–call parity for American options. Given that, without a specific model, the critical prices $S^*(t)$, $S^{**}(t)$ and consequently the time point for early exercise are largely unknown, we have to be satisfied with just an inequality.

Theorem 8.2 (Put–Call Parity for American Options) *We consider an American call and an American put with the same maturity date T and the same exercise price K on the same underlying asset. Let $C_{am}(S, \tau)$ and $P_{am}(S, \tau)$ denote the option prices at time t when the spot price is $S_t = S$ and the time to maturity is $\tau = T - t$. Then,*
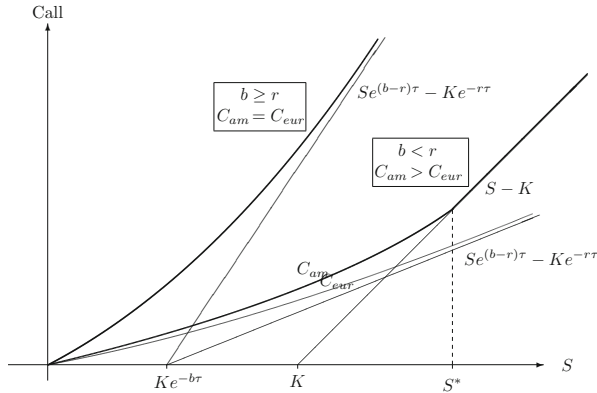
Fig. 8.3 A European call and an early exercised American Call

1. *if there are yields or costs during the time to maturity $\tau = T - t$ having discounted value D_t at time t, then*

$$P_{am}(S_t, \tau) + S_t - Ke^{-r\tau} \geq C_{am}(S_t, \tau) \geq P_{am}(S_t, \tau) + S_t - D_t - K \qquad (8.4)$$

2. *if there are continuous costs of carry with rate b on the underlying asset, then*

$$P_{am}(S_t, \tau) + S_t - Ke^{-r\tau} \geq C_{am}(S_t, \tau)$$
$$\geq P_{am}(S_t, \tau) + S_t e^{(b-r)\tau} - K$$
$$\text{if } b < r \qquad (8.5)$$
$$P_{am}(S_t, \tau) + S_t e^{(b-r)\tau} - Ke^{-r\tau} \geq C_{am}(S_t, \tau)$$
$$\geq P_{am}(S_t) + S_t - K$$
$$\text{if } b \geq r$$

Proof We suppose that, without loss of generality, the underlying asset is a stock paying dividends D_1 at time t_1 during time to maturity.

1. First, we show the left inequality. We consider a portfolio consisting of the following four positions:

   1. buy an American put
   2. buy a stock
   3. sell bonds (i.e. borrow money) with the nominal value K and the maturity date T
   4. sell an American call

Table 8.1 Portfolio value at time $\tilde{t}$, t_1 and T SFEamerican

Position	$\tilde{t} = t_1 - \Delta t$ early exercise of the call	t_1	T Call is exercised early $S_T \leq K$	$S_T > K$	Call is not exercised early $S_T \leq K$	$S_T > K$
1.	≥ 0	−	$K - S_T$	0	$K - S_T$	0
2.	$S_{\tilde{t}}$	D_1	−	−	S_T	S_T
3.	$-Ke^{-r(T-\tilde{t})}$	−	$-K$	$-K$	$-K$	$-K$
4.	$-(S_{\tilde{t}} - K)$	−	$Ke^{r(T-\tilde{t})}$	$Ke^{r(T-\tilde{t})}$	0	$-(S_T - K)$
Sum	≥ 0	≥ 0	≥ 0	≥ 0	0	0

Table 8.2 Portfolio value at time $\tilde{t}$, t_1 and T

Pos.	$\tilde{t} = t_1 - \Delta t$ early exercise of a put	T Put is exercised early $S_T \leq K$	$S_T > K$	Put is not exercised early $S_T \leq K$	$S_T > K$
1.	≥ 0	0	$S_T - K$	0	$S_T - K$
2.	$-S_{\tilde{t}} - D_t e^{r(\tilde{t}-t)}$	$-S_T$ $-D_t e^{r\tau}$	$-S_T$ $-D_t e^{r\tau}$	$-S_T$ $-D_t e^{r\tau}$	$-S_T$ $-D_t e^{r\tau}$
3.	$(D_t + K)e^{r(\tilde{t}-t)}$	$(D_t + K)e^{r\tau}$	$(D_t + K)e^{r\tau}$	$(D_t + K)e^{r\tau}$	$(D_t + K)e^{r\tau}$
4.	$-(K - S_{\tilde{t}})$	0	0	$-(K - S_T)$	0
Sum	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0

In this portfolio, position 1 is held until time T even if this should be suboptimal, i.e. the put is not exercised early, even when the call holder exercises his option early. Note from Theorem 8.1 that an early exercise of the call is only possible at time $\tilde{t} = t_1 - \Delta t$ where $\Delta t \approx 0$, i.e. directly before the dividend payment at time t_1. In this case we deliver the stock of the portfolio. The value of the portfolio at time T is given in the Table 8.1.

Therefore, we have for the portfolio value for every time t as claimed:

$$P_{am}(S_t, \tau) + S_t - Ke^{-r\tau} - C_{am}(S_t, \tau) \geq 0 \qquad (8.6)$$

The proof of the second inequality is analogous but with opposite positions where we allow that the put can be exercised early, see Table 8.2:

1. buy an American call
2. sell a stock
3. buy a bond (i.e. lend money) at present value $K + D_t$ with maturity date T
4. sell an American put

Therefore, the portfolio value at any time t satisfies, as claimed:

$$C_{am}(S_t, \tau) - P_{am}(S_t, \tau) - S_t + K + D_t \geq 0 \tag{8.7}$$

2. For continuous cost of carry, we first consider the case where $b \geq r \iff d \leq 0$. We prove the left inequality first:

$$P_{am}(S_t, \tau) + S_t e^{(b-r)\tau} - Ke^{-r\tau} \geq C_{am}(S_t, \tau) \tag{8.8}$$

Consider the following portfolio at time t:

1. buy an American put
2. buy $e^{(b-r)\tau}$ stocks
3. sell bonds with nominal value K and date of expiry T
4. sell an American call

As in part 1 it follows that the value of the portfolio at time T is zero if the call is *not* exercised early. The continuous costs of carry ($d \leq 0$) are financed through selling part of the stocks so that exactly one stock is left in the portfolio at time T.

If, on the other hand, the call is exercised early at time $\tilde{t}$, the whole portfolio is then liquidated and we get:

$$\begin{aligned} P_{am}(S_{\tilde{t}}, \tau) - (S_{\tilde{t}} - K) + S_{\tilde{t}} e^{(b-r)(T-\tilde{t})} - Ke^{-r(T-\tilde{t})} = \\ P_{am}(S_{\tilde{t}}, \tau) + K(1 - e^{-r(T-\tilde{t})}) + S_{\tilde{t}}(e^{(b-r)(T-\tilde{t})} - 1) \geq 0 \end{aligned} \tag{8.9}$$

The value of the portfolio at time t is:

$$P_{am}(S_t, \tau) + S_t e^{(b-r)\tau} - Ke^{-r\tau} - C_{am}(S_t, \tau) \geq 0 \tag{8.10}$$

If $b < r \iff d > 0$ the left inequality is similarly proved,

$$P_{am}(S_t, \tau) + S_t - Ke^{-r\tau} \geq C_{am}(S_t, \tau) \tag{8.11}$$

where it is enough to hold one stock in the portfolio as $d > 0$.
We now show for the case $b \geq r$ the right inequality

$$C_{am}(S_t, \tau) \geq P_{am}(S_t, \tau) + S_t - K \tag{8.12}$$

We consider the following portfolio at time t:

1. buy an American call
2. sell an American put
3. sell a stock (short sales)
4. buy a bond with nominal value $Ke^{r\tau}$ and expiring at time T

If the put is not exercised early, we have at time T:

$$
\begin{array}{cccc}
0 & - (K - S_T) & -S_T e^{-(b-r)\tau} + K e^{r\tau} \geq 0 \text{ if } S_T < K \\
(S_T - K) + & 0 & -S_T e^{-(b-r)\tau} + K e^{r\tau} \geq 0 \text{ if } S_T \geq K
\end{array} \tag{8.13}
$$

If the put is exercised early at time $\tilde{t}$, the whole portfolio is liquidated and we get:

$$
C_{am}(S_{\tilde{t}}, \tau) - (K - S_{\tilde{t}}) - S_{\tilde{t}} e^{-(b-r)(\tilde{t}-t)} + K e^{r(\tilde{t}-t)} \geq 0 \tag{8.14}
$$

Thus the value of the portfolio at time t is:

$$
C_{am}(S_t, \tau) - P_{am}(S_t, \tau) - S_t + K \geq 0 \tag{8.15}
$$

Analogously one gets for the right inequality when $b < r$

$$
C_{am}(S_t, \tau) \geq P_{am}(S_t, \tau) + S_t e^{(b-r)\tau} - K \tag{8.16}
$$

where the stock position is reduced to $e^{(b-r)\tau}$ stocks.   □

8.2 The Trinomial Model

In general, the American option price can only be determined numerically. Similar to European option, the binomial model after Cox-Ross-Rubinstein can be used. In this section we introduce a slightly more complex but numerically efficient approach based on trinomial trees, see Dewynne et al. (1993). It is related to the classical numerical procedures for solving partial differential equations, which, of course, also may be used to solve the Black–Scholes differential equation.

The trinomial model (see Sect. 4.2) follows largely the setup of the binomial model where, however, the price can move into three, instead of two directions, at each time point $t_j = j\Delta t$, $j = 0, \ldots, n$, with $\Delta t = T/n$, see Fig. 8.4. Starting from a value S_j^k at time t_j, the price process can attain the values $u_1 \cdot S_j^k, u_2 \cdot S_j^k, u_3 \cdot S_j^k$ at t_{j+1}, where $u_i > 0, i = 1, 2, 3$, are suitably chosen parameters of the model. The probability with which the price moves from S_j^k to $u_i \cdot S_j^k$ is denoted by $p_i, i = 1, 2, 3$. The resulting price process $S_j, j = 0, \ldots, n$, in discrete time is a geometric trinomial process, i.e. the logarithm of the price $Z_j = \log S_j, j = 0, \ldots, n$, is an ordinary trinomial process with possible increments $\log u_i, i = 1, 2, 3$.

As in the binomial model three conditions must be fulfilled: The sum of the probabilities $p_i, i = 1, 2, 3$, is one, and the expectation and variance of the logarithmic increments Z_j must be the same as those of the logarithms of the geometric Brownian motion over the time interval Δt. From these conditions we get three equations:
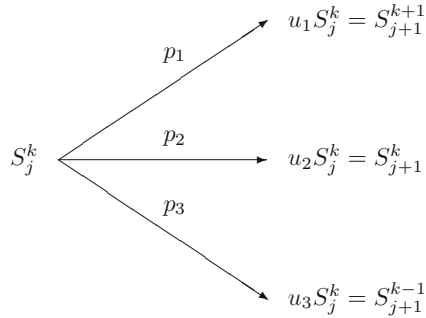
Fig. 8.4 Possible price movements in the trinomial model

$$p_1 + p_2 + p_3 = 1, \tag{8.17}$$

$$p_1 \log u_1 + p_2 \log u_2 + p_3 \log u_3 = (b - \frac{1}{2}\sigma^2)\Delta t, \tag{8.18}$$

$$p_1(\log u_1)^2 + p_2(\log u_2)^2 + p_3(\log u_3)^2 = \sigma^2 \Delta t + (b - \frac{1}{2}\sigma^2)^2 \Delta t^2 \tag{8.19}$$

For the last equation, we have used $\mathsf{E}(Z_j^2) = Var(Z_j) + (\mathsf{E}Z_j)^2$. Since there are 6 unknown parameters in the trinomial model, we have the freedom to introduce three extra conditions in order to identify a unique and possibly simple solution of the system of equations. For a clearly arranged price tree, we require the recombination property

$$u_1 u_3 = u_2^2.$$

From this, the number of possible prices at time t_n is reduced from maximally 3^n to only $2n + 1$ and consequently memory space and computation time are saved. To determine the parameters of the model we still need two more conditions. We discuss two approaches, one motivated by binomial models and the other one by numeric analysis of partial differential equations.

(a) The first approach requires that a time step of Δt in the trinomial model corresponds to two time steps in the binomial model: u_1 represents two upwards increments, u_3 two downwards increments and u_2 one upward and one downward increment (or vice versa). The binomial model fulfils the recombination condition $d = 1/u$. Since now the length of the time step is $\Delta t/2$, it follows from Sect. 7.1 that

$$u = e^{\sigma\sqrt{\Delta t/2}}$$

and the probability for the price moving upwards in the binomial model is:

$$p = \frac{1}{2} + \frac{1}{2}(b - \frac{1}{2}\sigma^2)\frac{\sqrt{\Delta t/2}}{\sigma}.$$

Then we get as conditions for the parameters of the trinomial model

$$u_1 = u^2, \; u_2 = 1, \; u_3 = 1/u^2,$$
$$p_1 = p^2, \; p_2 = 2p(1-p), \; p_3 = (1-p)^2.$$

With these parameters, the trinomial model performs as well as the corresponding binomial model, requiring however only half of the time steps. It converges therefore faster than the binomial model towards the Black–Scholes solution.

Example 8.1 Given the parameters from Table 7.1, the trinomial model provides a price of 30.769 for a European call option after $n = 50$ steps. This corresponds exactly to the value the binomial model attains after $n = 100$ steps, see Table 7.2.

                                                 Q SFEBiTree

American options differ from European options in that the options can be exercised at any time $t^*, 0 < t^* \leq T$. The value of a call falls back to the intrinsic value if it is exercised early:

$$C(S, t^*) = \max\{0, S(t^*) - K\}.$$

Mathematically, we have to solve a free boundary value problem which can only be done numerically in general.

Let again V_j^k denote the option price at time t_j if the spot price of the underlying stock is $S_j = S_j^k$. As in the binomial model for European options we determine V_j^k recursively as the discounted expectation of the option value at the next time step t_{j+1} which depends on the attainable prices V_{j+1}^{k+1}, V_{j+1}^k and V_{j+1}^{k-1}. In contrast to European options, this expectation may not fall below the intrinsic value. The recursion for the American call price $V_j^k = C_j^k$ is thus:

$$C_j^k = \max\{S_j^k - K, \; e^{-r\Delta t}[p_1 C_{j+1}^{k+1} + p_2 C_{j+1}^k + p_3 C_{j+1}^{k-1}]\}.$$

Example 8.2 Table 8.3 gives the parameters and the value of an American call option determined with $n = 50$ steps in a trinomial model. Consistent with Theorem 8.1 it gives the same value 30.769 as a European option because the underlying stock yields no dividend before maturity.

The American put is, on the other hand, more valuable than the European. With the parameters from Table 8.3 one gets $P_{eur} = 6.05140$ and $P_{am} = 6.21159$. These results are presented in Table 8.4.

Table 8.3 The value of an American call option

Spot stock price S_t	230.00
Exercise price K	210.00
Time to maturity τ	0.50
Volatility σ	0.25
Interest rate r	0.04545
Dividend	No
Steps	50
Option type	American call
Option price	30.769

Table 8.4 The value of an American call and put option

	Call	Put
European	30.769	6.051
American	30.769	6.212

(b) In the second approach the trinomial parameters p_i and u_i are determined through additional conditions. Here a certain upwards trend is shown in the whole price tree since we replace the condition $u_2 = 1$ by

$$u_2 = u \stackrel{\text{def}}{=} e^{(b-\frac{1}{2}\sigma^2)\Delta t}.$$

Furthermore we assume $p_1 = p_3$, and, together with the four conditions above, we get

$$p_1 = p_3 = p, \quad p_2 = 1 - 2p \quad \text{with } p = \frac{\Delta t}{Th^2},$$

$$u_1 = ue^{\sigma h\sqrt{T/2}}, \ u_2 = u, \ u_3 = ue^{-\sigma h\sqrt{T/2}},$$

where h is a yet free parameter. Those p_i and u_i fulfil the equation system (8.17) – (8.19) exactly. To guarantee that p_1, p_2, p_3 are not trivial probabilities, i.e they do not coincide with 0 and 1, we need $0 < p < 1/2$, and, therefore, h must fulfil the following condition:

$$h > \sqrt{\frac{2\Delta t}{T}}. \tag{8.20}$$

We now consider European options. Here the trinomial model results in the following recursion for the possible option values depending on the probabilities p_i and the change rates u_i:

$$V_j^k = e^{-r\Delta t}\left(\frac{\Delta t}{Th^2}V_{j+1}^{k+1} + (1 - 2\frac{\Delta t}{Th^2})V_{j+1}^k + \frac{\Delta t}{Th^2}V_{j+1}^{k-1}\right). \tag{8.21}$$

We consider $\Delta t = -(T - t_{j+1}) + (T - t_j)$ for all $j = 0, \ldots, n - 1$, and we put $h^* = \Delta t / T$ as well as

$$Z_j^k = V_j^k e^{-r(T-t_j)}, \; Z_{j+1}^k = V_{j+1}^k e^{-r(T-t_{j+1})}.$$

The recursion (8.21) for the option values V_j^k then becomes

$$\frac{Z_j^k - Z_{j+1}^k}{h^*} = \frac{Z_{j+1}^{k+1} - 2Z_{j+1}^k + Z_{j+1}^{k-1}}{h^2}. \tag{8.22}$$

This is the explicit difference approximation of the parabolic differential equation (6.15), see Samaskij (1984). The condition (8.20) corresponds to the well-known stability requirement for explicit difference schemes. Compared to the previously discussed approach, the probabilities p_i in this variant of the trinomial model and the calculation in (8.21) are not dependent on the volatility. The recursion (8.21) depends on σ only through the initial condition $S_n = S_T$, i.e. on the price of a stock following a geometric Brownian motion with volatility σ.

8.3   Recommended Literature

American options are classic financial derivative instruments and play a central role in the literature referenced in Chap. 6. The trinomial model as numerical procedures to approximate the option prices is introduced in detail in Dewynne et al. (1993) and Wilmott et al. (1995).

8.4   Exercises

Exercise 8.1 *Explain the relation between American call and put, and the following: value of the underlying asset, exercise price, stock volatility, interest rate, time to exercise date.*

Exercise 8.2 *Consider a stock whose price increases by* 20 % *or decreases by* 10 % *in each period. We are now at* $t = 0$ *and we have an American call option on this stock with an exercise price of* 10.5 *and a terminal value at* $T = 2$. *What will be the*

price of this American call option at $t = 0$*? Will it be different from the price of a European call option? (Set the interest rate and dividend equal to* 0*, and denote the intrinsic value at time* T *as* C_T*.)*

Exercise 8.3 *A stock price is currently* 50*. The price can increase by a factor of* 1.10*, or fall by a factor of* 0.90*. The stock pays no dividends and the yearly discrete compounding interest rate is* 0.05*. Consider American put and call options on this stock with strike price* 50*, and 2 years' time to maturity and 1 year's step length .*

 (a) *What will be the price of this American call option at* $t = 0$*? Will it be different from the price of a European call option?*
 (b) *What will be the price of this American put option with the same strike price?*
 (c) *Does the put call parity hold?*

Exercise 8.4 *Consider the same model in* 8.3*. There is no dividend, and we use continuous compounding interest rate.*

1. *Find the value of both European and American call options with strike prices of* 50 *and maturities of 2 years. The yearly compounding risk-free rate is 5 %.*
2. *Find the value of both European and American put options with strike prices of* 50 *and maturities of 2 years. The yearly compounding risk-free rate is 5 %.*
3. *Is the put–call parity relation satisfied by the European options? For the American ones? Would you predict that the American put price will be higher than its parity value in general? Explain.*

Exercise 8.5 *Consider the same model in* 8.3*. However, this time we know that there will be a dividend payment at* $t = 1$ *equal to* 5*.*

1. *Find the value of both European and American call options with strike prices of* 50 *and maturities of 2 years. The yearly compounding risk-free rate is 5 %.*
2. *Find the value of both European and American put options with strike prices of* 50 *and maturities of 2 years. The yearly compounding risk-free rate is 5 %.*
3. *Is the put–call parity relation satisfied by the European options? For the American ones? Would you predict that the American put price will be higher than its parity value in general? Explain.*

Exercise 8.6 *Consider the same model in* 8.3*, but extend the binomial model to* 10 *steps in 2 years for an American put option.*

Chapter 9
Exotic Options

There is a multitude of complex, so-called exotic options, which are mainly used in OTC (over-the-counter) trading to meet the special needs of institutional investors. The most important types of exotic options are:

- Compound Options
- Chooser Options
- Barrier Options
- Asian Options
- Lookback Options
- Cliquet Options
- Basket Options

In contrast, ordinary puts and calls, no matter whether European or American style, are frequently called *plain vanilla options*.

Figure 9.1 shows an example for DAX warrants issued by Sal. Oppenheim. These products payout at maturity a fraction (denoted as BV in Fig. 9.1) of the value of a barrier option. Pricing of a barrier option is a challenging task though if the barrier is not crossed, coincides with a corresponding European option. As there are European calls and puts there are two types of warrants: Long and Short. In the example, the strike price of the European options and the barrier (StopLoss) coincide. Moreover, the warrants all payoff 1 % of the value of the barrier options as the last but one column shows. The maturity is given in the last column of the table.

Basiswert:	DAX	4.130,81		+41,68	+1,02%	11.11.2004		Java-Applet:	aktiv	Neu Starten
WKN	**Typ**	**Bid**	**Zeit**	**Ask**	**Zeit**	**Strike**	**StopLoss**	**Währung**	**BV**	**Fälligkeit**
SAL60F	Long	2.470	7:05:14 PM	2.490	7:05:14 PM	3.900,00	3.900,00	XXP	0,01	23.12.2004
SAL6UC	Long	2.650	7:05:24 PM	2.670	7:05:24 PM	3.900,00	3.900,00	XXP	0,01	24.03.2005
SAL60G	Long	2.240	7:05:14 PM	2.260	7:05:14 PM	3.925,00	3.925,00	XXP	0,01	23.12.2004
SAL6U9	Long	1.970	7:05:14 PM	1.990	7:05:14 PM	3.950,00	3.950,00	XXP	0,01	23.12.2004
SAL6UA	Long	1.730	7:05:14 PM	1.750	7:05:14 PM	3.975,00	3.975,00	XXP	0,01	23.12.2004
SAL6UB	Long	1.470	7:05:14 PM	1.490	7:05:14 PM	4.000,00	4.000,00	XXP	0,01	23.12.2004
SAL4VM	Short	0.160	7:05:14 PM	0.180	7:05:14 PM	4.150,00	4.150,00	XXP	0,01	23.12.2004
SAL4VN	Short	0.410	7:05:14 PM	0.430	7:05:14 PM	4.175,00	4.175,00	XXP	0,01	23.12.2004
SAL1S6	Short	0.660	7:05:24 PM	0.670	7:05:24 PM	4.200,00	4.200,00	XXP	0,01	23.12.2004
SAL2GN	Short	0.610	7:05:27 PM	0.630	7:05:27 PM	4.200,00	4.200,00	XXP	0,01	24.03.2005

Fig. 9.1 Basic information on DAX warrants issued by Sal. Oppenheim

9.1   Compound Options, Option on Option

With a compound option one has the right to acquire an ordinary option at a later date. To illustrate such a compound option consider a *Call-on-a-Call* with the parameters:

$$\text{maturity dates } T_1 < T_2$$

$$\text{strike prices } K_1, K_2.$$

Such a compound option gives the owner the right at time T_1 to buy a call for the price K_1 that has a maturity date T_2 and a strike price K_2.

The value $V(S, t)$ of this option at time t with an actual price $S_t = S$ can be calculated by applying the Black–Scholes formula twice:

(1) Beginning at time $t = T_2$, calculate (explicitly or numerically) the value of the call, which can be bought at time $t = T_1$. Let $C(S, T_1)$ be this value.
(2) The purchasing option of the call at time T_1 is only exercised when

$$C(S, T_1) > K_1$$

Thus we get

$$V(S, T_1) = \max\{C(S, T_1) - K_1, \ 0\}.$$

Now, calculate $V(S, t)$ for $t < T_1$ using the Black–Scholes equation with the above constraint at $t = T_1$ analogously to an ordinary call.

Remark 9.1 Since $V(S, t)$ is only dependent on t and the price S_t at time t, this value function fulfils the Black–Scholes equation so that our approach is justified.

We illustrate the pricing of a compound option in the binomial tree model of Cox, Ross and Rubinstein. We consider an asset with an initial value of 230 and a volatility of 25 %. The corresponding tree for 1 year with time steps $\Delta t = 0.2$ is shown in Table 9.1. The corresponding prices C of a European call with a strike of $K_2 = 210$ that expires in $T_2 = 1$ year are presented in Table 9.2. The value of a compound option on this call with compound expiration at $T_1 = 0.4$ and compound strike $K_1 = 20$ is shown in Table 9.3.

Table 9.1 Binomial tree of a stock

					402.26
				359.71	
			321.66		321.66
		287.63		287.63	
	257.21		257.21		257.21
230		230		230	
	205.67		205.67		205.67
		183.91		183.91	
			164.46		164.46
				147.06	
					131.51

Table 9.2 Binomial tree of a call option

					192.25
				151.60	
			115.42		111.66
		83.75		79.53	
	58.14		51.99		47.21
38.91		32.30		23.98	
	19.39		12.18		
		6.19			

Table 9.3 Binomial tree of a call on a call option

		63.75
	38.32	$\max\{C^{uu} - K_1; 0\}$
22.49	PV(63.75, 12.30)	12.30
PV(38.32,6.25)	6.25	$\max\{C^{ud} - K_1; 0\}$
	PV(12.30, 0)	0
		$\max\{C^{dd} - K_1; 0\}$

9.2   Chooser Options or "As You Wish" Options

A *Chooser Option* is a form of the compound option, where the buyer can decide at a later date which type of option he would like to have. To illustrate this, consider a *regular Chooser Option* with the parameters:

$$\text{maturities } T_1 < T_2$$

$$\text{strike prices } K_1, K_2.$$

This option gives the right at time T_1 for the price K_1 to buy a call or a put (as preferred), which has a maturity T_2 and a strike price K_2: in the language of compound options this is referred to as a *Call-on-a-Call* or *Put*.

The value $V(S, t)$ can be found by applying the Black–Scholes formula three times:

(1) Determine the value $C(S, T_1)$ and $P(S, T_1)$ of a call and a put with maturity T_2, and strike price K_2.
(2) Solve the Black–Scholes equation for $V(S, t), t < T_1$ with the constraint at $t = T_1$:

$$V(S, T_1) = \max\{C(S, T_1) - K_1, \ P(S, T_1) - K_1, 0\}$$

9.3   Barrier Options

The value of a *Barrier Option* jumps as soon as the stock price reaches a given barrier, which in general may also be time dependent. As an example consider a simple European barrier option with

$$\text{maturity } T, \text{ strike price } K \text{ and barrier } B$$

which gives the holder the right to buy a stock at time T for the price K provided that

- *down-and-out:* $S_t > B$ for all $0 \leq t \leq T$
- *up-and-out:* $S_t < B$ for all $0 \leq t \leq T$

This type of *Knock-out Option* becomes worthless as soon as the stock price S_t reaches the barrier. Figure 9.2 represents the situation for a down-and-out option with two possible paths of the asset price. When the price hits the barrier (lower path), the option expires worthless regardless of any further evolution of the price.

A *Knock-in Option* is just the opposite to the knock-out option. It is worthless until the barrier is reached, and it behaves as the corresponding plain vanilla option afterwards.
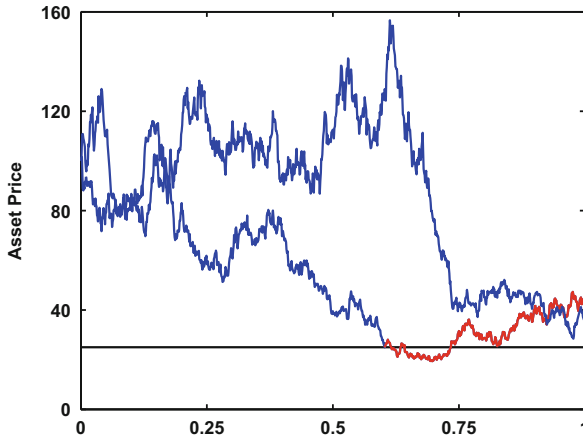
Fig. 9.2 Two possible paths of the asset price. When the price hits the barrier (*lower path*), the option expires worthless   Q SFEdown-and-out

For example, a European *Knock-in Call* consists of the right to buy stock provided that

- *down-and-in:* $S_t \leq B$ for some $0 \leq t \leq T$
- *up-and-in:* $S_t \geq B$ for some $0 \leq t \leq T$

The value of a barrier option depends no longer only on the stock price at a specific point in time, but on the overall evolution of the stock price during the option's whole life span, i.e. on the path of the price process up to maturity T. Thus in principle it does not fulfil the Black–Scholes differential equation. Here, the path dependence is, however, simple enough to work still with the conventional Black–Scholes approach. As an example consider a *Down-and-out Call* with $K > B$. As long as $S_t > B$, $V(S, t)$ fulfils the Black–Scholes equation with the restriction:

$$V(S, T) = \max(S_T - K,\ 0)$$

In case that the price reaches the barrier B, the option of course becomes worthless:

$$V(B, t) = 0\,, \qquad 0 \leq t \leq T,$$

is therefore an additional restriction that needs to be taken into account when solving the Black–Scholes differential equation. The explicit solution is given as:

$$V(S, t) = C(S, t) - \left(\frac{B}{S}\right)^{\alpha} C\left(\frac{B^2}{S}, t\right)$$

with $\alpha = \dfrac{2r}{\sigma^2} - 1$, where $C(S,t)$ represents the value of a common European call on the stock in question. The value $\bar{V}(S,t)$ of a European *Down-and-in Call* can be calculated analogously. If one already knows $V(S,t)$, one can also use the equation

$$\bar{V}(S,t) + V(S,t) = C(S,t).$$

It is fulfilled since a down-and-in and a down-and-out call together have the same effect as a plain vanilla call.

As an alternative to barrier options which depend on the whole path of the stock price between time 0 and T, frequently *discretely monitored barrier options* are considered which depend only on the stock prices at a finite number of time points, e.g. on the daily or weekly closing prices up to maturity. Given $0 = t_0 < t_1 < \ldots < t_n = T$, a corresponding discretely monitored European down-and-out call with maturity T, strike price K and barrier B gives the right to buy a stock at time T provided that

$$S_{t_i} > B \quad \text{for all } i = 0, \ldots, n$$

Those options require approximate methods for pricing, but they have the advantage that appropriate numerical schemes converge much faster than in the continuously monitored situation. This is in particular of interest if one leaves the Black–Scholes framework and uses models with more complex underlying dynamics. In Skaug and Naess (2007), e.g. a fast pricing algorithm based on numerical approximation of the transition probability density associated with the stochastic differential equation describing the stock price dynamics is described.

9.4 Asian Options

For Asian options, the value depends on the average stock price calculated over the entire life span of the option. With an *Average Strike Option* over the time period $0 \le t \le T$, for example, the payoff has at maturity the form

$$\max\left(S_t - \frac{1}{t}\int_0^t S_s\, ds\,,\ 0\right), t = T.$$

With an *American Average Strike Option* this is also the payoff when the option is exercised early at some arbitrary time $t \le T$.

To calculate the value of an *Asian Option*, we consider a more general class of European options with a payoff at time T which depends on S_T and I_T where for some given function f of stock price and time

$$I_t = \int_0^t f(S_s, s)ds.$$

Analogous to the Black–Scholes equation we derive an equation for the value of such a *path dependent option* $V(S, I, t)$ at time t. With a stock price S_t at time t we have

$$I_t + dI_t \overset{\text{def}}{=} I_{t+dt} = \int_0^{t+dt} f(S_s, s)ds$$

$$= I_t + \int_t^{t+dt} f(S_s, s)dt$$

$$= I_t + f(S_t, t)dt + \mathcal{O}(dt).$$

Thus the differential of I_t is equal to $dI_t = f(S_t, t)dt$.

Using Itôs Lemma it follows for $V_t = V(S_t, I_t, t)$ that:

$$dV_t = \sigma S_t \frac{\partial V_t}{\partial S}dW_t + f(S_t, t)\frac{\partial V_t}{\partial I}dt$$

$$+ \left(\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V_t}{\partial S} + \frac{\partial V_t}{\partial t}\right)dt$$

Analogous to the derivation of the Black–Scholes formula, continuous delta hedging produces a risk free portfolio from an option and $\Delta_t = \partial V_t/\partial S$ sold stocks. Together with the condition of no arbitrage it follows for the case without dividends $(b = r)$ that:

$$\frac{\partial V_t}{\partial t} + f(S_t, t)\frac{\partial V_t}{\partial I} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 V_t}{\partial S^2} + rS_t \frac{\partial V_t}{\partial S} - r\,V_t = 0$$

This is the Black–Scholes equation with an additional term $f(S_t, t)\frac{\partial V_t}{\partial I}$. The boundary condition in this case is

$$V(S, I, T) = g(S, I, T),$$

where, e.g., for an *Average Strike Call* we have

$$g(S, I, t) = \max(S - \frac{1}{t}I, \ 0) \quad \text{and} \quad f(S, t) = S.$$

For European options an explicit analytic solution of the differential equation exists in which complicated, specialized functions appear, the so-called confluent hypergeometric functions. The numerical solution, however, is easier and faster to obtain.

The integral $\int_0^t S_s ds$ in practice is calculated as the sum over all quoted prices, for example, at 30 s time intervals. As for barrier options, discrete time Asian options

are considered too, using a substantially larger time scale. Here, I_t changes only once a day or once a week:

$$I_t = \sum_{j=1}^{n(t)} S_{t_j} \ , \quad t_{n(t)} \leq t < t_{n(t)+1}$$

with $t_{j+1} - t_j = 1$ day or $= 1$ week and closing prices S_{t_j}.

Such a discretely monitored Asian option corresponds largely to an ordinary option with discrete dividend payments at time periods $t_1, t_2, \ldots$. From the assumption of no arbitrage follows a continuity constraint at t_j:

$$V(S, I, t_j -) = V(S, I + S, \ t_j +)$$

To determine the price of the option one begins as usual at the time of maturity where the value is known:

(1) $T = t_n$

$$V(S, I, T) = \max(S - \frac{1}{T} I_T, \ 0)$$

$$I_T = \sum_{j=1}^{n} S_{t_j}$$

Solve the Black–Scholes equation backwards to time t_{n-1} and obtain

$$V(S, I + S, \ t_{n-1} +)$$

(2) Using the continuity constraint, calculate the new terminal value $V(S, I, t_{n-1}-)$. Solve the Black–Scholes equation backwards to time t_{n-2} and obtain

$$V(S, I + S, \ t_{n-2} +)$$

etc.

9.5   Lookback Options

The value of a lookback option depends on the maximum or minimum of the stock price over the entire life span of the option, for example, a *Lookback Put* over the time period $0 \leq t \leq T$ has at maturity the payoff

$$\max(M_T - S_T, \ 0) \quad \text{with} \quad M_t = \max_{0 \leq s \leq t} S_s.$$
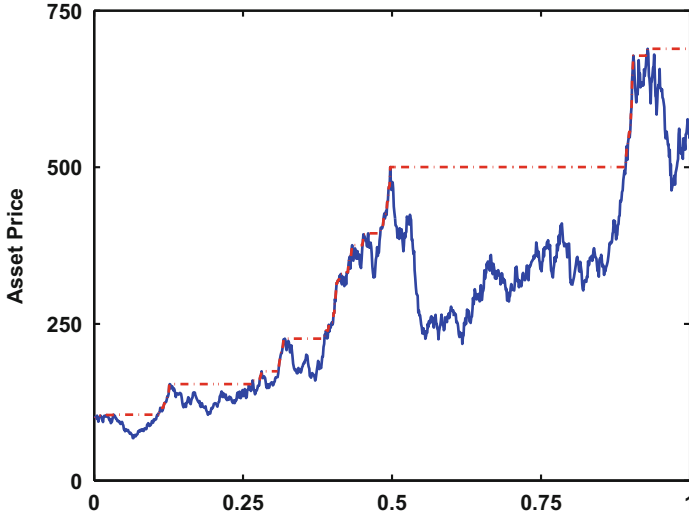
Fig. 9.3 Sample paths of price process and corresponding maximum process

SFEmaximumProcess

In the sample path of Fig. 9.3, we see a simulated price process (S_t) as a solid line and the maximum process (M_t) as increasing dotted line. The value of a lookback put is then given by the difference of the end value of the increasing curve and the final value of the price process.

To calculate the value of such an option, we first consider a sequence of path dependent options as in the previous section with

$$I_t(n) = \int_0^t S_s^n \, ds, \quad \text{i.e.} \quad f(S,t) = S^n, \quad n = 1, 2, \ldots$$

With $M_t(n) = \{I_t(n)\}^{\frac{1}{n}}$ it holds that:

$$M_t = \lim_{n \to \infty} M_t(n).$$

From the differential equation for $I_t(n)$ and $n \to \infty$ it follows that the value $V_t = V(S_t, M_t, t)$ of a European lookback put fulfils the following equation:

$$\frac{\partial V_t}{\partial t} + \frac{1}{2}\sigma^2 S_t^2 \frac{\partial^2 V_t}{\partial S^2} + r \, S_t \frac{\partial V_t}{\partial S} - r \, V_t = 0$$

This is the usual Black–Scholes equation. M only appears as an argument of V and in the boundary condition:

$$V(S, M, T) = \max(M - S, \, 0)$$

The solution for a time to expiry $\tau = T - t$ and $\alpha = 2r/\sigma^2$ is:

$$V(S, M, t) = S \left\{ \Phi(y_1) \cdot \left(1 + \frac{1}{\alpha} \right) - 1 \right\}$$

$$+ M \ e^{-r\tau} \left\{ \Phi(y_3) - \frac{1}{\alpha} \left(\frac{M}{S} \right)^{\alpha - 1} \Phi(y_2) \right\}$$

with $y_1 = \frac{1}{\sigma \sqrt{\tau}} \left\{ \log \frac{S}{M} + \left(r + \frac{1}{2}\sigma^2 \right) \tau \right\}$

$y_2 = \frac{1}{\sigma \sqrt{\tau}} \left\{ \log \frac{S}{M} - \left(r + \frac{1}{2}\sigma^2 \right) \tau \right\}$

$y_3 = \frac{1}{\sigma \sqrt{\tau}} \left\{ \log \frac{M}{S} - \left(r + \frac{1}{2}\sigma^2 \right) \tau \right\} .$

9.6   Cliquet Options

A cliquet option settles periodically and resets the strike at the level of the underlying during the time of settlement. It is therefore a series of at-the-money options, but where the total premium is determined in advance. A cliquet option can be thought of as a series of "pre-purchased" at-the-money options. The payoff on each option can either be paid at the final maturity, or at the end of each reset period. The number of reset periods is determined by the buyer in advance. All other parameters equal, more resets make the option more expensive.

The payoff of the cliquet call at maturity $T = t_n$ is given by

$$\max \{(S_{t_0}, S_{t_1}, \dots, S_{t_n}) - S_{t_0}\}$$

and the payoff of the cliquet put is given by

$$\max \{S_{t_0} - (S_{t_0}, S_{t_1}, \dots, S_{t_n})\} .$$

Example 9.1 Consider a cliquet call with maturity $T = 3$ years and strike price $K_1 = 100$ in the first year. If in the first year the underlying were $S_1 = 90$, the cliquet option would expire worthless in the first year. The new strike price for the second year will be set to $K_2 = 90$. Assume, in the second year $S_2 = 120$, then the contract holder will receive a payoff and the strike price would reset to the new level of $K_3 = 120$ for the third year. Thus, higher volatility provides better

conditions for investors to earn profits. Prices of cliquet options obey in a Black–Scholes model a partial differential equation that has one variable, in addition to the standard variables of the Black–Scholes partial differential equation.

9.7   Basket Options

The value of a common basket option is not path dependent, but, as for plain vanilla options, it depends only on the prices at maturity. Nevertheless, those options are frequently counted among the exotic options. Pricing is again complicated, as the option depends on several underlying assets forming the *basket*. Here, the main problem is not path dependency, but multidimensionality which does not allow for explicit option pricing formulas.

The value of the basket can be any weighted sum of the prices of the underlying assets as long as the weights are positive. To illustrate this type of option, let us consider m assets with prices $S_{1t}, \ldots, S_{mt}$ at time t and m weights $w_1, \ldots, w_m > 0$. The value of the basket is, then,

$$B_t = w_1 S_{1t} + \ldots + w_m S_{mt}.$$

Typical examples of such baskets are stock indices like the German DAX, but assets and weights may be chosen in an arbitrary manner. A basket call with maturity T and strike price K is an option to buy the basket, i.e. a portfolio of the underlying assets with portfolio weights given by $w_1, \ldots, w_m$. Its value at maturity is

$$V(B_T, T) = \max\{B_T - K, \, 0\}.$$

There are two main problems concerning the pricing of basket options. First, we have to model the dependencies between the assets forming the basket. If we stick to the Black–Scholes framework, we model the price vector $S_t = (S_{1t}, \ldots, S_{mt})$ as a m-dimensional geometric Brownian motion which is driven by a m-dimensional Wiener process. The $m \times m$-correlation matrix of the latter specifies the dependency of the model. Alternatively, the dependency is quantified using the concept of copulae, compare Chap. 17. Even for moderately large m, the model contains many parameters which have to be fitted to available data.

The second problem is originating from the observation that, in the Black–Scholes framework, stock prices are modelled as lognormal random variables. In contrast to normal random variables, a linear combination of lognormal variables is, however, not lognormal any more. Therefore, the Black–Scholes model does not hold exactly for the basket value B_t, and we have to use approximations or numerical procedures for pricing basket options. Due to the multidimensionality of the problem, numerical methods require a large computation time which is not always acceptable for practical purposes. Therefore, fast procedures which do not provide the exact option price under the assumed model but only a good approximation are

of interest. In Krekel and Man (2007) several such approximations are discussed and compared in a numerical study. For not too large m, a method of Ju (1992), which relies on characteristic functions as an alternative to specify probability distributions and on a Taylor expansion, turned out to be the best one except for underlying assets with considerable inhomogeneity in the volatilities. In the latter case, a lower bound for the basket option value proposed by Beisser (2001) turned out to be closer to the true price where the latter has been calculated by computationally expensive numerical techniques.

9.8   Recommended Literature

Exotic options such as bond options are discussed in detail in Hull (2006). The numerical methods necessary for valuing complex path-dependent derivatives are briefly reviewed in Wilmott et al. (1995) and are discussed in more detail in Dewynne et al. (1993).

9.9   Exercises

Exercise 9.1 (Compound Option) *A compound option is also called option on option. It allows the purchaser to acquire an ordinary option at a later date. Consider a European Call-on-a-Call option, with the first expiration date T_1, the second expiration date T_2, the first strike price K_1 and the second strike price K_2.*

(a) *Determine the value of the compound option at time T_1.*
(b) *Let $T_1 = 4$ months, $T_2 = 12$ months, $K_1 = 25$, $K_2 = 220$, initial value of the asset, volatility $\sigma = 0.23$, $r = 0.034$ and $S_{T_1} = 230$, calculate the value of the compound option at time T_1.*

Exercise 9.2 (Chooser Option) *A chooser option (preference option) is a path dependent option for which the purchaser pays an up-front premium and has the choice of exercising a vanilla put or call on a given underly at maturity. The purchaser has a fixed period of time to make the choice. At time $0 < T_0 < T$, the purchaser chooses the option with the higher value.*

(a) *Give the payoff of the chooser option.*
(b) *In a non-arbitrage framework, show that the price of the call and the put option is equivalent to the price of a chooser option:*

$$C^{ChooserOption} = C^{Call} + C^{Put} \qquad (9.1)$$

where C^{Call} denotes the call option with underlying price S_t, strike price K and maturity T. C^{Put} defines the put option with strike price $K(1+r)^{T_0-T}$ and maturity T_0. (Hint: $\{(1+r)^{-t} S_t\}$ is a martingale.)

Exercise 9.3 (Barrier Option) *A barrier Option changes its value in leaps as soon as the stock price reaches a given barrier, which can also be time dependent. A European down-and-in call is a barrier option which starts to be active only when the underlying $S_t \leq B$ at any time $0 \leq t \leq T$, and a European down-and-out call is a barrier option which expires as worthless if the underlying $S_t \leq B$ at any time $0 \leq t \leq T$. The two options share the same maturity time T and strike price K. Explain why the down-and-in and the down-and-out call together have the same effect as a normal European call (In-Out-Parity).*

Exercise 9.4 (Forward Start Option) *Forward start options are options whose strike is unknown at the beginning, but will be determined at an intermediate time t. So a forward start option is similar to a vanilla option except for not knowing the strike price at the moment of purchase. The strike is usually determined by the underlying price at time t. Let S_t denote a random path in a binomial tree. Let $a > 0$ and $b < 0$ denote the upward rate and downward rate. r is the risk free interest rate.*

Let $0 < t < T$. Calculate the price and a Delta hedge for the forward start option, whose payoff is given by

$$\max\{(S_T/S_t - K), 0\}$$

(see Susanne Kruse (2005)) where K is the strike price and S_T is the stock price at maturity T.

Exercise 9.5 (Forward Start Option) *Consider a call option with forward start $t = 4$ months from today ($t = 0$). The option starts at $K = 1.1 S_t$, time to maturity is $T = 1$ year from today, the initial stock price S_0 is 60, the risk free interest rate is $r = 9\%$, the continuous dividend yield is $d = 3\%$ and the expected volatility of the stock is $\sigma = 20\%$. What is the price of this forward start option?*

Exercise 9.6 (Cliquet Option) *In a cliquet option, the strike price periodically resets before the expiration date is reached. If the expiration date is reached, and the underlying price is below the strike price, the option will expire as worthless, and the strike will be reset to the lowest value of the underlying price. If the expiration date is reached and the underlying value is higher than the strike price, the purchaser of the option will earn the difference and the strike price will reset to the higher underlying price. Consider a cliquet call with maturity $T = 3$ years and strike price $K_1 = 100$ in the first year. Suppose that the underlying prices in the 3 years are $S_1 = 90$, $S_2 = 120$, $S_3 = 130$. What is the payoff of the cliquet option?*

Chapter 10
Interest Rates and Interest Rate Derivatives

The interest rate derivatives market is the largest derivatives market in the world. Mostly traded OTC, the interest rate securities are extremely popular especially among large institutional investors. Thus, the valuation of these instruments has been a major challenge of both practitioners and academics. Pricing interest rate derivatives fundamentally depends on the term structure of interest rates. Until now we have assumed:

- constant risk free domestic interest rate
- independence of the price of the option from the possibly stochastic interest rate r.

When considering interest rates in practice both of these assumptions will not be fulfilled. Particularly, during the recent Financial Crisis 2008–2013 one has observed rapid changes of the term structure driven by random components (risks) involved. Just as the dynamics of a stock price is unknown and has to be modeled via a stochastic process, the stochastics of interest rates are derived by modeling their dynamics. Here, however, there is no unique modeling framework like the Black–Scholes model for equities; practitioners have tended to use different models for valuing different interest-rate derivatives. These standard models for the term structure have become a core part of financial engineering.

This chapter is structured as follows. First, for beginners in the interest rate world, essential concepts are introduced. Being equipped with this background, we postulate the theoretical issues the interest rate model should cope with: no-arbitrage condition and the change of the numeraire technique. This framework is crucial for pricing interest rate derivatives which are discussed in the following part. Together with this fundamental instruments we analyze the Black pricing formula. We present an overview of different interest rate models of, e.g. one-factor and two-factor type. The Heath Jarrow Morton framework is covered and next we introduce the LIBOR Market Model. We conclude the chapter with empirical results.

10.1   Definitions and Notation

Definition 10.1 A bond $V\{r(t), t, T\}$ produces at the time of maturity T a fixed amount Z, the nominal value, and if applicable, at predetermined dates before T coupon payments.

For simplicity of notation we write $V\{r(t), t, T\}$ as $V(t, T)$. If there are no coupons, the bond is referred to as a zero coupon bond or zero bond for short. We consider $V(t, T)$ as a unit principal (i.e. $V(T, T) = 1$) zero coupon bond in the following sections. We further restrict ourselves to default free government debt.

The simple rate of return $R(t, T)$ from holding a bond over the time interval $\tau(t, T) = T - t$ equals:

$$R(t, T) = \frac{1 - V(t, T)}{\tau(t, T)V(t, T)} = \frac{1}{\tau(t, T)} \left\{ \frac{1}{V(t, T)} - 1 \right\}. \qquad (10.1)$$

The equivalent rate of return, with continuous compounding, is referred to as the yield to maturity on a bond.

Definition 10.2 The yield to maturity $Y(t, T)$ is the internal rate of return at time t on a bond with maturity date T.

$$Y(t, T) = -\frac{1}{\tau(t, T)} \log V(t, T). \qquad (10.2)$$

The rates $Y(t, T)$ considered as a function of time T will be referred to as the term structure of interest rates at time t. The straightforward relationship between the yield to maturity and the bond price is given by:

$$V(t, T) = \exp\{-Y(t, T)\tau(t, T)\}.$$

Remembering from Chap. 1, in order to transform $R(t, T)$ into a continuous compounding rate we need to calculate the relationship between $R(t, T)$ and the continuous compounded rate $Y(t, T)$:

$$Y(t, T) = \frac{1}{\tau(t, T)} \log\{1 + R(t, T)\tau(t, T)\}. \qquad (10.3)$$

We have just seen that the yield to maturity applies to a bond $V(t, T)$. The forward rate $f(t, T, S)$ corresponds to the internal rate of return of a bond $V(T, S)$.

Definition 10.3 The forward rate $f(t, T, S)$ is the internal rate of return at time t on a bond lasting from time T to the maturity date S, with $t < T < S$.

$$f(t, T, S) = \frac{\log V(t, T) - \log V(t, S)}{\tau(T, S)}. \qquad (10.4)$$

This shows that the forward rate can be thought of as the yield to maturity of a bond lasting from time T to time S, i.e. $f(t, T, S) = Y(T, S)$.

An intuitive approach to the forward rate is by considering the forward rate in terms of arbitrage free investments and in a simple compounding manner. If one would invest 1 EUR in a bond $V(t, T)$ and at maturity T re-invest the received amount in a bond $V(T, S)$, by no arbitrage this has to be equal to an investment of 1 EUR at time t in a bond $V(t, S)$. Therefore due to the no arbitrage condition and following the same path as Eq. (10.1):

$$V(T, S) = \frac{V(t, S)}{V(t, T)}$$

and therefore

$$
\begin{aligned}
F(t, T, S) &= \frac{1}{\tau(T, S)} \left\{ \frac{V(t, T) - V(t, S)}{V(t, S)} \right\} \\
&= \frac{1}{\tau(T, S)} \left\{ \frac{V(t, T)}{V(t, S)} - 1 \right\},
\end{aligned}
\tag{10.5}
$$

where $F(t, T, S)$ is the simple compounded forward rate.

By applying Eq. (10.3) to the forward rate we achieve equality with Eq. (10.4):

$$f(t, T, S) = \frac{1}{\tau(T, S)} \log\{1 + F(t, T, S)\tau(T, S)\}.$$

The instantaneous forward rate $f(t, T)$ is the limiting case of the forward rate $f(t, T, S)$, for $S \to T$:

$$f(t, T) = \lim_{S \to T} f(t, T, S).$$

The application of l'Hospital's rule gives:

Definition 10.4 The instantaneous forward rate $f(t, T)$ is the forward interest rate at time t for instantaneous risk free borrowing or lending at time T.

$$f(t, T) = -\frac{\partial \log V(t, T)}{\partial T}.
\tag{10.6}$$

The existence of $f(t, T)$ assumes that the continuum of bond prices is differentiable w.r.t T.

It holds that:

$$V(t, T) = \exp\left\{ -\int_t^T f(t, s)\, ds \right\}.$$

10.1.1   *Money Market Account*

One of the most basic instruments related to interest rates is the money market (or savings) account. The money market account represents a risk less investment at the prevailing instantaneous interest rate $r(t)$, where

$r(t)$ = spot rate = interest rate for the shortest possible investment.

Definition 10.5 Define $A(t)$ as the value of the money market account at time t. We assume $A(t) = 1$ and that the account develops according to the following differential equation:

$$dA(t) = r(t)A(t)dt,$$

with $r(t)$ as a positive function of time.

As a consequence:

$$A(T) = \exp\left\{\int_t^T r(s)\,ds\right\}. \tag{10.7}$$

At any time t, the current value $r(t)$ of the spot rate is the instantaneous rate of increase of the money market account value. One can also interpret $r(t)$ as the limiting case of the simple rate of return $R(t, T)$ from holding a bond. Similarly to the instantaneous forward rate we obtain:

$$V(t, T) = \exp\left\{-\int_t^T r(s)\,ds\right\}. \tag{10.8}$$

The short rate $r(t)$ can be derived at any time t, however subsequent values of the spot rate are unknown. In fact, it is assumed that $r(t)$ is an Itô process:

$$dr(t) = \mu\{r(t), t\}dt + \sigma\{r(t), t\}dW_t \tag{10.9}$$

with W_t being a Wiener process. For the moment we restrict ourselves to this basic set-up. More details are given in Sect. 10.4.1.

We have now presented the basic concepts of interest rates. In the next section we give a quick introduction to risk neutral pricing and equivalent martingale measures. One needs these tools in order to understand the following pricing of more complex derivatives such as caps, floors, swap options and bond options.

10.2   Risk Neutral Valuation and Numeraire Measures

It is vital to understand the principles of no-arbitrage valuation in a world with
stochastic interest rates. If interest rates are stochastic the market price of risk
becomes an essential factor in valuation of derivatives. Likewise, the numeraire
measure needs to be introduced. We show how a convenient choice of numeraire
and a corresponding choice for the market price of risk can greatly simplify the
valuation of certain interest rate derivatives. We first give a brief introduction to the
principles of risk neutral valuation and the market price of risk, followed by a first
look at techniques of measure change and a series of different possible numeraire
measures are considered. We see how the methods introduced here can be combined
with the Black model to value interest rate derivatives.

10.2.1   Principles of Risk Neutral Valuation

Definition 10.6 The market price of risk defines a value above the risk free return
for an asset.

Assume that the process θ_t is a geometric Brownian motion:

$$d\theta_t = m\theta_t dt + s\theta_t dW_t \qquad (10.10)$$

where dW_t is a Wiener Process.

Assume further that V_{1t} and V_{2t} are the prices of two derivatives dependent only
on θ_t and t. As a simplification, no payments are allowed during the observation
time period. The processes followed by $V_{jt} = V_j(\theta, t)$, $j = 1,2$ are:

$$dV_{jt} = \mu_{jt}V_{jt}dt + \sigma_{jt}V_{jt}dW_t, \qquad (10.11)$$

where μ_{jt}, σ_{jt} could be functions of θ_t and t. The random process W_t in Eqs. (10.10)
and (10.11) is the same, as there is only one source of uncertainty.

If we would construct a portfolio Π_t of $\sigma_{2t}V_{2t}$ units of V_{1t} and short sell $-\sigma_{1t}V_{1t}$
units of V_{2t} this portfolio is instantaneously risk less.

$$
\begin{aligned}
\Pi_t &= (\sigma_{2t}V_{2t})V_{1t} - (\sigma_{1t}V_{1t})V_{2t} \\
&= (\sigma_{2t}V_{2t})(\mu_{1t}V_{1t}dt + \sigma_{1t}V_{1t}dW_t) - (\sigma_{1t}V_{1t})(\mu_{2t}V_{2t}dt + \sigma_{2t}V_{2t}dW_t) \\
&= (\sigma_{2t}V_{2t}V_{1t}\mu_{1t} - \sigma_{1t}V_{1t}V_{2t}\mu_{2t})dt. \qquad (10.12)
\end{aligned}
$$

This portfolio would be risk free and thus in the time period dt it must produce
the risk free profit $r(t)dt$:

$$d\Pi_t = r(t)\Pi_t dt. \qquad (10.13)$$

Together with Eqs. (10.12) and (10.13) this gives:

$$(\sigma_{2t} V_{2t} V_{1t} \mu_{1t} - \sigma_{1t} V_{1t} V_{2t} \mu_{2t})dt = (\sigma_{2t} V_{2t} V_{1t} - \sigma_{1t} V_{1t} V_{2t})r(t)dt$$

$$\sigma_{2t} \mu_{1t} - \sigma_{1t} \mu_{2t} = r(t)\sigma_{2t} - r(t)\sigma_{1t}$$

$$\frac{\mu_{1t} - r(t)}{\sigma_{1t}} = \frac{\mu_{2t} - r(t)}{\sigma_{2t}}$$

The quantity:

$$\lambda_t = \frac{\mu_{1t} - r(t)}{\sigma_{1t}} = \frac{\mu_{2t} - r(t)}{\sigma_{2t}}$$

is called the market price of risk, dependent on both θ_t and t but not on the nature of the derivative V_t.

The general form for λ_t is:

$$\lambda_t = \frac{\mu(\theta_t, t) - r(t)}{\sigma(\theta_t, t)}. \tag{10.14}$$

We can rewrite this as:

$$\mu_t - r(t) = \lambda_t \sigma_t.$$

σ_t can be interpreted as the level of the θ_t-risk in V_t. The risk neutrality implies $\lambda_t = 0$, i.e. $\mu_t = r(t)$ and

$$dV_t = r(t)V_t dt + \sigma_t V_t dW_t.$$

By making other assumptions about the market price of risk we define other "risk" worlds that are internally consistent. If the market price of risk is λ_t, in combination with (10.14), it can be shown that:

$$dV_t = \{r(t) + \lambda_t \sigma_t\}V_t dt + \sigma_t V_t dW_t.$$

10.2.2   Change of Numeraire

A martingale is a zero drift stochastic process:

$$\mathsf{E}_t(\theta_T) = \theta_t.$$

Suppose now that V_t and Z_t are the prices of traded securities dependent on the same source of uncertainty where both securities produce no income during the time under consideration. Define the relative price of V_t w.r.t Z_t as:

$$\phi_t = \frac{V_t}{Z_t}.$$

We refer to Z_t as the numeraire.

Definition 10.7   A numeraire is any non-dividend paying asset.

A numeraire is chosen as to normalize all other asset prices with respect to it. Instead of considering the prices of V_t we are considering the relative prices $\phi_t = V_t/Z_t$.

A convenient choice of the market price of risk in combination with a particular numeraire can lead to a simplification of the valuation problems we encounter in a setting with stochastic interest rates. Combining the market price of risk with a numeraire measure leads to the equivalent martingale measure result.

10.2.3   Equivalent Martingale Measure

For a certain choice of the market price of risk ϕ_t will be a martingale. If we put $\sigma_{Z_t} = \lambda_t$, in combination with (10.14) and under the usual assumptions:

$$dV_t = \{r(t) + \sigma_{V_t}\sigma_{Z_t}\}V_t dt + \sigma_{V_t}V_t dW_t$$

and

$$dZ_t = \{r(t) + \sigma_{Z_t}^2\}Z_t dt + \sigma_{Z_t}Z_t dW_t.$$

Itô's lemma gives:

$$d\log V_t = \left\{r(t) + \sigma_{Z_t}\sigma_{V_t} - \frac{\sigma_{V_t}^2}{2}\right\} dt + \sigma_{V_t}dW_t$$

and

$$d\log Z_t = \left\{r(t) + \frac{\sigma_{Z_t}^2}{2}\right\} dt + \sigma_{Z_t}dW_t$$

so that

$$d(\log V_t - \log Z_t) = \left(\sigma_{Z_t}\sigma_{V_t} - \frac{\sigma_{V_t}^2}{2} - \frac{\sigma_{Z_t}^2}{2}\right) dt + (\sigma_{V_t} - \sigma_{Z_t})dW_t$$

or

$$d \left(\log \frac{V_t}{Z_t} \right) = -\frac{(\sigma_{V_t} - \sigma_{Z_t})^2}{2} dt + (\sigma_{V_t} - \sigma_{Z_t}) dW_t.$$

Again using Itô's Lemma:

$$d \frac{V_t}{Z_t} = (\sigma_{V_t} - \sigma_{Z_t}) \frac{V_t}{Z_t} dW_t,$$

all terms in the drift part cancel out and increments of $\frac{V_t}{Z_t}$ have the expected value equal to 0, $\mathrm{E}_t(d \frac{V_t}{Z_t}) = (\sigma_{V_t} - \sigma_{Z_t}) \frac{V_t}{Z_t} \mathrm{E}_t(dW_t) = 0$. Thus $\phi_t = \frac{V_t}{Z_t}$ is a martingale. A stochastic system where the market price of risk is σ_{Z_t} is referred to as a world that is forward risk neutral w.r.t. Z_t. Because ϕ_t is martingale it follows that:

$$\frac{V_t}{Z_t} = \mathsf{E}_{Z_t} \left(\frac{V_T}{Z_T} \right)$$

or

$$V_t = Z_t \mathsf{E}_{Z_t} \left(\frac{V_T}{Z_T} \right) \tag{10.15}$$

where E_{Z_t} denotes the expected value in a world that is forward risk neutral w.r.t Z_t.

We can generalize this reasoning to different numeraires. Choosing an appropriate numeraire can be helpful when valuing derivatives in the context of stochastic interest rates. In the following we see the basic choices for numeraire and the corresponding choice of the market price of risk, which will lead to an equivalent martingale measure result.

10.2.4 Traditional Risk Neutral Numeraire

A world that is forward risk neutral w.r.t. to the money market account $A(t)$ is a world where the market price of risk is equal to zero. This is the traditional risk neutral world we have considered so far. Denoting A_t as $A(t)$ It follows that:

$$V_t = A_t \mathsf{E}_t \left(\frac{V_T}{A_T} \right)$$

where E_t denotes the expectation in the traditional risk neutral world. In the case of $A_t = 1$ and applying (10.7) this reduces to:

$$V_t = \mathsf{E}_t \left[\exp \left\{ -\int_t^T r(s) \, ds \right\} V_T \right].$$

This is a crucial result. If we consider V_t as a bond $V(t, T)$ it shows that there is a clear relationship, as one might have assumed, between the price of a bond and the term structure of interest rates. However, a different choice of numeraire might be more practical when dealing with interest rate derivatives.

10.2.5   Other Choices of Numeraire

In order to ease the valuation of interest rate derivatives it can be helpful to deviate from the traditional risk neutral world. We will give an overview of the basic numeraire measures that are used in the valuation of interest rate derivatives.

Zero Bond as Numeraire

Setting Z_t equal to $V(t, T)$ will yield another martingale measure. To not confuse notations we will write S_t instead of V_t, where S_t now refers to some traded asset. Because $Z_T = V(T, T) = 1$ and $Z_t = V(t, T)$ we get:

$$S_t = V(t, T)\mathsf{E}_{V_t}(S_T), \qquad (10.16)$$

where we use E_{V_t} to denote the expectation in a world that is forward risk neutral w.r.t. $V(t, T)$.

   This result is helpful when considering the valuation of different interest rate derivatives. It shows that we can value any security that provides a payoff at time T by calculating the expected payoff in a world that is forward risk neutral w.r.t to a bond maturing at time T and discount it by multiplying by the value of $V(t, T)$. It is correct to assume that the expected value of the underlying asset equals its forward value when computing the expected payoff. These results are critical to our understanding of the standard market model for bond options.

Interest Rates with a Bond as Numeraire

Remember that $F(t, T, S)$ is the forward interest rate for the period between T and S as seen at time t. The forward price, as seen at time t, of a zero-coupon bond lasting between times T and S is:

$$V(T, S) = \frac{V(t, S)}{V(t, T)}.$$

It follows that:

$$F(t, T, S) = \frac{1}{\tau(T, S)} \left\{ \frac{V(t, T) - V(t, S)}{V(t, S)} \right\}.$$

Set

$$S_t = \frac{1}{\tau(T, S)} \{ V(t, T) - V(t, S) \}$$

and $Z_t = V(t, S)$. The equivalent martingale measure result shows that $F(t, T, S)$ is a martingale in a world that is forward risk neutral w.r.t. $V(t, S)$. This means that:

$$F(t, T, S) = \mathsf{E}_{V_S} \{ F(T, T, S) \}$$

where E_{V_S} denotes the expectation in a world that is forward risk neutral w.r.t. $V(t, S)$. Combining this result with Eq. (10.15) is decisive to understand the pricing of caps and floors.

Annuity Factor as Numeraire

We can also consider the annuity factor from Eq. (10.21) as a numeraire. This can be helpful when pricing swap options. We can rewrite the solution for the swap rate R_S from Eq. (10.21) as:

$$R_S(t, T) = \frac{1 - V(t, t_n)}{U(t, T)},$$

where

$$U(t, T) = \sum_{i=0}^{n-1} V(t, t_{i+1}) \tau_i.$$

If we now set S_t equal to $1 - V(t, t_n)$ and Z_t equal to $U(t, T)$ this leads to:

$$R_S(t, T) = \mathsf{E}_{U_t} \{ R_S(T, S) \} \tag{10.17}$$

where E_{U_t} denotes the expectation in a world that is forward risk neutral w.r.t. to $U(t, T)$. This result in combination with Eq. (10.15) will be critical to our understanding of the standard market model for European swap options.

This concludes our overview of the choices of numeraire measures and risk neutral pricing. These methods are applied in the valuation of interest rate derivatives in the following part of this chapter.

10.3   Interest Rate Derivatives

This section is devoted to fundamental interest rate products. First we introduce basic products, whose evaluation depends only on the initially given term structure and not on volatilities. These instruments are especially popular for institutional investors with customized cashflow needs. In a further step fundamental derivatives like bond options and swaptions are introduced. We also discuss the standard market formula for pricing simple, basic derivative structures: the Black model.

10.3.1   Forward Rate Agreement

Definition 10.8 A forward rate agreement $\text{FRA}_{R_K,S}\{r(t), t, T\}$ is an agreement at time t that a certain interest rate R_K will apply to a principal amount (for simplicity again equal to 1) for a certain period of time $\tau(T, S)$, in exchange for an interest rate payment at the future interest rates $R(T, S)$, with $t < T < S$.

The current value of a FRA paid-in-arrear is the discounted value of the payoff received at time S.

$$
\begin{aligned}
\text{FRA}_{R_K,S}\{r(t), t, T\} &= \{1 + R(t, S)\tau(t, S)\}^{-1}\tau(T, S)\{R_K - R(T, S)\} \\
&= V(t, S)\{1 + R_K\tau(T, S)\} - \{1 + R(t, T)\tau(t, T)\}^{-1} \\
&= V(t, S)\tau(T, S)R_K + V(t, S) - V(t, T)
\end{aligned}
$$

The payoff is negative if the floating rate is above the fixed rate and in the opposite case the payoff will be positive. When valuing a FRA we are considering three different time instants, namely the current time t, the time at which the FRA will come into place T and the maturity of the FRA S. However, all relevant interest rates can be observed at time t, so no knowledge of the future term structure of interest rates is needed.

Often the strike rate R_K is chosen so that the $\text{FRA}_{R_K,S}\{r(t), t, T\}$ is *fair*, i.e. $\text{FRA}_{R_K,S}\{r(t), t, T\} = 0$, at time t. In this case the strike rate will be equal to $F(t, T, S)$.

10.3.2   Interest Rate Swap

Definition 10.9 An Interest Rate Swap $\text{IRS}_{R_K,T}\{r(t), t\}$ is an agreement to exchange payments of a fixed rate R_K against a variable rate $R(t, t_i)$ over a period $\tau(t, T)$ at certain time points t_i, with $t \le t_i \le T$.

There are two basic types of IRS: a payer IRS and a receiver IRS. In the case of a payer IRS the fixed rate is paid and the floating rate is received. A receiver IRS functions exactly the other way around. The two parts of an IRS can also be referred to as "floating rate leg" and "fixed rate leg".

The value of a receiver IRS $\text{RIRS}_{R_K,T}\{r(t),t\}$ on the rate R_K starting at t and maturing at T with n payments between t and T is given by:

$$\text{RIRS}_{R_K,T}\{r(t),t\} = \sum_{i=0}^{n-1} V(t,t_{i+1})\tau_i \{R_K - R(t_i,t_{i+1})\}$$

$$= \sum_{i=0}^{n-1} V(t,t_{i+1})\tau_i \{R_K - F(t,t_i,t_{i+1})\}$$

$$= \sum_{i=0}^{n-1} V(t,t_{i+1})\tau_i R_K - 1 + V(t,T). \qquad (10.18)$$

where $t_0 = t$ and $t_n = T$, $i = 1, \ldots, n-1$, $t \le t_i \le T$, $\tau_i = \tau(t_i, t_{i+1})$ and $V(t,t) = 1$.

To simplify the valuation of an IRS, an IRS can be thought of as a portfolio of FRAs. By decomposing the IRS into a series of FRAs:

$$\text{RIRS}_{R_K,T}\{r(t),t\} = \sum_{i=0}^{n-1} \text{FRA}_{R_K,t_{i+1}}\{r(t),t,t_i\}$$

Alternatively an IRS can also be valued by considering the fixed and floating rate leg separately. This would correspond to thinking of an IRS as an agreement to exchange a coupon-bearing bond for a floating rate note.

The coupon payments c_i of a coupon bond paying n coupons at a rate of R_K would be:

$$c_i = \tau_i R_K$$

for $i = 1, \ldots, n$. The principal amount is repaid at the maturity of the bond. The value of the fixed leg is therefore the discounted value of the coupon payments plus the value of the principal amount received at time T:

$$\text{FixedLeg}_{R_K}\{r(t),t\} = \sum_{i=0}^{n-1}\{1 + R(t,t_{i+1})\tau_i\}^{-1} c_i + V(t,T)$$

$$= \sum_{i=0}^{n-1} V(t,t_{i+1})R_K\tau_i + V(t,T) \qquad (10.19)$$

For the floating leg we can use the fact that a floating rate note will always be traded at par at the reset dates. So far we have considered time t as the first reset date, therefore if the principal amount is repaid at maturity it follows that:

$$\text{FloatingLeg}\{r(t), t\} = 1. \tag{10.20}$$

The difference between Eqs. (10.19) and (10.20) is the value of the IRS:

$$\text{RIRS}_{R_K, T}\{r(t), t\} = \text{FixedLeg}\{r(t), t\} - \text{FloatingLeg}\{r(t), t\}$$

and we see that is equal to (10.18). In both valuation methods presented above no knowledge of the future term structure is needed because all relevant interest rates are known at time t. As with FRA the strike rate R_K, now referred to as the swap rate R_S, is often chosen so that the IRS has at time t a value of zero.

Definition 10.10 The swap rate $R_S(t, T)$ is the rate that makes the value of an IRS equal to zero at time t.

By setting Eq. (10.18) equal to zero and rearranging:

$$R_S(t, T) = \{1 - V(t, T)\} / \left\{ \sum_{i=0}^{n-1} V(t, t_{i+1}) \tau_i \right\}. \tag{10.21}$$

10.3.3 The Black Model

If the future term structure of interest rates is needed in order to price the derivative we apply the Black (1975) model. This includes caps, floors, swap options and bond options.

Consider a European Call $C_{K,T}(V, t)$ with payoff $\max(V_T - K, 0)$ at time T and K being the strike price.

Assuming:

- The value of the option today is its discounted expected payoff.
- V_T has a lognormal distribution with the standard deviation of $\log V_T$ being $\sigma \sqrt{\tau}$.
- The expected value of V_T at time t is the forward price F_t.

By means of the Black–Scholes framework that we presented in Chap. 6 this implies that:

$$C_{K,T}(V, T) = \mathsf{E}(V_T) \Phi(y + \sigma \sqrt{\tau}) - K \Phi(y)$$

with

$$y = \frac{\log \{\mathsf{E}(V_T)/K\} - \frac{1}{2}\sigma^2 \tau}{\sigma \sqrt{\tau}}.$$

Because interest rates now considered to be stochastic, we discount the expected payoff by multiplying with $V(t,T)$. With $\mathsf{E}(V_T) = F_t$ the value of the option at time t is:

$$C_{K,T}(V,t) = V(t,T)\left\{F_t\Phi(y + \sigma\sqrt{\tau}) - K\Phi(y)\right\} \tag{10.22}$$

with

$$y = \frac{\log\left(\frac{F_t}{K}\right) - \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}}.$$

10.3.4 Bond Option

Definition 10.11 A bond option is an agreement which gives the holder the right to buy or sell a particular bond at a specified time T for a specified strike price K.

As an example consider a European call with a strike price K and a maturity T on a zero bond with a maturity of $S > T$, i.e. the option is to buy the bond at time T at a price K. Therefore:

$$C_{K,T}\{r(t),t\} = \max\{V(T,S) - K, 0\}$$

To value a bond option we will again apply Black's model. We change the numeraire from the current cash amount to a bond $V(t,T)$ and thus we can, in combination with the result from Eq. (10.16), consider that the current value of any security as its expected future value at time T multiplied by $V(t,T)$. It can be shown that the expected value of any traded security at time T is equal to its forward price. Thus the price of an option with maturity T on a bond $V(t,S)$ with $S > T$ is given by:

$$C_{K,T}\{r(t),t\} = V(t,T)\mathsf{E}_T\left[\max\{V(T,S) - K, 0\}\right]$$

and

$$\mathsf{E}_T\{V(T,S)\} = F_t, \tag{10.23}$$

with E_T denoting the forward risk neutral expectation w.r.t to $V(t,T)$ as the numeraire. Again applying the same assumptions as above we find that:

$$C_{K,T}\{r(t),t\} = V(t,T)\left[\mathsf{E}_T\{V(T,S)\}\Phi(y + \sigma\sqrt{\tau}) - K\Phi(y)\right]$$

with

$$y = \frac{\log\left[\mathsf{E}_T\{V(T,S)\}/K\right] - \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}}.$$

By replacing $\mathsf{E}_T\{V(T,S)\}$ in the above equation with (10.23) we achieve equality with Eq. (10.22).

10.3.5   Caps and Floors

Definition 10.12 A Cap contract $\mathrm{Cap}_{R_K}, T\{r(t),t\}$ gives the holder the right to receive the floating interest rate $R(t_i, t_{i+1})$ at certain time points t_i in exchange for the fixed rate R_K, with $i = 1, \ldots, n$.

Definition 10.13 A floor contract $\mathrm{Flr}_{R_K}, T\{r(t),t\}$ gives the holder the right to receive the fixed rate R_K at certain time points t_i in exchange for the floating rate $R(t_i, t_{i+1})$, with $i = 1, \ldots, n$.

Caps and Floors are derivatives which can be used to insure the holder against interest rates rising above or dropping below a certain level, the cap rate or floor rate. Therefore these interest rate derivatives can be used to hedge against increasing or decreasing interest rates.

A cap is a contract that can be viewed as a payer IRS where each exchange payment is executed only if it has a positive value. A floor is equivalent to a receiver IRS where each exchange payment is executed only if it has a positive value.

A cap contract can be decomposed into a series of caplets. Floorlets are defined analogously. Valuing a cap contract can therefore be decomposed into the valuation of single caplets. For a cap contract with n resets, cap rate R_K and time to maturity $\tau(t, T)$ we have:

$$\mathrm{Cap}_{R_K,T}\{r(t),t\} = \sum_{i=1}^{n} \mathrm{Cpl}_i\{r(t),t\},$$

with $i = 1, \ldots, n$, $\tau_i = \tau(t_i, t_{i+1})$ and

$$\begin{aligned}
\mathrm{Cpl}_i\{r(t),t\} &= \{1 + R(t,t_{i+1})\tau(t,t_{i+1})\}^{-1}\tau_i \max\{R(t_i, t_{i+1}) - R_K, 0\} \\
&= \{1 + R(t,t_{i+1})\tau(t,t_{i+1})\}^{-1}\tau_i \max\{F(t,t_i, t_{i+1}) - R_K, 0\}
\end{aligned}$$

$$(10.24)$$

Equation (10.24) shows how caps and floors, essentially being options on interest rates, can be used to hedge against changes in the term structure. A caplet can be interpreted as a call option on the interest rate, while a floorlet would correspond to a put option. Unlike a FRA or IRS it is not possible to determine the value of a cap

(or floor) by knowing only the current term structure at time t. In order to calculate a price for a cap contract we need to apply Black's model.

With the usual assumptions on the strike rate and volatility, and considering a world that is forward risk neutral w.r.t. a bond $V(t, t_{i+1})$ we can consider:

$$\mathsf{E}_{V_{i+1}}\{R(t_i, t_{i+1})\} = F(t, t_i, t_{i+1})$$

The Black model for the ith caplet becomes:

$$\mathrm{Cpl}_i\{r(t), t\} = V(t, t_{i+1})\tau_i \left\{F(t, t_i, t_{i+1})\Phi(y + \sigma\sqrt{\tau_i}) - R_K\Phi(y)\right\} \qquad (10.25)$$

with

$$y = \frac{\log\left\{\frac{F(t, t_i, t_{i+1})}{K}\right\} - \frac{1}{2}\sigma_i^2\tau_i}{\sigma_i\sqrt{\tau_i}}.$$

Again equality with Eq. (10.22) is achieved. Floorlets can either be calculated by using the adapted Put–Call-Parity or by adapting the payoff function in the above equations.

Analogously to the Put–Call-Parity in Chap. 2 it holds that:

$$\mathrm{Cap}_{R_K, T}\{r(t), t\} = \mathrm{Flr}_{R_K, T}\{r(t), t\} + \mathrm{FRA}_{R_K, T}\{r(t), t\},$$

where $\mathrm{Flr}_{R_K, T, n}\{r(t), t\}$ is a floor contract with floor rate R_K and time to maturity $\tau(t, T)$.

10.3.6 Swaption

Definition 10.14 A European swap option or swaption $\mathrm{SWP}_{R_K, T}\{r(t), t)\}$ is an option giving the right to enter an IRS at a given future time T with a specified rate R_K lasting until S.

A market participant will only exercise this option if the market swap rate at maturity of the swaption is less favourable. Therefore, a swap option is essentially an option on the forward swap rate $R_S(T, S)$.

Like with IRS, we can distinguish between payer and receiver swaptions. The holder of a payer swaption has the right, but not the obligation to pay a fixed in exchange for a variable interest rate. The holder of the equivalent receiver swaption has the right, but not the obligation to receive interest at a fixed rate and pay the variable interest.

The value of a payer swaption at time t of is:

$$\text{SWP}_{R_K,T}\{r(t),t\} = V(t,T) \max\left[\sum_{i=0}^{n-1} V(T,T_{i+1})\tau_i\{R_S(T,S) - R_K\}, 0\right],$$

with $T \leq T_i \leq S$.

To determine the value of a swaption we again use the Black model. We consider a world that is risk neutral w.r.t. the numeraire measure $U(t,T)$ from Sect. 10.2.5. If we apply the usual assumptions on the distribution of swap rates and volatility we can show that:

$$\text{SWP}_{R_K,T}\{r(t),t\} = U(t,T)\mathsf{E}_{U_t}[\max\{R_S(T,S) - R_K, 0\}]$$

Thus by using the Black model the time t value is:

$$\text{SWP}_{R_K,T}\{r(t),t\} = U(t,T)[\mathsf{E}_{U_t}\{R_S(T,S)\}\Phi(y + \sigma\sqrt{\tau}) - R_K\Phi(y)],$$

with

$$y = \frac{\log\left[\mathsf{E}_{U_t}\{R_S(T,S)\}/R_K\right] - \frac{1}{2}\sigma^2\tau}{\sigma\sqrt{\tau}}.$$

Applying Eq. (10.17) we replace the expected future swap rate with the current forward swap rate and achieve equality with Eq. (10.22).

Note that the different versions of the Black model we have seen above are defined by different measures. Therefore the Black model for caps is not consistent with the approach to price swap or bond options and vice versa. This is due to the assumptions on the distribution of the underlying interest and swap rates. However, this fact is neglected in practice and remains one of the critical points in applying the Black model.

It is to be noted that the above tools are only of basic character. As with equity options, exotic interest rate derivatives exist. The Black formula, with no standard stochastic process for the interest rate evolution, cannot be applied.

10.4 Interest Rate Modeling

A good model of interest rates dynamics should be able to reproduce the historical stylized facts concerning the average shape of the yield curve. We prefer parsimonious models that are possibly analytically tractable and easy to calibrate. The empirical study suggests that any suitable model should satisfy the following interest rates stylized facts:

- Mean reversion: interest rates tend to return to an average level, as time grows.
- Interest rates should be non-negative.
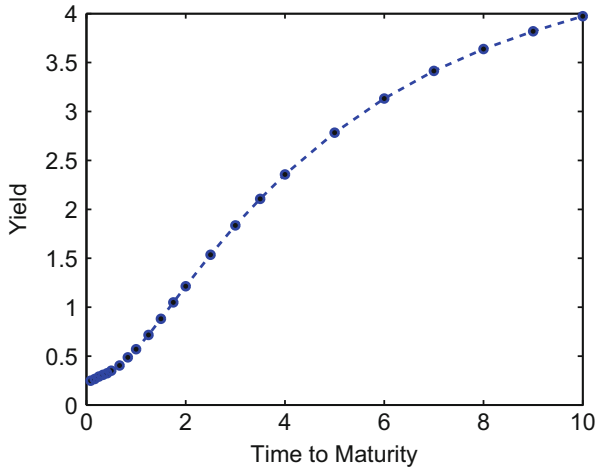
Fig. 10.1 Zero-coupon U.S Treasury bond yields with maturities from 1 month up to 10 years
observed on 30 March 2010   SFEusycplot

- Term structure of interest rates assumes a variety of shapes through time,
  including upward sloping, downward sloping, humped and inverted humped.

To satisfy those goals researchers have developed a vast literature with a wide
variety of models, the most popular of them are presented here. It is worth noting
that in times of recent financial distress (2008–2012 the European sovereign-debt
crisis) some above mentioned interest rates properties (e.g. non-negativity) were
violated. The bond and lending market turned out to be risky business, a fact
that stays in contradiction to the risk-free theoretical framework introduced before
(Fig. 10.1).

10.4.1   Short Rate Models

The first interest-rate modelling originated from the assumption of specific one-
dimensional dynamics for the short rate $r(t)$. Stochastic models of the instantaneous
spot rate $r(t)$ allow to define all fundamental quantities (rates and bonds) in terms
of properties of the process $r(t)$, see (10.8). In general, the dynamic of the short rate
is defined as the Itô Process (10.9):

$$dr(t) = \mu\{r(t), t\}dt + \sigma\{r(t), t\}dW_t,$$

where W_t represents as usual a standard Wiener process. Practice shows that $r(t)$
does not follow a geometric Brownian motion, so that the Black–Scholes approach
cannot be used.

Table 10.1   One-factor short rate models

Vasicek	$dr(t) = a\{b - r(t)\}dt + \sigma \, dW_t$
CIR	$dr(t) = a\{b - r(t)\}dt + \sigma \sqrt{r(t)}dW_t$
Ho–Lee	$dr(t) = \delta(t)dt + \sigma \, dW_t$
Hull–White one-factor	$dr(t) = \{\delta(t) - ar(t)\}dt + \sigma \, dW_t$

The short rate models can be divided into two main groups: equilibrium and arbitrage-free models. For the equilibrium approach today's term structure of interest rate is endogenously derived by the model. In the no-arbitrage approach today's term structure of interest rate is an input to the model.

One-Factor Short-Rate Models

A first approach to model the term structure of interest rates are the one factor short rate models. One factor short rate models consider only one factor of uncertainty in the dynamics of the interest rate (Table 10.1).

Vasicek Model

Vasicek (1977) introduced an interest rate model as:

$$dr(t) = a\{b - r(t)\}dt + \sigma \, dW_t$$

where a, b and σ are constants, W_t is a Wiener process. It is consistent with the mean reversion feature of the interest rate at a reversion rate a to the level b. The Vasicek model allows $r(t)$ to be negative.

Cox, Ingersoll and Ross (CIR) Model

Cox et al. (1985) proposed an alternative model from Vasicek as:

$$dr(t) = a\{b - r(t)\}dt + \sigma \sqrt{r(t)}dW_t \qquad (10.26)$$

where a, b and σ are constants, W_t is a standard Wiener process. The disadvantage of possible negative $r(t)$ is avoided here. The drift part does not differ from the Vasicek model. However, $\sqrt{r(t)}$ is included in the diffusion process as a proportion of the standard deviation. Therefore $r(t)$ has a positive impact on the standard deviation through this setting.

Ho–Lee Model

The former three models are all equilibrium models. The equilibrium approach yields today's term structure as output and is adapted to fit the term structure by choosing proper parameters. Sometimes it is difficult to find the parameters to fit today's term structure quite well. In order to overcome this problem the no-arbitrage approach was introduced. Other than the equilibrium approach, it takes today's term structure as an input to ensure that the model fits today's term structure perfectly by imposing a time function in the drift part.

Ho and Lee (1986) presented the first no-arbitrage model as:

$$dr(t) = \delta(t)dt + \sigma dW_t$$

where σ is constant, $\delta(t)$ is a deterministic function of time and W_t is a Wiener process. The time dependent variable $\delta(t)$ defines the trend of $r(t)$ at time t. The Ho–Lee model lacks mean reversion and $r(t)$ can be negative.

Hull–White Model

We have already discussed that it can be difficult to fit the Vasicek model to the initial term structure of interest rates. Hull and White (1990) proposed an extended Vasicek model to address this problem. The model is:

$$dr(t) = \{\delta(t) - ar(t)\}dt + \sigma dW_t$$

where a and σ are constants, $\delta(t)$ is a deterministic function of time, W_t is a Wiener process. Compared to the Vasicek model, it uses the time-dependent reversion level $\delta(t)/a$ instead of the constant b in Vasicek. It is also a special case of the Ho–Lee model with a mean reversion rate a.

Two-Factor Short-Rate Models

One-factor models imply that the instantaneous rates for all maturities in the yield curve are perfectly correlated which means that a shock on $r(t)$ at time t will transmit rigidly to all maturities in the curve. This property is clearly unrealistic. Yield curves observed on the market show a variety of different shapes and very different reactions to the market shocks. A more satisfactory method to model the interest rate process is needed. A possible improvement comes with involving more explanatory factors into the short rate model. We briefly introduce two representative two-factor models.

Longstaff–Schwartz Model

Longstaff and Schwartz (1992) developed a two-factor equilibrium model that is based on the CIR framework where $r(t)$ is a linear combination of X_t and Y_t as:

$$r_t = \alpha X_t + \beta Y_t$$

$$dX_t = (a - bX_t)dt + \sqrt{X_t}dW_{1t}$$

$$dY_t = (e - fY_t)dt + \sqrt{Y_t}dW_{2t}$$

a, b, e, $f > 0$, X and Y are state variables, W_{1t} and W_{2t} are Wiener process. The two factors are the short-term interest rate and the volatility of the short-term interest rate increment. This feature makes the contingent claim values to reflect both the current interest rate level and the interest rate volatility level. This model is proved to be quite tractable.

Hull–White Two-Factor Model

Hull and White (1994) presented a no-arbitrage two-factor model which assumed the short rate following the process:

$$dr(t) = \{\delta(t) + u(t) - ar(t)\}dt + \sigma_1 dW_{1t}$$

where

$$du(t) = -bu(t)dt + \sigma_2 dW_{2t}, \quad u(0) = 0$$

a, b, σ_1 and σ_2 are constants, W_{1t} and W_{2t} are Wiener processes, $dW_{1t}dW_{2t} = \rho dt$. The two factors are the short-term and long-term interest rates. $\delta(t)$ is deterministic and can be properly chosen to exactly fit the initial term structure. $u(t)$ is a reversion level component which mean reverts to zero.

10.4.2   *Heath Jarrow Morton Framework*

We have just seen a number of different possibilities to model the evolution of interest rates by means of modeling the short rate process. A main drawback of short rate models is its disability to reproduce the basket of shapes of the term structure. Moreover those simple models have a difficulty to calibrate to the current yield curve and an unrealistic presentation of the variance and covariance structure. To this end, Heath et al. (1992) have derived an arbitrage-free framework for a stochastic evolution of the entire yield curve from an instantaneous forward rate.

HJM Approach

By Eq. (10.6) we know that:

$$f(t, T) = -\frac{\partial \log V(t, T)}{\partial T}.$$

In a discrete time setting with $S = T + \Delta$ this will become

$$F(t, T, T + \Delta) = \frac{\log V(t, T + \Delta) - \log V(t, T)}{\Delta},$$

which corresponds to Eq. (10.4). As a bond is a traded security, its price development can be expressed as a SDE.

$$dV_t = \mu(t, T, V_t)V_t dt + \sigma(t, T, V_t)V_t dW_t,$$

where $V_t = V(t, T)$, $\sigma(t, T, V_t)$ is the volatility of $V(t, T)$ and W_t is a Wiener process w.r.t. the real world measure. Other than in the Black–Scholes world, the diffusion process depends on $V(t, T)$ and the Wiener process governing the SDE can be different for every maturity. Therefore every bond with different maturity can theoretically be driven by a different diffusion process.

As in the Black–Scholes framework, the drift coefficient can be modified according to Girsanov's theorem. In the risk neutral world the SDE becomes:

$$dV_t = r(t)V_t dt + \sigma(t, T, V_t)V_t dW_t^*, \qquad (10.27)$$

where W_t^* is now a Wiener Process under the risk neutral measure.

In order to find the dynamics for $F(t, T, T + \Delta)$ we apply Itô's Lemma to Eq. (10.27) and get:

$$d \log V(t, T) = \left\{ r(t) - \frac{1}{2}\sigma(t, T, V_t)^2 \right\} dt + \sigma(t, T, V_t)dW_t^*$$

and

$$d \log V(t, T + \Delta) = \left\{ r(t) - \frac{1}{2}\sigma(t, T + \Delta, V_t)^2 \right\} dt + \sigma(t, T + \Delta, V_t)dW_t^*.$$

Thus

$$dF(t, T, T + \Delta) = \frac{1}{2\Delta} \left[\sigma\{t, T + \Delta, V(t, T + \Delta)\}^2 - \sigma(t, T, V_t)^2 \right] dt$$

$$+ \frac{1}{\Delta} \left[\sigma\{t, T + \Delta, V(t, T + \Delta)\} - \sigma(t, T, V_t) \right] dW_t^*$$

For $\Delta \to 0$ we get the dynamics of the instantaneous forward rate.

$$df(t,T) = \sigma(t,T,V_t)\left\{\frac{\partial\sigma(t,T,V_t)}{\partial T}\right\} dt$$

$$+ \left\{\frac{\partial\sigma(t,T,V_t)}{\partial T}\right\} dW_t^*,$$

where $\sigma(\cdot)$ are the bond price volatilities. This can simplified to:

$$df(t,T) = \alpha(t,T)dt + \beta(t,T)dW_t^*,$$

where

$$\alpha(t,T) = \sigma(t,T,V_t)\left\{\frac{\partial\sigma(t,T,V_t)}{\partial T}\right\}$$

and

$$\beta(t,T) = \left\{\frac{\partial\sigma(t,T,V_t)}{\partial T}\right\}.$$

If the volatility term structure in the form of $\sigma(t,T,V_t)$ is developed from the underlying data set, the risk neutral process for $f(t,T)$ is known. By defining the volatility term structure accordingly we can formulate all of the short rate models considered in Sect. 10.4.1 within the HJM framework.

The advantages of the HJM are that it permits a large number of possible assumptions about the evolution of the yield curve and that the resulting models will, by definition, be consistent with the initial term structure that is observed in the market.

Short Rate Process in the HJM Framework

By integrating the process $df(t,T)$ we get:

$$f(t,T) = f(0,T) + \int_0^t \alpha(s,T)\,ds + \int_0^t \beta(s,T)\,dW_s^*$$

We can now set $T = t$ to receive the short rate $r(t)$ as:

$$r(t) = f(0,t) + \int_0^t \alpha(s,T)\,ds + \int_0^t \beta(s,t)\,dW_s^*.$$

This is somewhat problematic and poses the biggest problem of HJM. The process for short rate is in general no longer Markov as the drift term for the short rate process is now a function of all past volatilities. Only with a number of selected models, such as the Hull–White model or the Black, Derman and Toy model, the process for the short rate becomes a Markov process. In order to use the HJM framework in practice it requires Monte Carlo simulation.

Like with models for the short rate process, the HJM approach can be modified to incorporate multiple factors so to include, e.g. macroeconomic variables. This is however beyond the scope of this chapter. The interested reader is referred to either Heath et al. (1992) or to chapter 5 of Brigo and Mercurio (2001).

10.4.3 LIBOR Market Model

The term structure models we introduced before have a common drawback that neither the instantaneous spot rate nor the instantaneous forward rate can be directly observed in the market. Hence they are not compatible to price caps and swaptions with Black's formula. An alternative was proposed by Brace et al. (1997), Jamshidian (1997) and Miltersen et al. (1997) who modeled LIBORs instead of instantaneous rates. This approach is known as the LIBOR market model (LMM).

Definition 10.15 The London Interbank Offered Rate (LIBOR), $L_n(t)$, is the forward rate over the period $[t_n, t_{n+1}]$ as observed at time t with compounding period $\tau(t_n, t_{n+1})$.

Dynamics in the LMM

The relation between a zero-bond and the LIBOR forward rate is defined as:

$$1 + \tau_n L_n(t) = \frac{V_n(t)}{V_{n+1}(t)}.$$

where we define t_n, $n = 0, 1, \ldots, M$ to be the times at which M assets are traded in the market, $t_0 = 0 < t_1 < t_2 \ldots < t_{M+1}$ and $\tau_n = t_{n+1} - t_n$. Consider $V_n(t)$ as a bond maturing at time $t_n > t$. It is straightforward that the LIBOR forward rate can be represented as:

$$L_n(t) = \frac{1}{\tau_n} \left\{ \frac{V_n(t) - V_{n+1}(t)}{V_{n+1}(t)} \right\}.$$

This is similar to Eq. (10.5), with the simple difference that the forward rate $F(t, T, S)$ is replaced by the LIBOR forward rate $L_n(t)$.

LMM models the forward Libor rate $L_n(t)$ as a martingale by a (driftless) geometric Brownian:

$$\frac{dL_n(t)}{L_n(t)} = \sigma_n(t)dW^n, \tag{10.28}$$

where:

dW^n is a Wiener process under a measure Q^{n+1} defined by the numeraire $V_{n+1}(t)$.

$\sigma_n(t)$ is an instantaneous volatility of $L_n(t)$ at time t.

The Numeraire Measure

It is worth noting that $L_n(t)$ is a martingale only under the measure Q^{n+1}. In order to evolve all the LIBOR rates under the same measure, we consider a change of measure. The change leads to an additional drift term $\mu_n(t)$ in (10.28). If there are M LIBOR rates in the economy, we use the so called terminal measure which takes the numeraire under the measure Q^{M+1} associated with a bond $V_{M+1}(t)$. Under Q^{M+1} the (relative) price of $V_n(t)$ in an arbitrage free model has a drift free dynamics and follows:

$$\frac{V_n(t)}{V_{M+1}(t)} = \prod_{k=n}^{M} \frac{V_k(t)}{V_{k+1}(t)} = \prod_{k=n}^{M} \{1 + \tau_k L_k(t)\}. \tag{10.29}$$

By Itô lemma for a product of any Itô processes X and Y

$$d(X \cdot Y) = X \cdot dY + Y \cdot dX + dX \cdot dY, \tag{10.30}$$

applying iteratively to (10.29) yields

$$d\left[\prod_{k=n}^{M} \{1 + \tau_k L_k(t)\}\right] = \prod_{k=n}^{M} \{1 + \tau_k L_k(t)\} \cdot$$
$$\cdot \sum_{j=n}^{M} \left\{ \frac{\tau_j dL_j(t)}{1 + \tau_j L_j(t)} + \sum_{l=j+1}^{M} \frac{\tau_j dL_j(t)}{1 + \tau_j L_j(t)} \frac{\tau_l dL_l(t)}{1 + \tau_l L_l(t)} \right\}. \tag{10.31}$$

The drift free dynamics of $\prod_{k=n}^{M} \{1 + \tau_k L_k(t)\}$ under Q^{M+1} imply

$$\sum_{j=n}^{M} \underset{Q^{M+1}}{Drift} \left\{ \frac{\tau_j dL_j(t)}{1 + \tau_j L_j(t)} + \sum_{l=j+1}^{M} \frac{\tau_j dL_j(t)}{1 + \tau_j L_j(t)} \frac{\tau_l dL_l(t)}{1 + \tau_l L_l(t)} \right\} = 0. \tag{10.32}$$

It can be easily shown that for all $j = 1, \ldots, M$

$$\underset{Q^{M+1}}{Drift} \left\{ \frac{\tau_j dL_j(t)}{1 + \tau_j L_j(t)} + \sum_{l=j+1}^{M} \frac{\tau_j dL_j(t)}{1 + \tau_j L_j(t)} \frac{\tau_l dL_l(t)}{1 + \tau_l L_l(t)} \right\} = 0. \qquad (10.33)$$

The correlation between $L_n(t)$ and $L_j(t)$ is denoted by $\rho_{n,j}$,

$$dL_j(t) \cdot dL_n(t) = L_j(t) L_n(t) \sigma_j(t) \sigma_n(t) \rho_{n,j} dt. \qquad (10.34)$$

Combining Eqs. (10.33) and (10.34) the drift part of $L_n(t)$ under Q^{M+1} is

$$\mu_n(t) = - \sum_{l=n+1}^{M} \frac{\tau_l L_l(t)}{1 + \tau_n L_l(t)} \sigma_l(t) \sigma_n(t) \rho_{n,l} \qquad (10.35)$$

and the dynamics can be represented as:

$$\frac{dL_n(t)}{L_n(t)} = - \sum_{l=n+1}^{M} \frac{\tau_l L_l(t)}{1 + \tau_n L_l(t)} \sigma_l(t) \sigma_n(t) \rho_{n,l} dt + \sigma_n(t) dW^{M+1}(t). \qquad (10.36)$$

The most convenient choice of the numeraire (measure) depends on the derivative one is pricing. Repeating the above arguments for the spot measure (Q^0) yields

$$\frac{dL_n(t)}{L_n(t)} = \sum_{l=n+1}^{M} \frac{\tau_l L_l(t)}{1 + \tau_n L_l(t)} \sigma_l(t) \sigma_n(t) \rho_{n,l} dt + \sigma_n(t) dW^0(t). \qquad (10.37)$$

Comparing to the HJM model, the HJM model is the limit case when τ_n in the LIBOR model tends to zero. In the process of calibration, $\sigma_n(t), n = 1, \ldots, M$ as the deterministic functions are chosen to resemble the Black implied volatilities as good as possible. To price the caplets and swaptions by using the LIBOR model, Monte Carlo simulation has proven to be the most reliable method. The implementation is introduced in Sect. 10.6.

10.5   Bond Valuation

10.5.1   The Bond Valuation Equation

In this part, we take the simple one-factor short-rate model as an example to show how to derive a closed form of the bond valuation equation. Under the corresponding

numeraire, the value of a bond can be represented as:

$$V(t,T) = \mathsf{E}_t\left[\exp\left\{-\int_t^T r(s)ds\right\}V(T,T)\right].$$

We already saw that the general Itô Process of $r(t)$ can be written as:

$$dr(t) = \mu_r dt + \sigma_r dW_t, \tag{10.38}$$

where $\mu_r = \mu\{r(t), t\}$ and $\sigma_r = \sigma\{r(t), t\}$. Under the condition that $V(T, T) = 1$ and in combination with Itô's Lemma we can write:

$$dV(t,T) = \frac{\partial V(t,T)}{\partial t}dt + \frac{1}{2}\sigma^2\frac{\partial^2 V(t,T)}{\partial r^2}dt + \frac{\partial V(t,T)}{\partial r}dr(t).$$

If we plug in $dr(t)$ from Eq. (10.38), we get:

$$dV(t,T) = \left\{\frac{\partial V(t,T)}{\partial t} + \frac{1}{2}\sigma^2\frac{\partial^2 V(t,T)}{\partial r^2} + \mu_r\frac{\partial V(t,T)}{\partial r}\right\}dt + \sigma\frac{\partial V(t,T)}{\partial r}dW_t.$$

Under the risk-neutral measure, the PDE can thus be written as:

$$r(t)V(t,T) = \frac{\partial V(t,T)}{\partial t} + \frac{1}{2}\sigma_r^2\frac{\partial^2 V(t,T)}{\partial r^2} + \mu_r\frac{\partial V(t,T)}{\partial r}.$$

10.5.2   Solving the Zero Bond Valuation

We take the CIR model as an example (see Fig. 10.2). The PDE of the CIR model under the risk-neutral measure is

$$r(t)V(t,T) = \frac{\partial V(t,T)}{\partial t} + \frac{1}{2}r\sigma^2\frac{\partial^2 V(t,T)}{\partial r^2} + a(b-r)\frac{\partial V(t,T)}{\partial r}. \tag{10.39}$$

Assume the bond value $V(t,T) = \exp\{A(t) - rB(t)\}$ with a nominal value 1 EUR. Due to this we can consider the following:

$$\frac{\partial V(t,T)}{\partial t} = \{A'(t) - rB'(t)\}V(t)$$

$$\frac{\partial V(t,T)}{\partial r} = -B(t)V(t)$$

$$\frac{\partial^2 V(t,T)}{\partial r^2} = B^2(t)V(t)$$
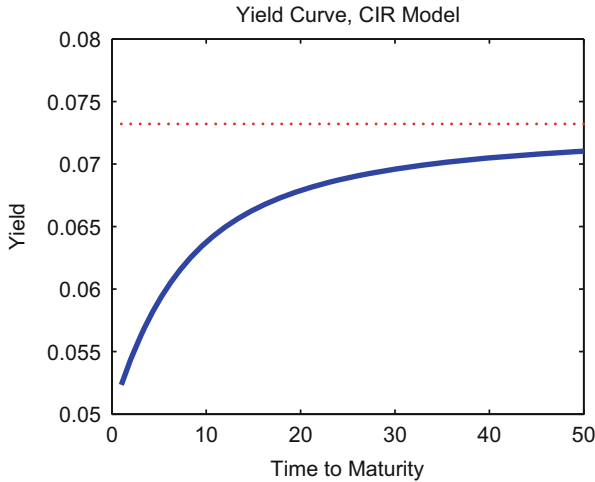
Fig. 10.2 Term structure according to the CIR model with $a = b = \sigma = 0.1$, $r = 0.05$

Q SFEcir

If we plug in these three functions into the PDE of Eq. (10.39) which we obtained above, then we get:

$$0 = \{A'(t) - B'(t)r(t)\}V(t) + \frac{1}{2}\sigma^2 r V(t)B^2(t)$$
$$-a(b-r)B(t)V(t) - r(t)V(t)$$
$$A'(t) = abB(t) - \sigma^2 r B^2(t)/2$$
$$B'(t) = aB(t) - 1$$

With the boundary condition $V(T, T) = 1$, $A(T, T) = B(T, T) = 0$, we can derive an explicit expression of:

$$V(t, T) = \exp\{A(t) - rB(t)\} \tag{10.40}$$

where

$$A(t) = \frac{2ab}{\sigma^2} \log \frac{2\psi \exp\{(a + \psi)(T - t)/2\}}{2\psi + (a + \psi)\exp\{\psi(T - t) - 1\}}$$

$$B(t) = \frac{2\exp\{\psi(T - t) - 1\}}{2\psi + (a + \psi)\exp\{\psi(T - t) - 1\}}$$

$$\psi = \sqrt{a^2 + 2\sigma^2}.$$

10.6   Calibrating Interest Rate Models

We conclude this chapter by a basic empirical application of the CIR and LIBOR Market models to real data. For the CIR model we use the maximum likelihood estimator (MLE) in order to estimate the underlying parameters. To achieve convergence to the global optimum we choose our starting values by considering the least squares estimator as proposed by Overbeck and Rydén (1997). Implementation of the LMM is done using Monte Carlo simulation of the SDE (10.37). The first step is to discretize the forward rate equation. As the LMM is driven by the instantaneous volatility, calibration of the LMM essentially refers to the calibration of the instantaneous volatility function. The goal is to assure that the modelled volatility resembles the Black implied volatility as good as possible. We calibrate the LMM to caplets.

10.6.1   CIR Model: Estimation

For maximum likelihood estimation of the parameter vector $\theta \stackrel{\text{def}}{=} (a, b, \sigma)$ transition densities are required. The CIR process is one of the few cases where the transition density has a closed form expression. Cox et al. (1985) show that the density of $r_{t+\Delta t}$ at time $t + \Delta t$ is:

$$p(r_{t+\Delta t}|r_t, \theta, \Delta t) = c \exp(-u - v) \left(\frac{v}{u}\right)^{\frac{q}{2}} I_q(2\sqrt{uv}),$$

where

$$c = \frac{2a}{\sigma^2\{1 - \exp(-a\Delta t)\}},$$

$$u = cr_t \exp(-a\Delta t),$$

$$v = cr_{t+\Delta t},$$

$$q = 2ab/\sigma^2 - 1,$$

and I_q is the modified Bessel function of the first order q.

Initial Estimates

The success of the MLE approach is dependent on the availability of good starting values for the numerical optimization algorithm. We choose the starting values by means of a conditional least squares estimation as applied by Overbeck and Rydén

(1997). Following the notation we have used above, the conditional mean function for the CIR model is derived as:

$$m(r; \theta) = \mathsf{E}_\theta(r_t | r_{t-1} = r) = \gamma_0 + \gamma_1 r$$

with

$$\gamma_0 = -b\{\exp(-a\Delta t) - 1\}$$

and

$$\gamma_1 = \exp(-a\Delta t).$$

Overbeck and Rydén (1997) show that the conditional least squares estimators for a and b are given by:

$$\hat{a} = -\frac{1}{\Delta t}\left[\left\{n^{-1}\sum_{t=1}^{n}(r_t - \bar{r}_n)(r_{t-1} - \bar{r}'_n)\right\} / \left\{n^{-1}\sum_{t=1}^{n}(r_{t-1} - \bar{r}'_n)^2\right\}\right]$$

and

$$\hat{b} = -\frac{\bar{r}_n - \exp(-a\Delta t)\bar{r}'_n}{\exp(-a\Delta t) - 1},$$

where $\bar{r}_n = n^{-1}\sum_{t=1}^{n} r_t$ and $\bar{r}'_n = n^{-1}\sum_{t=1}^{n} r_{t-1}$.

The estimator for b is based on the conditional second moment function which is given by:

$$v(r; \theta) = \mathsf{E}_\theta[\{r_t - \mathsf{E}_\theta(r_t | r_{t-1} = r)\}^2 | r_{t-1} = r] = \sigma^2(\eta_0 + \eta_1 r)$$

with

$$\eta_0 = \frac{b}{2a}\{\exp(-a\Delta t) - 1\}^2$$

and

$$\eta_1 = -\frac{1}{a}\exp(-a\Delta t)\{\exp(-a\Delta t) - 1\}.$$

As an estimator for sigma we use:

$$\hat{\sigma}^2 = n^{-1}\sum_{t=1}^{n}\frac{\{r_t - m(r_{t-1}; \hat{a}, \hat{b})\}^2}{\hat{\eta}_0 + \hat{\eta}_1 r_{t-1}}$$

where $\hat{\eta}_0$ and $\hat{\eta}_1$ are evaluated at $(\hat{a}, \hat{b})$.

Maximum Likelihood Estimator

The likelihood function for interest rate time series with n observations is:

$$L(\theta) = \prod_{t=1}^{n-1} p(r_{t+1}|r_t, \theta, \Delta t).$$

Thus the log-likelihood function is:

$$\log L(\theta) = \sum_{t=1}^{n-1} \log p(r_{t+1}|r_t, \theta, \Delta t).$$

The log-likelihood function of the CIR process is given by:

$$\log L(\theta) = (n-1)\log c$$
$$+ \sum_{t=1}^{n-1}\left[-u_t - v_{t+1} + 0.5q\log\frac{v_{t+1}}{u_t} + \log\{I_q(2\sqrt{u_t v_{t+1}})\}\right],$$

$$\tag{10.41}$$

where $u_t = cr_t\exp(-a\Delta t)$, and $v_{t+1} = cr_{t+1}$. We find the MLE $\hat\theta$ by maximizing the log-likelihood function in Eq. (10.41) over its parameter space.

$$\hat\theta = (\hat a, \hat b, \hat\sigma) = \arg\max_\theta \log L(\theta).$$

10.6.2   CIR Model: Implementation Results

Our empirical work uses daily observations of the annualized yield on U.S. Treasury Bills with 3 months to maturity. The series was constructed from a daily series available from the Federal Reserve. We have 3,754 observations, ranging from 02 January 1998 to 31 December 2012. We consider the process for $r(t)$ to be given by Eq. (10.26). Our objective is to estimate the parameters a, b and σ from the observations of $r(t)$ at time intervals of Δt by MLE and to simulate a CIR process based on these results (Fig. 10.3).

The CIR process is calibrated to the analyzed data, and the MLE estimates are displayed in Table 10.2. The estimates are applied for pricing of a bond by Eq. (10.40). The results are presented in Table 10.3. It is worth noting that in the same spirit one can evaluate any derivative based on our underlying process. For simple instruments closed analytical formula may be derived.
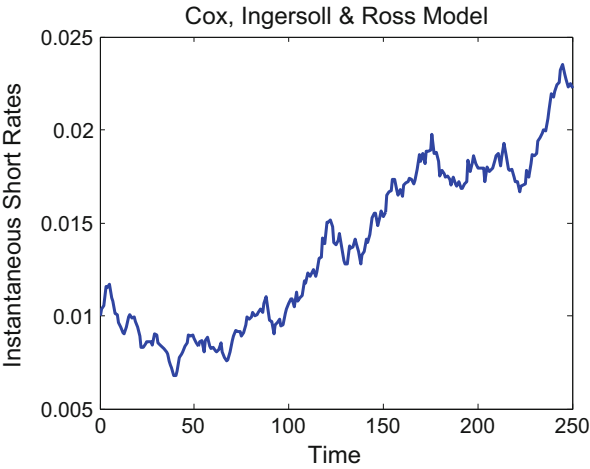
Fig. 10.3 Simulated CIR process with $a = 0.221$, $b = 0.02$, $\sigma = 0.055$ and $r_0 = 0.01$

Q SFEsimCIR

Table 10.2 MLE results for the CIR model with 3,754 observations from 19980102 to 20121231   Q SFECIRmle

Number of observations n	3,754
Time step Δt	1/252
Estimated $\hat{a}$	0.111
Estimated $\hat{b}$	0.006
Estimated $\hat{\sigma}$	0.088

Table 10.3 Results of bond pricing with the CIR model   Q SFECIRpricing

Face value	1 EUR
Time to maturity T	3 months
Short rate at time t	0.02
Bond price $V(t, T)$	0.99 EUR

10.6.3 LMM: Discretization of the Forward Rate

For the LMM under the spot measure it is not possible to derive known transition densities. Therefore the model dynamics have to be discretized in order to perform simulation. In line with Glasserman (2004), application of the Euler method to the SDE of the LIBOR market model, see (10.37), leads to:

$$L_n(t_{i+1}) = L_n(t_i) \sum_{l=n+1}^{M} \frac{\tau_j L_j(t_i)\sigma_n(t_i)\sigma_j(t_i)\rho_{n,j}}{1 + \tau_j L_j(t_i)} L_n(t_i)\tau_i + L_n(t_i)$$
$$\times \sqrt{\tau_i}\sigma_n(t_i)\varepsilon_{i+1},$$

where $\varepsilon_1, \varepsilon_2, \ldots$ are independent $N(0, 1)$ distributed variables. Given a set of bond prices we can initialize the simulation with:

$$L_n(0) = \frac{V_n(0) - V_{n+1}(0)}{\tau_n V_{n+1}(0)},$$

where $n = 1, \ldots, M$. However, this approach can lead to negative rates. An alternative can be to apply the Euler scheme to $\log L_i$. This approach leads to the following approximates for the LIBOR rates under the spot measure, where all rates produced are positive:

$$L_n(t_{i+1}) = L_n(t_i) \exp \left\{ \sum_{j=n+1}^{M} \frac{\tau_j L_j(t_i)\sigma_n(t_i)\sigma_j(t_i)\rho_{n,j}}{1 + \tau_j L_j(t_i)} - \frac{1}{2}\sigma_n(t_i)^2 \right\}$$
$$\times \tau_i + \sqrt{\tau_i}\sigma_n(t_i)\varepsilon_{i+1}.$$

This corresponds to approximating the LIBOR rates by a geometric Brownian motion over $[t_i, t_{i+1}]$, with drift and volatility parameters fixed at t_i.

10.6.4   LMM: Instantaneous Volatility Function

After LIBOR discretization one needs to determine the instantaneous volatility function. There are several suggestions for both parametric and non-parametric forms of the instantaneous volatility function in the literature. General requirements, as suggested by Rebonato (2002), to the functional form of the instantaneous volatility function are that:

- the chosen form should be able to reproduce either a monotonically decreasing or a humped shaped volatility
- the parameters should be economically interpretable.

As shown in Rebonato (2004), the following parametric form fulfils these criteria:

$$\sigma_n(t) = g(t_n)f(t_n - t) = k_n[\{a + b(t_n - t)\} \exp\{-c(t_n - t)\} + d], \qquad (10.42)$$

where $g(t_n)$ is a function of the individual forward rate and $f(t_n - t)$ is a time-homogeneous component. This type of instantaneous volatility function allows a humped shaped form for the instantaneous volatility and includes the possibility to modify the volatility for each maturity separately.

For a more thorough treatment of the different possibilities for the choice of the instantaneous volatility function we refer to Brigo and Mercurio (2001), Chap. 6. When calibrating this function to the given data set, we must make sure that the parameters a, b, c and d are chosen such that the volatility function is well-behaved. Therefore the following conditions have to be given:

- $a + d > 0$
- $d > 0$
- $c > 0.$

10.6.5 LMM: Implementation Results

The parameters are chosen so that the instantaneous volatility function is consistent with the implied volatilities of the Black model. We infer the implied volatilities from a series of N traded caplets. We know by Eq. (10.28) that under the correct measure we can express the forward rate process as a driftless diffusion process:

$$\frac{dL_n(t)}{L_n(t)} = \sigma_n(t)dW^{n+1}(t),$$

where the instantaneous volatility function is connected with the average volatility of the Black model by:

$$(\sigma^{Black})^2 T = \int_0^T \sigma(t)^2 dt.$$

We therefore choose the instantaneous volatility function as:

$$(\sigma_n^{Black})^2 t_n = \int_0^{t_n} \sigma_n^2(s)ds,$$

where σ_n^{Black} is the Black implied volatility for the caplet associated with the forward rate L_i.

If we want to price N different caplets we start by calibrating the time-homogeneous part. To do this, first set $g(t_n) = 1$ for $n = 1, \ldots, N$. We perform a least-squares minimization of the following equation:

$$\min \sum_{i=1}^N \left\{ (\sigma_i^{Black})^2 t_i - \int_0^{t_i} f^2(t_i - s)ds \right\}^2.$$

The conditions for the parameters a, b, c and d are checked at the end of the minimization process and are fulfilled (Table 10.4). After $f(t_n - t)$ is derived we focus on the forward rate specific function $g(t_n)$. In particular one has to find the values of $g(\cdot)$ for each individual forward rate given by t_n, e.g. $g(t_n) = k_n$. k_n is a

Table 10.4 Estimated parameters for time-homogeneous component of the volatility function ▣ SFEcapvolalibor

a	b	c	d
−0.0029	−0.8172	0.0112	15.2314
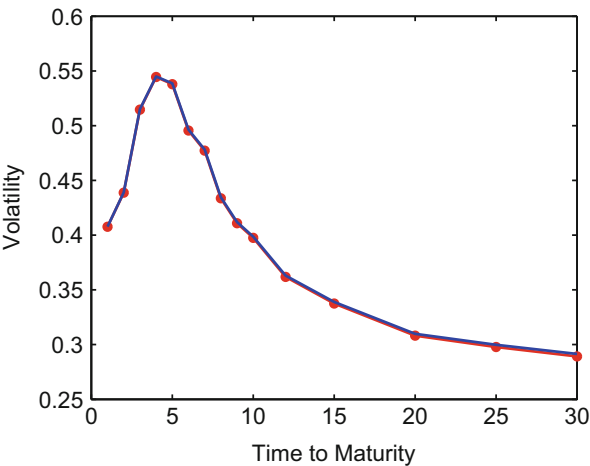
Fig. 10.4 Black implied volatility structure (*blue*) of LIBOR caplets observed on 20120508 and calibrated volatility structure (*red*) with parameters by Table 10.4 ▣ SFEcapvplotlibor

constant (does not depend on t), the time-homogeneous features of the instantaneous volatility function in (10.42) are preserved. By the formula:

$$k_n^2 = \frac{(\sigma_n^{Black})^2 t_n}{\int_0^{t_n} f^2(t_n - s)ds}$$

k_n can be derived from the Black implied volatility and the time-homogeneous component $f(t_n - t)$. Therefore we explicitly ensure that the instantaneous volatility function is consistent with the implied volatilities of the Black model, i.e. the model implied Caplet prices are consistent with the market (Fig. 10.4). This concludes out calibration of the LIBOR market model. The results are used to price traded caps with the Black Model [see Eq. (10.25)].

10.7 Recommended Literature

An introduction to the mathematical concepts and the quantitative tools used in pricing interest rate derivatives is given in Neftci (1996). Standard text books which cover interest rate derivatives and models are Hull (2006) and, in more detail, Brigo and Mercurio (2001). A thorough coverage of volatility and correlation in

Table 10.5 Dataset

Maturity	0.5 years	1 year	1.5 years	2 years
Bond value	0.97	0.94	0.91	0.87
Strike rate	7.50 %	7.50 %	7.50 %	7.50 %

the LIBOR Market Model is given in Rebonato (2004). The classical stochastic interest rate models are introduced in Vasicek (1977), Cox et al. (1985) and Hull and White (1990). A standard work taking a modern point of view of the interest rate in financial mathematics is Heath et al. (1992). The work by Jamshidian (1997) derives the LIBOR Market Model for swap options, whereas Miltersen et al. (1997) and Brace et al. (1997) discuss the LIBOR Market Model for caps and floors.

10.8   Exercises

Exercise 10.1 *Consider the set-up in Table 10.5 with the face value of the considered bonds being equal to* 1 *EUR.*

(a) Calculate the value of the forward rate agreements.
(b) Calculate the value of a receiver interest rate swap.
(c) Determine the swap rate.

Exercise 10.2 *Consider the Hull–White model*

$$dr(t) = \mu_r dt + \sigma dW_t \qquad\qquad (10.43)$$

$$\mu_r = \delta(t) - ar$$

Derive the price of zero-coupon bond at time t with a nominal face value of 1 *EUR under risk-neutral measures.*

Exercise 10.3 *Consider the Vasicek model*

$$dr(t) = a(b - r)dt + \sigma dW_t$$

where a, b, σ are known, W_t is a Wiener process. Derive the price of the pure-discount bond with real-world measure.

Exercise 10.4 *Use the Vasicek in Exercise 10.3,*

(a) Calculate $\mathsf{E}[r_t|\mathcal{F}_s]$ *and* $\mathsf{Var}[r_t|\mathcal{F}_s]$ *where* $s < t$ *and* $\mathcal{F}_s$ *denotes the past information set.*
(b) the yield to maturity is defined as:

$$Y_T(t) = -\log P_T(t)/\tau$$

where $\tau = T - t$, $P_T(t) = V(r, t)$. *Calculate the* $Y_{lim} = \lim\limits_{\tau \to \infty} Y(\tau)$ *and what does it imply?*

Part II
Statistical Models of Financial Time Series

Chapter 11
Introduction: Definitions and Concepts

Financial markets can be regarded from various points of view. Firstly there are economic theories which make assertions about security pricing; different economic theories exist in different markets (currency, interest rates, stocks, derivatives, etc.). Well-known examples include the purchasing power parity for exchange rates, interest rate term structure models, the *capital asset pricing model* (CAPM) and the Black–Scholes option pricing model. Most of these models are based on theoretical concepts which, for example, involve the formation of expectations, utility functions and risk preferences. Normally it is assumed that individuals in the economy act "rationally", have rational expectations and are averse to risk. Under these circumstances prices and returns can be determined in equilibrium models (as, for example, the CAPM) which clear the markets, i.e. supply equals aggregate demand. A different Ansatz pursues the arbitrage theory (e.g., the Black–Scholes model), which assumes that a riskless profit would be noticed immediately by market participants and be eliminated through adjustments in the price. Arbitrage theory and equilibrium theory are closely connected. The arbitrage theory can often get away with fewer assumptions, whereas the equilibrium theory reaches more explicitly defined solutions for complex situations.

The classic econometric models are formulated with economically interpreted parameters. One is interested in the following empirical questions:

1. How well can a specific model describe a given set of data (cross section or time series)?
2. Does the model help the market participants in meeting the relative size of assertions made on future developments?
3. What do the empirical findings imply for the econometric model? Will it eventually have to be modified? Can suggestions actually be made which will influence the function and structural organization of the markets?

In order to handle these empirical questions, a statistical inquiry is needed. Since, as a rule with financial market data, the dynamic characteristics are the most important,

we will mainly concentrate on the time series analysis. First of all, we will introduce the concepts of univariate analysis and then move to the multivariate time series. The interdependence of financial items can be modelled explicitly as a system.

Certain terms, which are often used in time series analysis and in the analysis of financial time series, are introduced in a compact form. We will briefly define them in the next section.

11.1   Some Definitions

First we will need to look closer at stochastic processes, the basic object in time series analysis.

Definition 11.1 (Stochastic Process) A stochastic process X_t, $t \in \mathbb{Z}$, is a family of random variables, defined in a probability space $(\Omega, \mathcal{F}, \mathsf{P})$.

At a specific time point t, X_t is a random variable with a specific density function. Given a specific $\omega \in \Omega$, $X(\omega) = \{X_t(\omega), t \in \mathbb{Z}\}$ is a realization or a path of the process.

Definition 11.2 (Cumulative Distribution Function (CDF) of a Stochastic Process) The joint cumulative distribution function (CDF) of a stochastic process X_t is defined as

$$F_{t_1,\ldots,t_n}(x_1,\ldots,x_n) = \mathsf{P}(X_{t_1} \le x_1, \ldots, X_{t_n} \le x_n).$$

The stochastic process X_t is clearly identified, when the system of its density functions is known. If for any $t_1, \ldots, t_n \in \mathbb{Z}$ the joint distribution function $F_{t_1,\ldots,t_n}(x_1,\ldots,x_n)$ is known, the underlying stochastic process is uniquely determined.

Definition 11.3 (Conditional CDF) The conditional CDF of a stochastic process X_t for any $t_1, \ldots, t_n \in \mathbb{Z}$ with $t_1 < t_2 < \cdots < t_n$ is defined as

$$F_{t_n|t_{n-1},\ldots,t_1}(x_n \mid x_{n-1},\ldots,x_1) = \mathsf{P}(X_{t_n} \le x_n \mid X_{t_{n-1}} = x_{n-1}, \ldots, X_{t_1} = x_1).$$

Next we will define moment functions of the real stochastic process. Here we will assume that the moments exist. If this is not the case, then the corresponding function is not defined.

Definition 11.4 (Mean Function) The mean function μ_t of a stochastic process X_t is defined as

$$\mu_t = \mathsf{E}[X_t] = \int_{\mathbb{R}} x \, dF_t(x). \tag{11.1}$$

In general μ_t depends on time t, as, for example, processes with a seasonal or periodical structure or processes with a deterministic trend.

Definition 11.5 (Autocovariance Function) The autocovariance function of a stochastic process X is defined as

$$\gamma(t, \tau) = \mathsf{E}[(X_t - \mu_t)(X_{t-\tau} - \mu_{t-\tau})]$$
$$= \int_{\mathbb{R}^2} (x_1 - \mu_t)(x_2 - \mu_{t-\tau})dF_{t,t-\tau}(x_1, x_2) \qquad (11.2)$$

for $\tau \in \mathbb{Z}$.

The autocovariance function is symmetric, i.e., $\gamma(t, \tau) = \gamma(t - \tau, -\tau)$. For the special case $\tau = 0$ the result is the variance function $\gamma(t, 0) = \mathsf{Var}(X_t)$. In general $\gamma(t, \tau)$ is dependent on t as well as on τ. Below we define the important concept of stationarity, which will simplify the moment functions in many cases.

Definition 11.6 (Stationary) A stochastic process X is covariance stationary if

1. $\mu_t = \mu$, and
2. $\gamma(t, \tau) = \gamma_\tau$.

A stochastic process X_t is strictly stationary if for any $t_1, \ldots, t_n$ and for all $n, s \in \mathbb{Z}$ it holds that

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) = F_{t_1+s,\ldots,t_n+s}(x_1, \ldots, x_n).$$

For covariance stationary the term "weakly stationary" is often used. One should notice, however, that a stochastic process can be strictly stationary without being covariance stationary, namely that, when the variance (or covariance) does not exist. If the first two moment functions exist, then covariance stationary follows from strictly stationary.

Definition 11.7 (Autocorrelation Function (ACF)) The autocorrelation function ρ of a covariance stationary stochastic process is defined as

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0}.$$

The ACF is normalized on the interval $[-1, 1]$ and thus simplifies the interpretation of the autocovariance structure from various stochastic processes. Since the process is required to be covariance stationary, the ACF depends only on one parameter, the lag τ. Often the ACF is plotted as a function of τ, the so called *correlogram*. This is an important graphical instrument to illustrate the linear dependency structures of the process.

Next we define two important stochastic processes which build the foundation for further modelling.

Definition 11.8 (White Noise (WN)) The stochastic process X_t is white noise if the following holds

1. $\mu_t = 0$, and
2. $\gamma_\tau = \begin{cases} \sigma^2 & \text{when } \tau = 0 \\ 0 & \text{when } \tau \neq 0. \end{cases}$

◓ SFEtimewr

If X_t is a process from i.i.d. random values with expectation 0 and finite variance, then it is a white noise. This special case is called *independent white noise*. On the contrary the white noise could have a dependent third or higher moments, and in this case it would not be independent.

Definition 11.9 (Random Walk) The stochastic process X_t follows a *random walk*, if it can be represented as

$$X_t = c + X_{t-1} + \varepsilon_t$$

with a constant c and white noise ε_t.

If c is not zero, then the variables $Z_t = X_t - X_{t-1} = c + \varepsilon_t$ have a non-zero mean. We call it a random walk with a drift (see Sect. 4.1). In contrast to Sect. 4.3 we do not require the variables here to be independent. The random walk defined here is the boundary case for an AR(1) process introduced in Example 11.1 as $\alpha \to 1$. When we require, as in Sect. 4.3, that ε_t is independent white noise, then we will call X_t a random walk with independent increments. Historically the random walk plays a special role, since at the beginning of the last century it was the first stochastic model to represent the development of stock prices. Even today the random walk is often assumed as an underlying hypothesis. However the applications are rejected in their strongest form with independent increments.

In order to determine the moment functions of a random walk, we will simply assume that the constant c and the initial value X_0 are set to zero. Then, through recursive substitutions we will get the representation:

$$X_t = \varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_1.$$

The mean function is simply:

$$\mu_t = \mathsf{E}[X_t] = 0, \tag{11.3}$$

and for the variance function, since there is no correlation of ε_t, we obtain:

$$\mathsf{Var}(X_t) = \mathsf{Var}(\sum_{i=1}^{t} \varepsilon_i) = \sum_{i=1}^{t} \mathsf{Var}(\varepsilon_i) = t\sigma^2. \tag{11.4}$$

The variance of the random walk increases linearly with time. For the autoco-variance function the following holds for $\tau < t$:

$$\gamma(t, \tau) = \text{Cov}(X_t, X_{t-\tau})$$

$$= \text{Cov}\left(\sum_{i=1}^{t} \varepsilon_i, \sum_{j=1}^{t-\tau} \varepsilon_j\right)$$

$$= \sum_{j=1}^{t-\tau} \sum_{i=1}^{t} \text{Cov}(\varepsilon_i, \varepsilon_j)$$

$$= \sum_{j=1}^{t-\tau} \sigma^2 = (t - \tau)\sigma^2.$$

For $\tau < t$ the autocovariance is thus strictly positive. Since the covariance function depends on time t (and not only on the lags τ), the random walk is not covariance stationary. For the autocorrelation function ρ we obtain

$$\rho(t, \tau) = \frac{(t - \tau)\sigma^2}{\sqrt{t\sigma^2(t-\tau)\sigma^2}} = \frac{(t-\tau)}{\sqrt{t(t-\tau)}} = \sqrt{1 - \frac{\tau}{t}}.$$

Again ρ depends on t as well as on τ, thus the random walk is not covariance stationary.

As a further illustration we consider a simple, but important stochastic process.

Example 11.1 (AR(1) Process)   The stochastic process X_t follows an autoregressive process of first order, written AR(1) process, if

$$X_t = c + \alpha X_{t-1} + \varepsilon_t$$

with a constant parameter α, $|\alpha| < 1$. The process X_t can also, through iterative substitutions, be written as

$$X_t = c(1 + \alpha + \alpha^2 + \cdots + \alpha^{k-1})$$
$$\quad + \alpha^k X_{t-k} + \varepsilon_t + \alpha\varepsilon_{t-1} + \cdots + \alpha^{k-1}\varepsilon_{t-k+1}$$

$$= c\left(\sum_{i=0}^{k-1} \alpha^i\right) + \alpha^k X_{t-k} + \sum_{i=0}^{k-1} \alpha^i \varepsilon_{t-i}$$

$$= c\frac{1 - \alpha^k}{1 - \alpha} + \alpha^k X_{t-k} + \sum_{i=0}^{k-1} \alpha^i \varepsilon_{t-i}$$

If X_{t-k} is given for a particular k (e.g., the initial value of the process), the characteristics of the process are obviously dependent on this value. This influence disappears, however, over time, since we have assumed that $|\alpha| < 1$ and thus $\alpha^k \to 0$ for $k \to \infty$. For $k \to \infty$ a limit in the sense of squared deviation exists, thus we can write the process X_t as

$$X_t = c\frac{1}{1-\alpha} + \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i}.$$

For the moment functions we then have

$$\mu_t = c\frac{1}{1-\alpha},$$

and

$$\gamma_\tau = \frac{\sigma^2}{1-\alpha^2}\alpha^\tau.$$

The ACF is thus simply $\rho_\tau = \alpha^\tau$. For positive α this function is strictly positive, for negative α it alternates around zero. In every case it converges to zero, but with $\alpha = 0.5$, for example, convergence is very fast, and with $\alpha = 0.99$ it is quite slow.

Q SFEacfar1

Definition 11.10 (Markov Process) A stochastic process has the Markov property if for all $t \in \mathbb{Z}$ and $k \geq 1$

$$F_{t|t-1,\dots,t-k}(x_t|x_{t-1},\dots,x_{t-k}) = F_{t|t-1}(x_t|x_{t-1}).$$

In other words, the conditional distribution of a Markov process at a specific point in time is entirely determined by the condition of the system at the previous date. One can also define Markov processes of a higher order, from which the conditional distribution only depends on the finite number of past levels. Two examples of the Markov process of the first order are, the above mentioned random walk with independent variables and the AR(1) process with independent white noise.

Definition 11.11 (Martingale) The stochastic process X_t is a martingale if the following holds

$$\mathsf{E}[X_t|X_{t-1} = x_{t-1},\dots,X_{t-k} = x_{t-k}] = x_{t-1}$$

for every $k > 0$.

The martingale is also a frequently used instrument in describing prices in financial markets. One should notice that for a martingale process only one statement about the conditional expectation is made, while for a Markov process, statements on the entire conditional distribution are made. An example of a martingale is the random walk without a drift. The AR(1) process with $0 < \alpha < 1$ is not a Martingale, since $\mathsf{E}[X_t | x_{t-1}, \ldots, x_{t-k}] = c + \alpha x_{t-1}$.

Definition 11.12 (Fair Game) The stochastic process X_t is a fair game if the following holds

$$\mathsf{E}[X_t | X_{t-1} = x_{t-1}, \ldots, X_{t-k} = x_{t-k}] = 0$$

for every $k > 0$.

Sometimes a fair game is also called a martingale difference. If X_t is namely a martingale, then $Z_t = X_t - X_{t-1}$ is a fair game.

Definition 11.13 (Lag-Operator) The operator L moves the process X_t back by one unit of time, i.e. $LX_t = X_{t-1}$ and $L^k X_t = X_{t-k}$. In addition we define the difference operator Δ as $\Delta = 1 - L$, i.e., $\Delta X_t = X_t - X_{t-1}$ and $\Delta^k = (1 - L)^k$.

After these mathematical definitions we arrive at the more econometric definitions, and in particular, at the term return. We start with a time series of prices $P_1, \ldots, P_n$ and are interested in calculating the return between two periods.

Definition 11.14 (Simple Return) The simple return R_t is defined as

$$R_t = \frac{P_t - P_{t-1}}{P_{t-1}}.$$

Should the average return $R_t(k)$ need to be calculated over k periods, then the geometric mean is taken from the simple gross return, i.e.

$$R_t(k) = \left(\prod_{j=0}^{k-1} (1 + R_{t-j}) \right)^{1/k} - 1.$$

In general the geometric mean is not equal to the arithmetic mean $k^{-1} \sum_{j=0}^{k-1} R_{t-j}$.

Definition 11.15 (Log Return) The log return r_t is defined as

$$r_t = \log \frac{P_t}{P_{t-1}} = \log(1 + R_t).$$

The log return is defined for the case of continuous compounding. For the average return over several periods we have

$$r_t(k) = \log\{1 + R_t(k)\} = \frac{1}{k} \log \prod_{j=0}^{k-1} (1 + R_{t-j})$$

$$= \frac{1}{k} \sum_{j=0}^{k-1} \log(1 + R_{t-j})$$

$$= \frac{1}{k} \sum_{j=0}^{k-1} r_{t-j},$$

i.e. for log returns the arithmetic average return is applied.

For small price changes the difference of the simple return and log return is negligible. According to the Taylor approximation it follows that:

$$\log(1 + x) = \log(1) + \frac{\partial \log x}{\partial x}(1)x + \frac{\partial^2 \log x}{\partial x^2}(1)\frac{x^2}{2!} + \cdots$$

$$= x - \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots .$$

For x close to zero, a first order approximation is sufficient, i.e. $\log(1 + x) \approx x$. As a general rule one could say that with returns under 10 % it does not really matter whether the simple or the log returns are used. This is above all the case when one is studying financial time series with a high frequency, as, for example, with daily values.

11.2 Statistical Analysis of German and British Stock Returns

In this section we describe several classical characteristics of financial time series using daily returns of the DAX index, FTSE 100 index and 40 selected German and British stocks from 1 January 1998 to 31 December 2007. We will concentrate, on the one hand, on the linear, chronological (in)dependence of the returns, and on the other hand, on the distribution characteristics. Tables 11.1 and 11.2 display estimated values of selected parameters for the German and British stock market, respectively. The autocorrelation of first order is for all stock returns close to zero. The largest autocorrelation is for MUENCHNER RUECK (0.1011), and the smallest autocorrelation is for BG GROUP (−0.0903). This is an empirical phenomenon which is also documented for the American market.

Table 11.1 Estimated first order autocorrelation of the returns, $\hat{\rho}_1(r_t)$, the squared returns, $\hat{\rho}_1(r_t^2)$, and the absolute returns, $\hat{\rho}_1(|r_t|)$, as well as estimated skewness, $\hat{S}$, estimated kurtosis, $\widehat{Kurt}$, and the Bera–Jarque test statistic, BJ, for the daily returns of DAX index and German stocks from 1998 to 2007

Q SFEReturns

	$\hat{\rho}_1(r_t)$	$\hat{\rho}_1(r_t^2)$	$\hat{\rho}_1(|r_t|)$	$\hat{S}$	$\widehat{Kurt}$	BJ
DAX	−0.0226	0.1781	0.2194	−0.16	5.99	983.7
ADIDAS	0.0759	0.1867	0.2206	0.00	6.50	1328.7
ALLIANZ	0.0700	0.2035	0.2698	−0.05	8.12	2848.1
BASF	0.0233	0.1517	0.1631	0.09	5.76	833.4
BAYER	0.0147	0.0546	0.1740	0.98	28.98	73721.6
BMW	0.0435	0.1828	0.2255	0.06	6.21	1120.8
COMMERZBANK	0.0642	0.2081	0.2347	0.07	8.17	2905.8
CONTINENTAL	0.0048	0.1139	0.1737	0.05	6.50	1335.5
DEUTSCHE BANK	0.0689	0.2635	0.2838	−0.11	6.32	1203.0
DEUTSCHE TELEKOM	0.0250	0.1760	0.2708	0.09	6.20	1115.0
EON	0.0010	0.1567	0.2218	0.08	5.57	720.5
HENKEL	0.0124	0.2104	0.2266	0.15	6.27	1174.4
LINDE	−0.0479	0.1446	0.1969	0.08	5.83	870.4
LUFTHANSA	0.0275	0.0894	0.1698	−0.04	7.32	2025.0
MAN	0.0179	0.1212	0.1528	−0.02	5.02	444.3
MUENCHENER RUECK	0.1011	0.2085	0.2532	−0.10	9.26	4255.3
RWE	−0.0216	0.2054	0.2042	0.31	6.41	1309.3
SAP	0.0557	0.1447	0.2650	0.36	9.61	4802.8
SIEMENS	0.0771	0.1351	0.2016	0.15	5.43	650.5
THYSSENKRUPP	0.0477	0.0801	0.1389	−0.07	6.13	1069.7
VOLKSWAGEN	0.0889	0.1890	0.1843	−0.09	5.12	489.6

While the first order autocorrelation of the returns of all stock returns are all close to zero, the autocorrelations of the squared and absolute returns of all stocks are positive and significantly larger than zero. Obviously there is a linear relationship in the absolute and squared values of the chronologically sequential returns. Since the autocorrelation is positive, it can be concluded that small absolute returns are followed sequentially by small absolute returns and large absolute returns are followed by large ones again. In other words, there are quiet periods with small price changes and turbulent periods with large oscillations. Indeed one can further conclude that these periods are of relatively longer duration, i.e. the autocorrelations of squared returns from mainly very large orders are still positive. These effects were examined by Mandelbrot and Fama as long ago as the 1960s. They can be modelled using, among others, the ARCH models studied in Chap. 13. Furthermore we will consider estimates for the skewness and kurtosis. Whereas the estimated skewness is in most cases close to zero and is sometimes positive, sometimes negative, the estimated kurtosis is in every case significantly larger than 3. The smallest estimated kurtosis is by MAN $\left(\widehat{Kurt} = 5.02\right)$, and the largest by BRITISH

Table 11.2 Estimated first order autocorrelation of the returns, $\hat{\rho}_1(r_t)$, the squared returns, $\hat{\rho}_1(r_t^2)$, and the absolute returns, $\hat{\rho}_1(|r_t|)$, as well as estimated skewness, $\hat{S}$, estimated kurtosis, $\widehat{Kurt}$ and the Bera–Jarque test statistic, BJ, for the daily returns of FTSE 100 index and British stocks from 1998 to 2007   ⬛ SFEReturns

	$\hat{\rho}_1(r_t)$	$\hat{\rho}_1(r_t^2)$	$\hat{\rho}_1(|r_t|)$	$\hat{S}$	$\widehat{Kurt}$	BJ
FTSE 100	−0.0222	0.2266	0.2390	−0.19	5.66	785.7
ASTRAZENECA	0.0571	0.1935	0.2296	−0.04	7.55	2251.3
BARCLAYS	0.0761	0.2333	0.2718	0.10	5.15	506.4
BG GROUP	−0.0903	0.1623	0.1767	0.15	5.16	517.0
BHP BILLITON	0.0147	0.0725	0.1274	0.38	6.21	1183.2
BP	−0.0144	0.1734	0.1699	0.00	5.42	637.6
BRIT. AMER. TOBACCO	−0.0140	0.0564	0.2548	1.80	30.49	83515.6
BT GROUP	0.0254	0.0971	0.2032	−0.05	7.19	1909.4
DIAGEO	−0.0033	0.1426	0.2310	0.11	7.35	2063.5
GLAXO SMITH KLINE	−0.0010	0.0857	0.1987	0.23	11.50	7877.6
HBOS	0.0164	0.3433	0.3049	0.38	9.03	4012.4
HSBC	0.0152	0.1367	0.2579	0.04	8.37	3128.2
LLOYDS TSB GROUP	0.0270	0.1997	0.2719	0.22	6.07	1048.4
NATIONAL GRID	−0.0403	0.0920	0.1883	−0.04	10.89	6761.3
RIO TINTO	0.0382	0.0989	0.1188	0.42	6.65	1522.4
ROYAL BANK OF SCOT	0.0431	0.2249	0.3040	−0.01	6.84	1603.6
ROYAL DUTCH SHELL A	0.0017	0.2236	0.2521	−0.28	6.25	1179.8
STANDARD CHARTERED	0.0285	0.1606	0.2554	0.06	7.52	2224.0
TESCO	−0.0902	0.1435	0.1931	0.11	5.84	883.0
UNILEVER	0.0235	0.1177	0.2102	−0.48	9.33	4459.0
VODAFONE GROUP	−0.0047	0.1739	0.2253	0.19	5.55	723.3

AMERICAN TOBACCO $\left(\widehat{Kurt} = 30.49\right)$. Under the null hypothesis of the normal distribution, the estimates in (3.2) and (3.3) are independent and asymptotically normally distributed with

$$\sqrt{n}\hat{S} \xrightarrow{\mathcal{L}} N(0, 6)$$

and

$$\sqrt{n}(\widehat{Kurt} - 3) \xrightarrow{\mathcal{L}} N(0, 24).$$

From this the combined test of the normal distribution from Bera and Jarque (BJ) can be derived:

$$BJ = n\left(\frac{\hat{S}^2}{6} + \frac{(\widehat{Kurt} - 3)^2}{24}\right).$$

BJ is asymptotically χ^2 distribution with two degrees of freedom. The last column in Tables 11.1 and 11.2 shows that in all cases the normal distribution hypothesis is clearly rejected by a significance level of 1 % (critical value 9.21). This is above all caused by the value of the kurtosis. Typically in financial time series, the kurtosis is significantly larger than 3, which is caused by frequent appearance of outliers. Furthermore, there are more frequent appearances of very small returns than what one would expect under the normal distribution hypothesis. One says that the empirical distribution of the returns is *leptokurtic*, which means that the distribution has more mass around the centre and in the tails than the normal distribution. The opposite, a weaker asymmetry or *platykurtic* distribution rarely appears in financial markets.

11.3   Expectations and Efficient Markets

Market efficiency is a very general concept in economic theory. A market is called efficient if at every point in time all relevant information is completely reflected in the price of the traded object. This general definition must be defined more concretely, in order to say what "completely reflected" means. To this end we require the concept of *rational expectations*. In general one speaks of rational expectations when by the forecast of a stochastic process P_t all relative and available information $\mathcal{F}_{t-1}$ (see Definition 5.1) is "optimally" used. Optimal means that the mean squared error of the forecast is minimized. This is the case when the conditional expectation (see Sect. 3.5) $\mathsf{E}[P_t|\mathcal{F}_{t-1}]$ is used as the forecast.

Theorem 11.1 *For every* $h > 0$ *using the conditional expectation* $\mathsf{E}[P_{t+h} \mid \mathcal{F}_t]$ *as a forecast,* $P^*_{t+h|t}$ *minimizes the mean squared error* $\mathsf{E}[(P_{t+h} - P^*_{t+h|t})^2]$ *given all relevant information* $\mathcal{F}_t$ *at time* t.

Proof Given any forecast $P^*_{t+h|t}$ that can be written as a (in general nonlinear) function of the random variables at time t, which determines the information set $\mathcal{F}_t$, then the mean squared error can be written as:

$$\mathsf{E}[(P_{t+h} - P^*_{t+h|t})^2] = \mathsf{E}[(P_{t+h} - \mathsf{E}[P_{t+h}|\mathcal{F}_t] + \mathsf{E}[P_{t+h}|\mathcal{F}_t] - P^*_{t+h|t})^2]$$

$$= \mathsf{E}[(P_{t+h} - \mathsf{E}[P_{t+h}|\mathcal{F}_t])^2]$$

$$+ \mathsf{E}[(\mathsf{E}[P_{t+h}|\mathcal{F}_t] - P^*_{t+h|t})^2], \tag{11.5}$$

since the cross product is equal to zero:

$$2\mathsf{E}[(P_{t+h} - \mathsf{E}[P_{t+h}|\mathcal{F}_t]) \, (\mathsf{E}[P_{t+h}|\mathcal{F}_t] - P^*_{t+h|t})] =$$

$$2\mathsf{E}[\mathsf{E}[P_{t+h} - \mathsf{E}[P_{t+h}|\mathcal{F}_t]|\mathcal{F}_t](\mathsf{E}[P_{t+h}|\mathcal{F}_t] - P^*_{t+h|t})] =$$

$$2\mathsf{E}[0 \cdot (\mathsf{E}[P_{t+h}|\mathcal{F}_t] - P^*_{t+h|t})] = 0.$$

The second term on the right-hand side of (11.5) is non-negative and is equal to zero when $\mathsf{E}[P_{t+h}|\mathcal{F}_t] = P^*_{t+h|t}$.   □

Not all economic variables have sufficient information available to estimate $\mathsf{E}[P_t \mid \mathcal{F}_{t-1}]$. This has to do with the type of underlying process that determines P_t and the relative level of the necessary information for the forecast. In order to shed light upon this conceptual problem, hypotheses have been developed in the macro-economic theory, which do not require the use of mathematical expectations $\mathsf{E}[P_t \mid \mathcal{F}_{t-1}]$. The hypothesis on *adaptive expectations* assumes, for instance, that the forecast at time $t-1$ of P_t, $\mathsf{E}^a_{t-1}[P_t]$, is generated by the following mechanism:

$$\mathsf{E}^a_{t-1}[P_t] - \mathsf{E}^a_{t-2}[P_{t-1}] = \theta(P_{t-1} - \mathsf{E}^a_{t-2}[P_{t-1}]) \tag{11.6}$$

with a constant parameter θ, $0 < \theta < 1$. Changes in the forecast result from the last forecast error weighted by θ.

Theorem 11.2 *The adaptive expectation in (11.6) is optimal in the sense of the mean squared error exactly when P_t follows the process*

$$P_t = P_{t-1} + \varepsilon_t - (1-\theta)\varepsilon_{t-1} \tag{11.7}$$

where ε_t is white noise.

Proof With the Lag-Operator L (see Definition 11.13), (11.6) can be represented as

$$\{1 - (1-\theta)L\}\,\mathsf{E}^a_{t-1}[P_t] = \theta P_{t-1}.$$

Since $0 < \theta < 1$ and $\{1 - (1-\theta)z\}^{-1} = \sum_{i=0}^{\infty}(1-\theta)^i z^i$ this can be written as

$$\mathsf{E}^a_{t-1}[P_t] = \theta \sum_{i=0}^{\infty}(1-\theta)^i P_{t-i-1}.$$

The process (11.7) can be rewritten as

$$\{1 - (1-\theta)L\}\,\varepsilon_t = P_t - P_{t-1}$$

and

$$\varepsilon_t = \sum_{j=0}^{\infty}(1-\theta)^j (P_{t-j} - P_{t-j-1})$$

$$= P_t - \theta \sum_{j=0}^{\infty}(1-\theta)^j P_{t-j-1},$$

so that $P_t - \mathsf{E}^a_{t-1}[P_t]$ is white noise. Thus $\mathsf{E}^a_{t-1}[P_t]$ is the best forecast for P_t in the sense of the mean squared error.   □

The process (11.7) is also referred to as the integrated autoregressive moving average process (ARIMA) of order (0,1,1). The family of ARIMA models will be discussed in more detail in Chap. 12. In general exogenous factors, for example, supply shocks, could also be involved in determining the equilibrium prices. In this case adaptive expectations would be suboptimal. If X_t is the stochastic exogenous factor and $\mathcal{F}_t$ is a family of results which are determined from the observations $\{p_t, p_{t-1}, \ldots, x_t, x_{t-1}, \ldots\}$ available at time t, then the optimal process $\mathsf{E}[P_t \mid \mathcal{F}_{t-1}]$ is in general a function of $\{p_t, p_{t-1}, \ldots\}$ and of $\{x_t, x_{t-1}, \ldots\}$. Special cases exist in which adaptive expectations coincide with rational expectations, for example, in a linear supply/demand system with X_t as an exogenous shock that follows a random walk. If X_t is instead an AR(1) process, then forecasts with adaptive expectations have a larger mean squared error than forecasts with rational expectations. If the factor X_t is *common knowledge*, i.e. available to the public, then rational expectations in this example would mean that the price would be optimally forecasted by using this information.

However, when the factor X_t is not observable for everyone, in principle the uninformed agent could *learn* from the prices offered by the informed agent. This means that through the observation of prices they could obtain information on the status of ω, above and beyond what is in their private information set F_i. Here it is assumed that the information function of prices is correctly interpreted.

In order to illustrate what role the price plays in forming expectations, imagine purchasing a bottle of wine. In the store there are three bottles to choose with the prices EUR 300, EUR 30 and EUR 3. Since the bottle for EUR 300 exceeds the budget, only two bottles for EUR 3 and EUR 30 are considered. Now assume that someone who is not a wine expert could not evaluate the quality of the wine from the information on the label. Since one is often pressed for time, collecting information from other sources is time consuming, so what remains is the information implied by the price. Assume further that one has learned through previous shopping experiences that the more expensive wine tends to be better than the cheaper wine. Thus one constructs a function of the price with respect to the quality, i.e. how good the wine is. One would choose the wine for EUR 30 if the better quality and more expensive wine was valued more in the utility function than the price advantage of the cheaper wine. The buyer behaved *rationally*, since he optimized his decision (here maximizing his utility function) with the help of the available information and the price function, assuming that the function was right.

In addition let's take a look at another example of an experimental market which is taken from the literature. We have a security that is traded in two periods P1 and P2 and in each period it pays various dividends according to the type of investor. The trading system is an auction in which at the time of an offer both bid and ask prices are verbally given. There are three types of investors and from each type there are three investors, i.e. a total of nine investors can trade the security, among other instruments. Each investor has an initial capital of 10,000 Franks (1 "Frank"=0.002

Table 11.3 Payments in periods P1
and P2 according to type of investor

Type	P1	P2
I	300	50
II	50	300
III	150	250

USD) and two securities. The initial capital of 10,000 Franks must be paid back at
the end of the second period. Every profit which results from trading the security
may be kept. When the investor is in possession of the security at the end of P1 or P2,
he will receive the dividend with respect to what type of investor he is. Table 11.3
displays information on the dividend payments.

Every investor knows only his own dividend payment, no one elses. The question
is, whether, and if so, how fast, the investors "learn" about the pricing structure, i.e.
gain information on the value of the security to the other investors. There are two
underlying hypotheses:

1. Investors tell each other through their bids about their individual dividends only
   in P1 and P2 ("naive behaviour").
2. The investors draw conclusions through the observed price on the value of
   the security for the other investors and use this information in their own bids
   ("rational behaviour").

Since the experiment is over after the period P2, only the individual dividend
payments of each investor are of interest, so that in P2 both hypotheses coincide:
The equilibrium price is 300 Franks, since type II is just willing to buy at this price
and there is competition among the type II investors. At the beginning of P1, before
any trading begins, each investor has information only on his own dividends, so that
at first one applies naive behaviour: type I and type II would offer a maximum of
350, type III would offer a maximum of 400, thus the equilibrium price according
to the hypothesis of naive behaviour is 400 Franks. This hypothesis performed well
in empirical experiments. When the experiment is repeated with the same dividend
matrix, the investors can learn through the prices of the previous experiment, which
value the security has for the other types of investors in P2. In particular under the
hypothesis of rational behaviour, type I could learn that the equilibrium price in
P2 is higher than what his own dividend would be, thus he could sell the security
at a higher price. The equilibrium price in P1 is under the rational hypothesis 600
Franks. Type I buys at the price in P1 and sells in P2 to type II at a price of 300.

In repeated experiments it was discovered that the participants actually tended
from naive behaviour to rational behaviour, although the transition did not occur
immediately after the first experiment, it was gradual and took about eight repeti-
tions. Other experiments were run, including a *forward and futures market* in which
in the first period P1 the price of the security in P2 could already be determined.
Here it was shown that through the immediate transparency of the security's value in
future periods the transition to rational expectations equilibrium was much quicker.

The observed market price is created through the interaction of various supplies and demands an aggregation of the individual heterogenous information sets. Assume that the price at time t is a function of the state of the economy, the *price function* $p_t(\omega)$, $\omega \in \Omega$. Below we define an equilibrium with rational expectations.

Definition 11.16 (RE-Equilibrium) An equilibrium at t with rational expectations (RE-equilibrium) is an equilibrium in which every agent i optimizes their objective function given the information set $\mathcal{F}_{i,t}$ and the price function $p_t(\omega)$.

Definition 11.16 assumes in particular that every agent includes the information function of the prices correctly in his objective function.

The concept of *efficient markets* is closely related to the concept of rational expectations. According to the original and general definition, a market is efficient if at every point in time all relevant information is reflected in the price. This means, for example, that new information is immediately incorporated into the price. Below we define efficient markets with respect to an information set $\mathcal{G}$.

Definition 11.17 (Efficient Markets) A market is efficient with respect to $\mathcal{G} = (G_t)$, $t \in \mathbb{N}$, $G_t \subset \mathcal{F}_t$, if at every time t the market is in RE-equilibrium with the price function $p_t(\omega)$ and if for every agent i and every time t the following holds

$$G_t \subset \{\mathcal{F}_{i,t} \cup p_t(\omega)\}.$$

Typically three cases are identified as weak, semi-strong and strong efficiency.

1. The market is *weak efficient*, when efficiency refers only to historical prices, i.e. the set $\mathcal{G} = (G_t)$,
   $G_t = \{p_t, p_{t-1}, p_{t-2}, \ldots\}$. This is, for example, achieved when for all i it holds that $\{p_t, p_{t-1}, p_{t-2}, \ldots\} \subset \mathcal{F}_{i,t}$, that is when the historical prices are contained in every private information set.
2. The market is *semi-strong efficient*, when efficiency refers to the set $\mathcal{G} = (G_t)$, $(\cap_i \mathcal{F}_{i,t}) \subset G_t \subset (\cup_i \mathcal{F}_{i,t})$, which includes all publicly available information.
3. The market is *strong efficient*, when efficiency refers to the set $\mathcal{G} = (G_t)$, $G_t = \cup_i \mathcal{F}_{i,t}$, i.e. when all information (public and private) is reflected in the price function. In this case one speaks of a *fully revealing* RE-equilibrium.

An equivalent definition says that under efficient markets no abnormal returns can be achieved. In order to test it one must first determine what a "normal" return is, i.e. one must define an econometric model. Efficient markets can then be tested only with respect to this model. If this combined hypothesis is rejected, it could be that markets are inefficient or that the econometric model is inadequate.

The following is a brief summary of the typical econometric models that have been proposed for financial data. For each of the most interesting financial instruments—stocks, exchange rates, interest rates and options—a corresponding theory will be presented; each is considered to be classic theory in its respective area. In later chapters we will refer back to these theories when discussing empirically motivated expansions.

11.4   Econometric Models: A Brief Summary

11.4.1   Stock Prices: The CAPM

The *CAPM*, developed independently by various authors in the 1960s, is a classical equilibrium model for the valuation of risky securities (stocks). It is based on the following assumptions:

1. Homogenous information among the market participants exists. This belief can be weakened by assuming that under homogenous information a rational equilibrium is *fully revealing* (see the strong version of Definition 11.17).
2. The market has no friction, i.e. there are no transaction costs, no taxes, no restrictions on short selling or on the divisibility of stocks.
3. There is complete competition.
4. There are no arbitrage opportunities.
5. There are a finite number of stocks (K) and risk free security with return r.
6. Every investor has a strictly concave utility function as a function of risky future cash flow. This means that every investor is risk averse.
7. Every investor maximizes their expected utility, which is dependent only on the expectation and variance of risky future cash flow. This is the crucial assumption of the CAPM. Sufficient conditions for this ($\mu - \sigma$)-criterion are either of the following:

   (a) Every investor has a quadratic utility function.
   (b) The stock returns are normally distributed.

Below $X_{i,t}$ and $\alpha_{i,t}$ represent the price and the number of ith stock supplied in equilibrium at time t. We define the *market portfolio $X_{m,t}$* as

$$X_{m,t} = \sum_{i=1}^{K} \alpha_{i,t} X_{i,t}. \tag{11.8}$$

The relative weight $w_{i,t}$ of the ith stock in this portfolio is as follows:

$$w_{i,t} = \frac{\alpha_{i,t} X_{i,t}}{\sum_k \alpha_{k,t} X_{k,t}}.$$

Most of the well-known stock indices are value weighted indices, nevertheless often only the largest stocks on the market are included in the index (DAX, for example, contains only the 30 largest stocks). As in Definition 11.15, we define the stock return as $R_{i,t} = \log(X_{i,t}/X_{i,t-1})$ and the market return as $R_{m,t} = \log(X_{m,t}/X_{m,t-1})$. We assume that the underlying process of the return is covariance stationary. In equilibrium according to the CAPM it holds for every stock i that

$$\mathsf{E}[R_{i,t}] = r + \beta_i (\mathsf{E}[R_{m,t}] - r), \tag{11.9}$$

with the "beta" factor

$$\beta_i = \frac{\mathrm{Cov}(R_{i,t}, R_{m,t})}{\mathrm{Var}(R_{m,t})}.$$

Equation (11.9) says that in equilibrium the expected return of the ith stock is comprised of two components: the return of risk free security and a risk premium which is specifically determined for each stock through the beta factor. Stocks that are positively correlated with the market have a positive risk premium. The larger the correlation of a stock with the market portfolio is, the larger is the premium in CAPM for portfolio risk.

Since the CAPM can be derived using theories on utilities, it is sometimes described as a demand oriented equilibrium model. In contrast to this there are other models that explain the stock returns in terms of various aggregate variables, so called *factors*, and are referred to as being supply oriented. In Sect. 13.4 we will relax the assumptions of time constant variance and covariance implicit in Eq. (11.9).

11.4.2   Exchange Rate: Theory of the Interest Rate Parity

For stocks one can find a large variety of econometric models and for exchange rates there are even more. There are two standard and quite simple theories. However they are not sufficient to explain the considerable price movements in currency markets, especially in the short-run. The *purchasing power parity* (PPP) assumes that identical goods in different countries must have the same relative price, i.e. a relative price given in units of currency. It has been empirically determined that in the long-run this theory describes reality well, but in the short-term price movements could not be explained. The second simple theory, the *theory of interest rate parity*, performs better as capital flows faster than goods. The difference in interest rates can thus resemble the exchange of capital in other currencies, as does the exchange rate. The theory of interest rate parity assumes that domestic and foreign securities are perfect substitutes with respect to duration and risk structure.

Assume that along with forward and futures markets, currency can be traded over time. The *spot price* is calculated by W_t^K, the forward and future price by W_t^T, each is given in units of the foreign currency, i.e. EUR/USD. An internationally acting investor has two choices. Either he holds a domestic capital investment with the domestic interest rate r_t^i or he chooses a foreign investment with the foreign interest rate r_t^a. If he chooses the foreign investment, he must first exchange his capital into the foreign currency at the spot price and at the end, exchange it back again. The uncertainty about the future developments of the exchange rate can be avoided by purchasing a forward or future contract. In this case the return on the foreign investment is $(1/W_t^K)(1 + r_t^a)W_t^T - 1$. If this return is not equal to the domestic interest rate, then an equilibrium has not been reached. Through immediate price

adjustments the interest rate arbitrage disappears and then equilibrium is reached. Thus in equilibrium it must hold that

$$\frac{W_t^T}{W_t^K} = \frac{1 + r_t^i}{1 + r_t^a}, \tag{11.10}$$

i.e. the relationship between forward and future markets and spot markets corresponds exactly to the relationship between domestic and foreign gross interest rates. The relationship in (11.10) is also called the *covered interest rate parity*, since at the time of investment it deals with risk free exchange and interest rates.

In addition to the interest rate arbitrageur, there are the so-called forward and future speculators that compare the expected future exchange rate with the forward and future price and the corresponding risk of purchasing (selling) currency below or above the equilibrium. Consider a simple case where forward and future speculators are risk neutral. Then in equilibrium the expected exchange rate is equal to the forward and future price, i.e.

$$W_t^T = \mathsf{E}[W_{t+1}^K \mid \mathcal{F}_t], \tag{11.11}$$

with the information set $\mathcal{F}_t$ which contains all relevant and available information. Here we assume that the speculators have rational expectations, i.e., the true underlying process is known and is used to build the optimal forecast by the speculators. This can also be written as the relationship:

$$W_{t+1}^K = \mathsf{E}[W_{t+1}^K \mid \mathcal{F}_t] + \varepsilon_{t+1} \tag{11.12}$$

which says that the deviations of the speculator's forecast $\mathsf{E}[W_{t+1}^K \mid \mathcal{F}_t]$ from the realized exchange rates is white noise ε_t (see Definition 11.8). The market is inefficient when the speculators actually are risk neutral and ε_t is not white noise. In this case the set $\mathcal{F}_t$ does not reflect all of the relevant information in the expectations of the speculators—they do not have rational expectations. In order to test for market efficiency (that is, in order to test whether ε_t is white noise) we first need a model for $\mathsf{E}[W_{t+1}^K \mid \mathcal{F}_t]$. This can be formulated from (11.11) and (11.10).

Substituting (11.11) into (11.10) we obtain the so-called *uncovered interest rate parity*,

$$\frac{\mathsf{E}[W_{t+1}^K \mid \mathcal{F}_t]}{W_t^K} = \frac{1 + r_t^i}{1 + r_t^a}. \tag{11.13}$$

This interest rate parity is risky because future exchange rates are uncertain and enter the relationship as expectations.

Together with (11.12) the uncovered interest rate parity (11.13) implies that the following holds

$$W_{t+1}^K = \frac{1 + r_t^i}{1 + r_t^a} W_t^K + \varepsilon_{t+1}. \tag{11.14}$$

When the difference in the long-term interest rates is zero on average, then (11.14) is a random walk (see Definition 11.9). The random walk is the first model to describe exchange rates.

It should be emphasized that the derivation of this simple model occurred under the assumption of risk neutrality of the speculators. In the case of risk aversion, a risk premium must be included. If, for example, we want to test the efficiency of the currency markets, we could then test the combined hypothesis of efficiency and uncovered interest rate parity using risk neutrality. A rejection of this hypothesis indicates market inefficiency or that the interest rate parity model is a poor model for currency markets.

11.4.3 Term Structure: The Cox–Ingersoll–Ross Model

Term structure models are applied to model the chronological development of bond returns with respect to time to maturity. The classic starting point is to identify one or more factors which are believed to determine the term structure. Through specification of the dynamic structure and using specific expectation hypotheses, an explicit solution can be obtained for the returns.

As a typical example we briefly introduce the Cox, Ingersoll and Ross (CIR) model, which has already been mentioned in Sect. 10.4.1. The price of a *Zero Coupon Bond* with a nominal value of 1 EUR is given by $P_T(t)$ at time t, i.e. a security with no dividend payments that pays exactly one EUR at maturity date T. The log return of the zero coupon bond is given by $Y_T(t)$. We assume that continuous compounding holds. The process $Y_T(t)$ is frequently referred to as the *yield to maturity*. The relationship between the price and the return of the zero coupon bond is:

$$P_T(t) = \exp\{-Y_T(t)\tau\}$$

with the remaining time to maturity $\tau = T - t$. This can be easily seen from the definition of a log return (Definition 11.15). For very short time intervals the *short rate $r(t)$* is defined as:

$$r(t) = \lim_{T \to t} Y_T(t).$$

In practice the short rate corresponds to the spot rate, i.e. the interest rate for the shortest possible investment (see Sect. 10.4.1). Consider, intuitively, the choice between an investment in a zero bond with the return $Y_T(t)$ and repeatedly investing at a (risky) short-term interest rate in future periods. An important expectation hypothesis says that the following holds

$$P_T(t) = \mathsf{E}\left[\exp\left(-\int_t^T r(s)ds\right)|\mathcal{F}_t\right].\tag{11.15}$$

The short rate is frequently seen as the most important predicting factor of the term structure. As the CIR model, most one factor models use the short rate as a factor. The CIR model specifies the dynamic of the short rate as a continuous stochastic process

$$dr(t) = a\{b - r(t)\}dt + \sigma\sqrt{r(t)}dW_t\tag{11.16}$$

with a Wiener process W_t and constant parameters a, b and σ—see also Sect. 10.2. The process (11.16) has a so-called *mean reversion* behaviour, i.e. once deviations from the stationary mean b occur, the process is brought back to the mean value again through a positive a. The volatility, written as $\sigma\sqrt{r(t)}$, is larger whenever the interest level is higher, which can also be shown empirically.

Since in Eq. (11.16) $r(t)$ is specified as a Markov process, $P_T(t)$ is, as a consequence of Eq. (11.15), a function of the actual short rate, i.e.

$$P_T(t) = V\{r(t), t\}.$$

With Itô's lemma (5.12) and (10.39) we obtain from (11.16) the differential equation

$$a(b - r)\frac{\partial V(r, t)}{\partial r} + \frac{1}{2}\sigma^2 r\frac{\partial^2 V(r, t)}{\partial r^2} + \frac{\partial V(r, t)}{\partial t} - rV(r, t) = 0.$$

With the bounding constraint $V(r, T) = P_T(T) = 1$ the following solution is obtained:

$$P_T(t) = V\{r(t), t\} = \exp\{A(T - t) + B(T - t)r(t)\},\tag{11.17}$$

where (see Sect. 10.5)

$$A(\tau) = \frac{2ab}{\sigma^2}\log\frac{2\psi\exp\{(a + \psi)\tau/2\}}{g(\tau)},$$

$$B(\tau) = \frac{2\{1 - \exp(\psi\tau)\}}{g(\tau)}$$

$$\psi = \sqrt{a^2 + 2\sigma^2}$$

$$g(\tau) = 2\psi + (a + \psi)\{\exp(\psi\tau) - 1\}.$$

For increasing time periods $T - t$ the term structure curve $Y_T(t)$ converges to the value

$$Y_{lim} = \frac{2ab}{\psi + a}.$$

Fig. 11.1 Term structure curve (*solid line*) according to the Cox–Ingersoll–Ross model with $a = b = \sigma = 0.1$, $n = 50$, short rate of $r_t = 0.2$ and $Y_{lim} = 0.073$ (*dotted line*)   Q SFEcir
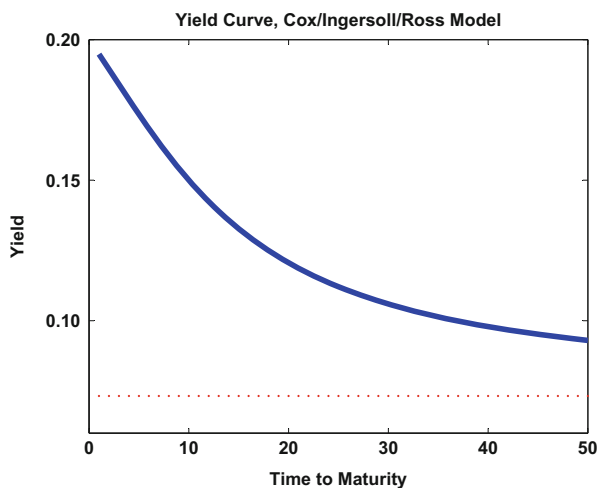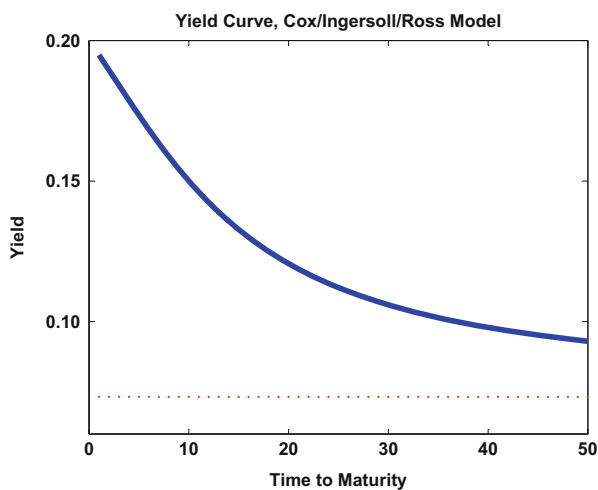

Fig. 11.2 Term structure curve (*solid line*) according to the Cox–Ingersoll–Ross model with $a = b = \sigma = 0.1$, $n = 50$, short rate of $r_t = 0.01$ and $Y_{lim} = 0.073$ (*dotted line*)   Q SFEcir

If the short-term interest lies above b, then the term structure decreases, see Fig. 11.1; if it lies below Y_{lim}, then the term structure increases, see Fig. 11.2. If the short-term interest rate lies between b and Y_{lim}, then the curve could first rise and then fall.

11.4.4 Options: The Black–Scholes Model

Since we have thoroughly covered the Black–Scholes model on option pricing in the first part of this book, only a brief summary of the model will be covered here. Options are not only theoretically interesting for financial markets, but also from an empirical point of view. Just recently there have been indications of a systematic deviation of actual market prices from the Black–Scholes prices. These deviations will be discussed in more detail in later chapters, specifically in dealing with ARCH models.

As an example let's consider a European call option on a stock which receives no dividends in the considered time periods and has the spot price S_t at time t. $C(S,t)$ is the option price at time t, when the actual price is $S_t = S$. The payoff at the time to maturity T is $C(S_T,T) = \max(0, S_T - K)$, where K is the strike price. The option price is determined from general no arbitrage conditions as

$$C(S_t,t) = \mathsf{E}[e^{-r\tau}C(S_T,T)|\mathcal{F}_t],$$

where expectations are built on an appropriate risk neutral distribution—see also (6.23). r is the fixed riskless interest rate.

Special results can only be derived when the dynamics of the stock prices are known. The assumptions made by Black and Scholes are that the stock prices S_t are geometric Brownian motion, i.e.

$$dS_t = \mu S_t dt + \sigma S_t dW_t. \tag{11.18}$$

The option price $C(S,t)$ thus satisfies the Black–Scholes differential equation (6.3) as a function of time and stock prices

$$\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS\frac{\partial C}{\partial S} + \frac{\partial C}{\partial t} = rC$$

Black and Scholes derive the following solutions (see Sect. 6.2):

$$C(S,t) = S\Phi(y + \sigma\sqrt{\tau}) - e^{-r\tau}K\Phi(y), \tag{11.19}$$

where $\tau = T - t$ is the time to maturity for the option and y is an abbreviation for

$$y = \frac{\log\frac{S}{K} + (r - \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}}.$$

11.4.5   The Market Price of Risk

In a risk neutral world the market price of risk, see Sect. 10.2, is equal to zero. In the following section, we will consider the market price of risk and derive once again the Black–Scholes formula. To do this we will consider derivatives of financial instruments that are determined by a single random process θ_t. We will assume that the process θ_t is geometric Brownian motion:

$$d\theta_t = m\theta_t dt + s\theta_t dW_t \; . \tag{11.20}$$

The variable θ_t does not necessarily represent a financial value. It could be the state of the market, a measure of the popularity of a politician or the frequency of an ad hoc announcement at time t. Assume that V_{1t} and V_{2t} are the prices for two derivatives of financial instruments that depend only on θ_t and t. As a simplification, no payments are allowed during the observed time period. This process $V_{jt} = V_j(\theta, t)$, $j = 1, 2$ also follows the schema (11.20) with the *same* Wiener process W_t

$$dV_{jt} = \mu_{jt} V_{jt} dt + \sigma_{jt} V_{jt} dW_t, \qquad j = 1, 2 \tag{11.21}$$

where μ_{jt}, σ_{jt} could be functions of θ_t and t. The random process W_t in (11.20) and (11.21) is always the same since we assume that this is the only source that creates uncertainty.

The observation of (11.21) in discrete time leads to:

$$\Delta V_{1t} = \mu_{1t} V_{1t} \Delta t + \sigma_{1t} V_{1t} \Delta W_t \tag{11.22}$$

$$\Delta V_{2t} = \mu_{2t} V_{2t} \Delta t + \sigma_{2t} V_{2t} \Delta W_t \tag{11.23}$$

We could "eliminate the random variable ΔW_t" by constructing a risk free portfolio that continually changes. To do this we take $\sigma_{2t} V_{2t}$ units of the first instrument and $-\sigma_{1t} V_{1t}$ of the second instrument, i.e. we short sell the second instrument. Letting Π_t be the total value of the portfolio at time t we have:

$$\Pi_t = (\sigma_{2t} V_{2t}) V_{1t} - (\sigma_{1t} V_{1t}) V_{2t} \tag{11.24}$$

and

$$\Delta \Pi_t = (\sigma_{2t} V_{2t}) \Delta V_{1t} - (\sigma_{1t} V_{1t}) \Delta V_{2t} \tag{11.25}$$

Substituting in (11.22) and (11.23) we have:

$$\Delta \Pi_t = (\mu_{1t} \sigma_{2t} V_{1t} V_{2t} - \mu_{2t} \sigma_{1t} V_{1t} V_{2t}) \Delta t. \tag{11.26}$$

This portfolio should be risk free, thus in time period Δt it must produce the risk free profit $r\Delta t$:

$$\frac{\Delta\Pi_t}{\Pi_t} = r\Delta t. \tag{11.27}$$

Substituting (11.24) and(11.26) into this equation produces:

$$(\mu_{1t}\sigma_{2t}V_{1t}V_{2t} - \mu_{2t}\sigma_{1t}V_{1t}V_{2t})\Delta t = (\sigma_{2t}V_{1t}V_{2t} - \sigma_{1t}V_{1t}V_{2t})r\Delta t$$

$$\mu_{1t}\sigma_{2t} - \mu_{2t}\sigma_{1t} = r\sigma_{2t} - r\sigma_{1t}$$

$$\frac{\mu_{1t} - r}{\sigma_{1t}} = \frac{\mu_{2t} - r}{\sigma_{2t}}$$

Equating this as in (10.5) to λ_t we see that the price V_t of a derivative instrument, an instrument that depends only on θ_t and t, follows the dynamics

$$dV_t = \mu_t V_t dt + \sigma_t V_t dW_t, \tag{11.28}$$

the value

$$\lambda_t = \frac{\mu_t - r}{\sigma_t} = \frac{\mu(\theta_t, t) - r}{\sigma(\theta_t, t)} \tag{11.29}$$

represents the *market price of risk*. This market price of risk can depend on θ_t (using μ_t, σ_t), but not on the actual price of the instrument V_t! We can rewrite Eq. (11.29) as:

$$\mu_t - r = \lambda_t \sigma_t \tag{11.30}$$

Furthermore we can interpret σ_t, which in this interpretation can also be negative, as the level of the θ_t-risk in V_t. Equation (11.30) has strong ties to the CAPM, which we discussed in Sect. (11.4.1)—for further details see also Hafner and Herwartz (1998).

Example 11.2 Assume that there are two objects, both are dependent on a 90-day interest rate. The first instrument has an expected return of 6 % per year and a volatility of 20 % per year. For the second instrument a volatility of 30 % per year is assumed. Furthermore, $r = 3$ % per year. The market price of risk for the first instrument according to (11.29) is:

$$\frac{0.06 - 0.03}{0.2} = 0.15 \tag{11.31}$$

By substituting into Eq. (11.30) for the second object we obtain:

$$0.03 + 0.15 \cdot 0.3 = 0.075 \tag{11.32}$$

or 7.5 % expected value.

Since V_t is a function of θ_t and t, we can determine the dependence on θ_t using Itô's lemma. The direct application of Itô's lemma (5.12) on $V(\theta, t)$ gives, in comparison to (11.28), the parameters in this equation

$$\mu_t V_t = m\theta_t \frac{\partial V_t}{\partial \theta_t} + \frac{\partial V_t}{\partial t} + \frac{1}{2} s^2 \theta_t^2 \frac{\partial^2 V_t}{\partial \theta^2}$$

$$\sigma_t V_t = s\theta_t \frac{\partial V_t}{\partial \theta}.$$

Due to Eq. (11.30) we have $\mu_t V_t - \lambda_t \sigma_t V_t = rV_t$, so that we obtain the following differential equation for V_t:

$$\frac{\partial V_t}{\partial t} + (m - \lambda_t s)\theta_t \frac{\partial V_t}{\partial \theta} + \frac{1}{2} s^2 \theta_t^2 \frac{\partial^2 V_t}{\partial \theta^2} = r \cdot V_t \tag{11.33}$$

This equation (11.33) is very similar to the Black–Scholes differential equation and is in fact identical to (6.3) for $\theta_t = S_t$, where S_t denotes the stock price with no dividends. In this case θ_t itself is the price of the risk bearing instrument and must therefore satisfy (11.30), like the price V_t of any derivative based on the stock. Thus we obtain

$$m - r = \lambda_t s, \tag{11.34}$$

so that the second term in (11.33) is equal to

$$r\theta_t \frac{\partial V_t}{\partial \theta}. \tag{11.35}$$

Thus we have a differential equation:

$$\frac{1}{2} s^2 \theta_t^2 \frac{\partial^2 V_t}{\partial \theta^2} + r\theta_t \frac{\partial V_t}{\partial \theta} - rV_t + \frac{\partial V_t}{\partial t} = 0 \tag{11.36}$$

which is identical to (6.3) after renaming the variables. More explicitly, let $S_t = \theta_t, b = r$ (since there are no dividends) and let $\sigma = s$ using the notation in Sect. 6.1.

11.5   The Random Walk Hypothesis

We have seen that econometric models, at least with stock prices and exchange rates, motivate the use of a random walk as a statistical model. With exchange rates we saw that as a consequence of the uncovered interest rate parity and the assumption of risk neutrality of forward and future speculators the model in (11.14) follows a random walk. Assuming a geometric Brownian motion for stock price as in (11.18), it then follows from Itô's lemma that the log of stock price follows a Wiener process with a constant drift rate:

$$d \log S_t = \mu^* dt + \sigma dW_t \tag{11.37}$$

where $\mu^* = \mu - \sigma^2/2$. If one observes (11.37) in time intervals of length $\Delta > 0$, i.e. at discrete points in time $0, \Delta, 2\Delta, \ldots$, then one obtains

$$\log S_{t\Delta} = \log S_{(t-1)\Delta} + \Delta\mu^* + \sqrt{\Delta}\sigma\xi_t \tag{11.38}$$

with independent, standard normally distributed $\xi_t, t = 1, 2, \ldots$. The process (11.38) is a random walk with a drift for the logged stock prices. The log returns (see Definition 11.15) over the time interval of length Δ are also independently normally distributed with expected value $\Delta\mu^*$ and variance $\Delta\sigma^2$.

   With long interest rate time series the random walk appears to be less plausible, since it is assumed that in the long-run there is a stationary level around which interest rates fluctuate in the short run. Let's consider once again the process for the short rate in (11.16), the Cox–Ingersoll–Ross (CIR) model. A discrete approximation is

$$r_t - r_{t-1} = \alpha + \beta r_{t-1} + \sigma\sqrt{r_{t-1}}\xi_t$$

or

$$r_t = \alpha + (1 + \beta)r_{t-1} + \sigma\sqrt{r_{t-1}}\xi_t. \tag{11.39}$$

If β in (11.39) is negative (and larger than -2), then the process is a stationary AR(1) process with heteroscedastic error terms. In Example 11.1 we encountered such a process with heteroscedastic error terms.

   There is also the interpretation that interest rates are, at least in the short-term, explained well by a random walk. It is therefore of general interest to test whether a random walk exists. Below we show the distinguishing features of the three versions of the random walk hypothesis. In general we consider a random walk with a drift:

$$P_t = \mu + P_{t-1} + \varepsilon_t. \tag{11.40}$$

1. The stochastic errors in (11.40) are independent and identically distributed (i.i.d.) with expectation zero and variance σ^2. This hypothesis has already been tested on multiple data sets in the 1960s and was empirically determined to be unsupported. For example, distinct volatility clusters were discovered which under the i.i.d. hypotheses are statistically not expected.
2. The stochastic errors in (11.40) are independent but not necessarily identically distributed with an expectation of zero. This hypothesis is weaker than the first one since, for example, it allows for heteroscedasticity. Nonetheless, even here the empirical discoveries were that a dependence between the error terms must be assumed.
3. The stochastic errors in (11.40) are uncorrelated, i.e. $\gamma_\tau(\varepsilon_t) = 0$ for every $\tau \neq 0$. This is the weakest and most often discussed random walk hypothesis. Empirically it is most often tested through the statistical (in)significance of the estimated autocorrelations of ε_t.

The discussion of the random walk hypotheses deals with, above all, the predicability of financial time series. Another discussion deals with the question of whether the model (11.40) with independent, or as the case may be, with uncorrelated, error terms is even a reasonable model for financial time series or whether it would be better to use just a model with a deterministic trend. Such a *trend-stationary* model has the form:

$$P_t = \nu + \mu t + \varepsilon_t \tag{11.41}$$

with constant parameters ν and μ. The process (11.41) is non-stationary since, for example, $\mathsf{E}[P_t] = \nu + \mu t$, the expected value is time dependent. If the linear time trend is filtered from P_t, then the stationary process $P_t - \mu t$ is obtained.

To compare the difference stationary random walk with a drift to the trend stationary process (11.41) we write the random walk from (11.40) through recursive substitution as

$$P_t = P_0 + \mu t + \sum_{i=1}^{t} \varepsilon_i, \tag{11.42}$$

with a given initial value P_0. One sees that the random walk with a drift also implies a linear time trend, but the cumulative stochastic increments $(\sum_{i=1}^{t} \varepsilon_i)$ in (11.42) are not stationary, unlike the stationary increments (ε_t) in (11.41). Due to the representation (11.42) the random walk with or without a drift will be described as *integrated*, since the deviation from a deterministic trend is the sum of error terms. Moreover, every error term ε_t has a permanent influence on all future values of the process. For the best forecast in the sense of the mean squared error it holds for every $k > 0$,

$$\mathsf{E}[P_{t+k} \mid \mathcal{F}_t] = P_0 + \mu(t + k) + \sum_{i=1}^{t} \varepsilon_i.$$

In contrast, the impact of a shock ε_t on the forecast of the trend-stationary process (11.41) could be zero, i.e.

$$\mathsf{E}[P_{t+k} \mid \mathcal{F}_t] = \nu + \mu(t + k).$$

It is thus of utmost importance to distinguish between a difference stationary and a trend-stationary process. It is worth emphasizing here that the random walk is only a special case of a difference stationary process. If, for example, the increasing variables in (11.40) are stationary but are autocorrelated, then we have a general difference stationary process. There are many statistical tests which test whether a process is difference stationary or not. Two such tests are discussed in the next section.

11.6   Unit Root Tests

In Example 11.1 we discussed that the AR(1) process is:

$$X_t = c + \alpha X_{t-1} + \varepsilon_t. \tag{11.43}$$

Given $|\alpha| < 1$, the process is stationary when $E[X_0] = \frac{l}{1-\alpha}$ or after the "decaying process". The case where $\alpha = 1$ corresponds to the random walk which is non-stationary. The relationship between a stationary AR(1) process and α close to one is so similar to a random walk that it is often tested whether we have the case $\alpha = 1$ or $\alpha < 1$. To do this the so-called *unit root tests* were developed.

11.6.1   Dickey–Fuller Test

The unit root test developed by Dickey and Fuller tests the null hypothesis of a *unit root*, that is, there is a root for the characteristic equation (12.6) of the AR(1) process with $z = 1$, against the alternative hypothesis that the process has no unit roots. As a basis for the test the following regression used is:

$$\Delta X_t = (\alpha - 1)X_{t-1} + \varepsilon_t, \tag{11.44}$$

which is obtained by rearranging (11.43) with $c = 0$. If X_t is a random walk, then the coefficient of X_{t-1} is equal to zero. If, on the other hand, X_t is a stationary AR(1) process, then the coefficient is negative. The standard t-statistic is formed

$$\hat{t}_n = \frac{1 - \hat{\alpha}}{\sqrt{\hat{\sigma}^2(\sum_{t=2}^{n} X_{t-1}^2)^{-1}}}, \tag{11.45}$$

Table 11.4 ADF-Test: simulated rejection probabilities for the process (11.48) at a nominal significance level of 5 % (according to Friedmann (1992))

α	p	β			
		-0.99	-0.9	0	0.9
1	3	0.995	0.722	0.045	0.034
	11	0.365	0.095	0.041	0.039
0.9	3	1.000	0.996	0.227	0.121
	11	0.667	0.377	0.105	0.086

where $\hat{\alpha}$ and $\hat{\sigma}^2$ are the least squares estimators for α and the variance σ^2 of ε_t. For increasing n the statistic (11.45) converges not to a standard normal distribution but instead to the distribution of a functional of Wiener process,

$$\hat{t}_n \xrightarrow{\mathcal{L}} \frac{W^2(1) - 1}{2\left\{\int_0^1 W^2(u)du\right\}^{1/2}},$$

where W is a standard Wiener process. The critical value of the distribution, for example, at the 1, 5 and 10 % significance levels, -2.58, -1.95 and -1.62, respectively. A problem with this test is that the normal test significance level (e.g. 5 %) is not reliable when the error terms ε_t in (11.44) are autocorrelated. The larger the autocorrelation of ε_t, the larger, in general, the distortion will be of the test significance. Ignoring then that autocorrelations could lead to the rejection of the null hypothesis of a unit root at low significance levels of 5 %, when in reality the significance level lies at, for example, 30 %. In order to prohibit these negative effects, Dickey and Fuller suggest another regression which contains lagged differences. The regression of this *augmented Dickey–Fuller Test* (ADF) is thus:

$$\Delta X_t = c + (\alpha - 1)X_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta X_{t-i} + \varepsilon_t \tag{11.46}$$

where, as with the simple Dickey–Fuller Test, the null hypothesis of a unit root is rejected when the test statistic (11.45) is smaller than the critical value (which are summarized in Table 11.4). Naturally the choice of p is problematic. In general it holds that the size of the test is better when p gets larger, but which causes the test to lose *power*. This is illustrated in a simulated process. The errors ε_t are correlated through the relationship:

$$\varepsilon_t = \beta \xi_{t-1} + \xi_t$$

where ξ_t are i.i.d. $(0, \sigma^2)$. In the next chapter these processes will be referred to as *moving average* processes of order 1, MA(1). It holds that $\mathsf{Var}(\varepsilon_t) = \sigma^2(1 + \beta^2)$,

$\gamma_1(\varepsilon_t) = \mathsf{Cov}(\varepsilon_t, \varepsilon_{t-1}) = \beta\sigma^2$ and $\gamma_\tau(\varepsilon_t) = 0$ for $\tau \geq 2$. For the ACF of ε_t we then get

$$
\rho_\tau(\varepsilon_t) = \begin{cases} \frac{\beta}{1+\beta^2} & \text{wenn } \tau = 1 \\ 0 & \text{wenn } \tau \geq 2. \end{cases} \tag{11.47}
$$

For the process

$$
X_t = \alpha X_{t-1} + \beta\xi_{t-1} + \xi_t \tag{11.48}
$$

simulations of the ADF Tests were done and are summarized in an abbreviated form in Table 11.4.

As one can see, the nominal significance level of 5 % under the null hypothesis ($\alpha = 1$) is held better, if p is larger. However the power of the test decreases, i.e. the test is no longer capable of distinguishing between a process with unit roots and a stationary process with $\alpha = 0.9$. Thus in choosing p there is also the conflict between validity and power of the test.

If X_t is a trend-stationary process as in (11.41), the ADF test likewise does not often enough reject the (false) null hypothesis of a unit root. Asymptotically the probability of rejecting goes to zero. The ADF regression (11.46) can be extended by a linear time trend, i.e. run the regression

$$
\Delta X_t = c + \mu t + (\alpha - 1)X_{t-1} + \sum_{i=1}^{p} \alpha_i \Delta X_{t-i} + \varepsilon_t \tag{11.49}
$$

and test the significance of α. The critical values are contained in tables. The ADF test with a time trend (11.49) has power against a trend-stationary process. On the other hand, it loses power as compared to the simple ADF test (11.46), when the true process, for example, is a stationary AR(1) process.

As an empirical example, consider the daily index values for DAX and FTSE 100, as well as the stock prices for the 20 largest German and 20 largest British companies from 1 January 1998 to 31 December 2007. Tables 11.5 and 11.6 display the ADF test statistics for the log index values and log stock prices for the German and British stock market, respectively. Results are provided for the case $p = 0$ and for the case $p = 4$. The tests were run without and with a linear time trend, always including a constant.

For 16 stocks one rejects the unit root hypothesis without trend, and for five stocks one rejects the null hypothesis when a trend was included into the estimation procedure at significance level of 5 %. Since in all other cases the unit root hypothesis is not rejected, one concludes that these processes are integrated. By taking the first differences of the analysed time series, one obtains stationary processes. This implies that the return processes are stationary, and that they can be modelled further. These results will be put into question in the next section using another test.

Table 11.5 Unit root tests: ADF Test (Null hypothesis: unit root) and KPSS Test (Null hypothesis: stationarity) in the case of the German stock market

| | ADF | | | | KPSS | | | |
| | Without time trend | | With time trend | | Without time trend | | With time trend | |
p and T	0	4	0	4	11	36	11	36
DAX	−1.03	−0.95	−1.07	−0.99	3.32*	1.09*	3.40*	1.12*
ADIDAS	−0.74	−0.76	−2.08	−2.08	11.28*	3.72*	3.90*	1.30*
ALLIANZ	−1.13	−1.22	−1.27	−1.30	11.86*	3.88*	2.90*	0.95*
BASF	−0.73	−0.65	−2.10	−2.07	13.99*	4.66*	3.38*	1.14*
BAYER	−0.94	−0.91	−0.89	−0.86	3.69*	1.22*	3.55*	1.17*
BMW	−3.28*	−3.19*	−4.05*	−3.99*	9.86*	3.39*	0.93*	0.33*
COMMERZBANK	−1.44	−1.56	−1.26	−1.39	4.87*	1.60*	3.74*	1.23*
CONTINENTAL	−0.11	−0.09	−1.52	−1.60	16.17*	5.28*	4.77*	1.57*
DEUTSCHE BANK	−2.03*	−2.06*	−2.29	−2.31	4.36*	1.46*	2.04*	0.68*
DEUTSCHE TEL.	−1.14	−1.05	−2.05	−2.04	12.29*	4.02*	1.86*	0.62*
EON	−0.24	−0.07	−1.87	−1.68	12.20*	4.04*	4.05*	1.35*
HENKEL	−1.73	−1.58	−2.45	−2.37	9.48*	3.17*	3.79*	1.28*
LINDE	−1.01	−0.79	−1.74	−1.56	7.53*	2.50*	4.04*	1.34*
LUFTHANSA	−1.73	−1.83	−1.70	−1.80	6.44*	2.13*	3.09*	1.03*
MAN	0.17	0.09	−0.85	−0.95	9.31*	3.07*	4.48*	1.48*
MUENCH. RUECK	−1.23	−1.35	−1.66	−1.80	11.82*	3.87*	2.26*	0.75*
RWE	−0.43	−0.48	−1.37	−1.46	7.92*	2.61*	4.18*	1.38*
SAP	−3.09*	−3.11*	−3.08	−3.09	1.47*	0.51*	1.28*	0.44*
SIEMENS	−1.76	−1.73	−1.92	−1.92	3.61*	1.21*	1.63*	0.54*
THYSSENKRUPP	−0.83	−1.01	−1.28	−1.44	5.23*	1.73*	4.01*	1.33*
VOLKSWAGEN	−0.01	−0.10	−0.29	−0.37	3.90*	1.30*	3.73*	1.24*

The augmented portion of the ADF regression is of order $p = 0$ and $p = 4$. The KPSS statistic was calculated with the reference point $T = 11$ and $T = 36$. The asterisk indicates rejection of the null hypothesis at 5 % significance level   🔍 SFEAdfKpss

11.6.2   The KPSS Test

The KPSS Test from Kwiatkowski et al. (1992) tests for stationarity, i.e. for a unit root. The hypotheses are thus exchanged from those of the ADF test. As with the ADF test, there are two cases to distinguish between, whether to estimate with or without a linear time trend. The regression model with a time trend has the form

$$X_t = c + \mu t + k \sum_{i=1}^{t} \xi_i + \eta_t, \tag{11.50}$$

with stationary η_t and i.i.d. ξ_t with an expected value 0 and variance 1. Obviously for $k \neq 0$ the process is integrated and for $k = 0$ trend-stationary. The null hypothesis is $H_0 : k = 0$, and the alternative hypothesis is $H_1 : k \neq 0$.

Table 11.6 Unit root tests: ADF Test (Null hypothesis: unit root) and KPSS Test (Null hypothesis: stationarity) in the case of the British stock market

	ADF				KPSS			
	Without time trend		With time trend		Without time trend		With time trend	
p and T	0	4	0	4	11	36	11	36
FTSE 100	−1.60	−1.34	−1.57	−1.29	4.46*	1.47*	4.01*	1.32*
ASTRAZENECA	−3.17*	−2.98*	−3.15	−2.96	1.09*	0.37*	1.12*	0.38*
BARCLAYS	−2.71*	−2.50*	−3.67*	−3.44*	12.34*	4.15*	0.74*	0.26*
BG GROUP	0.12	0.85	−1.61	−0.88	17.50*	5.78*	4.43*	1.48*
BHP BILLITON	−0.44	−0.31	−3.39	−3.11	20.14*	6.62*	1.50*	0.51*
BP	−3.02*	−2.77*	−3.11	−2.82	3.31*	1.12*	1.68*	0.57*
BRIT. AM. TOB.	−0.70	−0.38	−2.94	−2.44	19.48*	6.42*	1.86*	0.63*
BT GROUP	−1.17	−1.05	−1.71	−1.53	13.43*	4.38*	3.78*	1.25*
DIAGEO	−2.12*	−1.64	−3.56*	−2.94	13.28*	4.44*	1.44*	0.49*
GLAXO SMITH	−2.32*	−2.11*	−3.46*	−3.11	12.13*	4.03*	2.26*	0.77*
HBOS	−3.00*	−2.65*	−3.43*	−3.03	8.14*	2.76*	2.14*	0.74*
HSBC	−2.75*	−2.75*	−3.09	−2.96	9.94*	3.36*	1.33*	0.46*
LLOYDS TSB GR.	−2.14*	−1.96*	−2.93	−2.65	13.33*	4.42*	2.80*	0.95*
NATIONAL GRID	−1.96*	−1.50	−2.44	−1.99	8.79*	2.95*	2.91*	0.98*
RIO TINTO	0.01	0.03	−2.20	−2.13	17.86*	5.93*	2.75*	0.94*
ROYAL BANK	−2.85*	−2.65*	−2.94	−2.68	13.58*	4.53*	2.19*	0.76*
ROYAL DUTCH	−2.27*	−2.14*	−2.26	−2.13	2.71*	0.90*	2.09*	0.70*
STANDARD CH.	−1.65	−1.50	−2.90	−2.66	12.52*	4.22*	2.81*	0.97*
TESCO	−1.04	−0.53	−3.11	−2.38	17.17*	5.67*	2.09*	0.71*
UNILEVER	−2.39*	−2.36*	−2.79	−2.76	3.65*	1.27*	1.45*	0.50*
VODAFONE GR.	−2.28*	−2.06*	−2.79	−2.58	6.66*	2.20*	1.82*	0.61*

The augmented portion of the ADF regression is of order $p = 0$ and $p = 4$. The KPSS statistic was calculated with the reference point $T = 11$ and $T = 36$. The asterisk indicates rejection of the null hypothesis at 5 % significance level   🇶 SFEAdfKpss

Under H_0 the regression (11.50) is run with the method of the least squares obtaining the residuals $\hat{\eta}_t$. Using these residuals the partial sum

$$S_t = \sum_{i=1}^{t} \hat{\eta}_i,$$

is built which under H_0 is integrated of order 1, i.e. the variance S_t increases linearly with t. The KPSS test statistic is then

$$KPSS_T = \frac{\sum_{t=1}^{n} S_t^2}{n^2 \hat{\omega}_T^2}, \tag{11.51}$$

where

$$\hat{\omega}_T^2 = \hat{\sigma}_\eta^2 + 2 \sum_{\tau=1}^{T} (1 - \frac{\tau}{T-1}) \hat{\gamma}_\tau$$

is an estimator of the spectral density at a frequency of zero where $\hat{\sigma}_\eta^2$ is the variance estimator of η_t and $\hat{\gamma}_\tau = 1/n \sum_{t=\tau+1}^{n} \hat{\eta}_t \hat{\eta}_{t-\tau}$ is the covariance estimator. The problem again is to determine the reference point T: for T that is too small the test is biased when there is autocorrelation, for T that is too large it loses power.

The results of the KPSS test in Tables 11.5 and 11.6 clearly indicate that both analysed indices (log values) and all log stock price processes are not stationary nor trend-stationary, since in every case the null hypothesis was rejected at a significance level of 5 %. Thus, all analysed processes are integrated, and therefore the stationary return processes can be used in further modelling.

11.6.3   Variance Ratio Tests

If one wants to test whether a time series follows a random walk, one can take advantage of the fact that the variance of a random walk increases linearly with time, see (11.4). Considering the log prices of a financial time series, $\log S_t$, the null hypothesis would be

$$H_0 : r_t = \mu + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma^2)$$

with log returns $r_t = \log S_t - \log S_{t-1}$, constant μ and ε_t white noise. An alternative hypothesis is, for example, that r_t is stationary and autocorrelated. The sum over the returns is formed

$$r_t(q) = r_t + r_{t-1} + \cdots + r_{t-q+1}$$

and the variance of $r_t(q)$ is determined. For $q = 2$ it holds that, for example,

$$\begin{aligned} \text{Var}\{r_t(2)\} &= \text{Var}(r_t) + \text{Var}(r_{t-1}) + 2\text{Cov}(r_t, r_{t-1}) \\ &= 2\text{Var}(r_t) + 2\gamma_1 \\ &= 2\text{Var}(r_t)(1 + \rho_1), \end{aligned}$$

where taking advantage of the stationarity of r_t, generally

$$\text{Var}\{r_t(q)\} = q\text{Var}(r_t) \left(1 + 2 \sum_{\tau=1}^{q-1} (1 - \frac{\tau}{q}) \rho_\tau \right). \tag{11.52}$$

Under H_0 it holds that $\rho_\tau = 0$ for all $\tau > 0$, so that under H_0

$$\frac{\mathsf{Var}\{r_t(q)\}}{q\mathsf{Var}(r_t)} = 1.$$

A test statistic can now be constructed where the consistent estimator

$$\hat{\mu} = \frac{1}{n}(\log S_n - \log S_0)$$

for μ,

$$\widehat{\gamma_0} = \frac{1}{n-1}\sum_{t=2}^{n}(\log S_t - \log S_{t-1} - \hat{\mu})^2$$

for $\mathsf{Var}(r_t)$ and

$$\widehat{\gamma_0}(q) = \frac{n}{q(n-q)(n-q+1)}\sum_{t=q+1}^{n}(\log S_t - \log S_{t-q} - q\hat{\mu})^2$$

for $\frac{1}{q}\mathsf{Var}\{r_t(q)\}$ are substituted into (11.52). The test statistic is then

$$VQ(q) = \frac{\widehat{\gamma_0}(q)}{\widehat{\gamma_0}} - 1.$$

It can be shown that the asymptotic distribution is

$$\sqrt{n}VQ(q) \xrightarrow{\mathcal{L}} \mathrm{N}\left(0, \frac{2(2q-1)(q-1)}{3q}\right).$$

The asymptotic variance can be established through the following approximation: Assume that $\hat{\mu} = 0$ and $n \gg q$. Then we have that $\log S_t - \log S_{t-q} = \sum_{j=0}^{q-1}r_{t-j}$ and

$$
\begin{aligned}
VQ(q) &\approx \frac{1}{qn}\sum_{t=q+1}^{n}\left(\left(\sum_{j=0}^{q-1}r_{t-j}\right)^2 - q\widehat{\gamma_0}\right)/\widehat{\gamma_0} \\
&= \frac{1}{qn}\sum_{t=q+1}^{n}\frac{1}{\widehat{\gamma_0}}\left(\sum_{j=0}^{q-1}r_{t-j}^2 + 2\sum_{j=0}^{q-2}r_{t-j}r_{t-j-1} + \cdots + 2r_t r_{t-q+1} - q\widehat{\gamma_0}\right) \\
&\approx \frac{1}{q}\left(q\widehat{\gamma_0} + 2(q-1)\widehat{\gamma_1} + \cdots + 2\widehat{\gamma_{q-1}} - q\widehat{\gamma_0}\right)/\widehat{\gamma_0} \\
&= 2\sum_{j=1}^{q-1}\frac{q-j}{q}\widehat{\rho_j}.
\end{aligned}
$$

Since under H_0 the estimated autocorrelation $\widehat{\rho}_j$ scaled with $\sqrt{n}$ is asymptotically standard normal and is independent, see Sect. 12.5, the asymptotic variance is thus:

$$
\begin{aligned}
\mathsf{Var}\{\sqrt{n}VQ(q)\} &= \mathsf{Var}\left(2\sum_{j=1}^{q-1}\frac{q-j}{q}\sqrt{n}\widehat{\rho}_j\right) \\
&= 4\sum_{j=1}^{q-1}\frac{(q-j)^2}{q^2}\mathsf{Var}(\sqrt{n}\widehat{\rho}_j) \\
&= 4\sum_{j=1}^{q-1}\frac{(q-j)^2}{q^2} \\
&= 4(q-1) - \frac{8}{q}\sum_{j=1}^{q-1}j + \frac{4}{q^2}\sum_{j=1}^{q-1}j^2.
\end{aligned}
$$

With the summation formulas

$$
\sum_{j=1}^{q-1}j = (q-1)q/2
$$

and

$$
\sum_{j=1}^{q-1}j^2 = q(q-1)(2q-1)/6
$$

we finally obtain

$$
\mathsf{Var}\{\sqrt{n}VQ(q)\} = \frac{2(2q-1)(q-1)}{3q}.
$$

11.7   Recommended Literature

Four current textbooks in the area of empirical financial market analysis are Mills (1993), Gouriéroux (1997), Campbell et al. (1997) and Gouriéroux and Jasiak (2002). The focus of Gouriéroux and Mills is more towards the econometric/time series analysis applications (which will also be followed in this book), whereas Campbell, Lo and MacKinlay discuss many economic applications that do not always end with statistical or econometric models. As an introduction and yet a comprehensive book on time series analysis Schlittgen and Streitberg (1995)

is recommended. The same can be found with Copeland and Weston (1992), an introductory book on finance theory.

The experiment of the expectation hypotheses comes from Forsythe et al. (1982). The definition of expectations and efficient markets is based on Jarrow (1992). The CAPM is developed in Sharpe (1964), Lintner (1965) and Mossin (1966). The discussion on interest rate parity follows Jarchow and Rühmann (1994)[p. 236] and that of the term structure models of Cox–Ingersoll–Ross follow Ingersoll (1987). The standard option pricing model originated in Black and Scholes (1973). The market price of risk is discussed in Hull (2006).

A good overview of unit root tests is given in Hassler (1994). The ADF test is taken from Dickey and Fuller (1979).

11.8 Exercises

Exercise 11.1 *Let X be a random variable with $\mathsf{E}(X^2) < \infty$ and define a stochastic process*

$$X_t \stackrel{\text{def}}{=} (-1)^t X, \ t = 1, 2, \ldots \tag{11.53}$$

(a) What do the paths of this process look like?
(b) Find a necessary and sufficient condition for X such that the process $\{X_t\}$ is strictly stationary.
(c) Find a necessary and sufficient condition for X such that $\{X_t\}$ is covariance (weakly) stationary.
(d) Let X be such that $\{X_t\}$ is covariance (weakly) stationary. Calculate the autocorrelation ρ_τ.

Exercise 11.2 *Perform an empirical analysis using the data on DAX and Dow Jones index from the period 1 January 1997 to 30 December 2004.*

1. Display a time plot of the given indices data, its returns and log returns.
2. Calculate mean, skewness, kurtosis, autocorrelation of the first order, autocorrelation of squared returns and autocorrelation of absolute returns for the given data.

Exercise 11.3 *Consider the data on the DAX and Dow Jones index from Exercise 11.2. Which empirical distribution do the returns follow? Are they normally distributed? Perform an appropriate test of normality.*

Exercise 11.4 *According to the stylized facts, the autocorrelation of first order is close to zero for all stock returns; the autocorrelation of squared and absolute returns are positive and different from zero. In addition, small (positive or negative) returns are followed by small (positive or negative) returns and large returns are followed by large returns. Can you approve this fact applied to your data? Plot the*

autocorrelation function (ACF) for returns, absolute returns and for squared log returns from the DAX and Dow Jones data from Exercise 11.1.

Exercise 11.5 *Use the data of DAX and Dow Jones indices from Exercise 11.1. Apply the Augmented Dickey–Fuller test (ADF) of stationarity to the*

1. raw data, i.e. $I(1)$
2. log returns, i.e. $I(0)$.

Exercise 11.6 *Use the data on DAX and Dow Jones index from Exercise 11.2. Apply the KPSS of stationarity to the*

1. raw data, i.e. $I(1)$
2. log returns, i.e. $I(0)$

Exercise 11.7 *Is there a difference between the ADF and KPSS tests?*

Chapter 12
ARIMA Time Series Models

In this chapter we will deal with classic, linear time series analysis. At first we will define the general linear process.

Definition 12.1 (Linear Process) If the process X_t has the representation

$$X_t = \mu + \sum_{i=-\infty}^{\infty} a_i \varepsilon_{t-i}$$

with absolute summability of the filter (a_i): $\sum_{i=-\infty}^{\infty} |a_i| < \infty$, and white noise process ε_t, then it is a linear process.

The linear process X_t is covariance stationary, since $\mathsf{E}(X_t) = \mu$ and

$$\mathsf{Cov}(X_t, X_{t+\tau}) = \sigma^2 \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} a_i a_j \mathbf{1}(\tau = i - j) = \sigma^2 \sum_{i=-\infty}^{\infty} a_i a_{i-\tau}$$

with $\mathsf{Var}(\varepsilon_t) = \sigma^2$.

In general in econometrics, especially in the area of financial markets, series are observed which indicate non-stationary behaviour. In the previous chapter we saw that econometric models, which are based on assumptions of rational expectations, frequently imply that the relevant levels of, for example, prices, follow a random walk. In order to handle these processes within the framework of the classic time series analysis, we must first form the differences in order to get a stationary process. We generalize the definition of a difference stationary process in the following definition.

Definition 12.2 (Integrated Process) We say that the process X_t is integrated of order d, $I(d)$, when $(1 - L)^{d-1} X_t$ is non-stationary and $(1 - L)^d X_t$ is stationary.

White noise is, for example, $I(0)$, a random walk $I(1)$. In only a few cases processes are observed that are $I(d)$ with $d > 1$. This means that in most cases first differences are enough to form a stationary process. In the following we assume that the observed process Y_t is $I(d)$ and we consider the transformed process $X_t = (1 - L)^d Y_t$, i.e. we will concentrate on stationary processes.

12.1   Moving Average Processes

The moving average process of order q, MA(q), is defined as

$$X_t = \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} + \varepsilon_t \tag{12.1}$$

with white noise ε_t. With the Lag-Operator L (see Definition 11.13) instead of (12.1) we can write

$$X_t = \beta(L)\varepsilon_t \tag{12.2}$$

with $\beta(L) = 1 + \beta_1 L + \ldots + \beta_q L^q$. The MA($q$) process is stationary, since it is formed as the linear combination of a stationary process. The mean function is simply $\mathsf{E}(X_t) = 0$. Let $\beta_0 = 1$, then the covariance structure is

$$\begin{aligned}
\gamma_\tau &= \mathsf{Cov}(X_t, X_{t+\tau}) \\
&= \mathsf{Cov}(\sum_{i=0}^{q} \beta_i \varepsilon_{t-i}, \sum_{j=0}^{q} \beta_j \varepsilon_{t+\tau-j}) \\
&= \sum_{i=0}^{q} \sum_{j=0}^{q} \beta_i \beta_j \mathsf{Cov}(\varepsilon_{t-i}, \varepsilon_{t+\tau-j}) \\
&= \sum_{i=0}^{q-|\tau|} \beta_i \beta_{i+|\tau|} \sigma^2, \quad |\tau| \le q.
\end{aligned}$$

For the ACF we have for $|\tau| \le q$

$$\rho_\tau = \frac{\sum_{i=0}^{q-|\tau|} \beta_i \beta_{i+|\tau|}}{\sum_{i=0}^{q} \beta_i^2}, \tag{12.3}$$

and $\rho_\tau = 0$ for $|\tau| > q$, i.e. the ACF breaks off after q lags.

As an example consider the MA(1) process

$$X_t = \beta \varepsilon_{t-1} + \varepsilon_t,$$

which according to (12.3) holds that $\rho_1 = \beta/(1 + \beta^2)$ and $\rho_\tau = 0$ for $\tau > 1$. Figure 12.1 shows the correlogram of a MA(1) process.
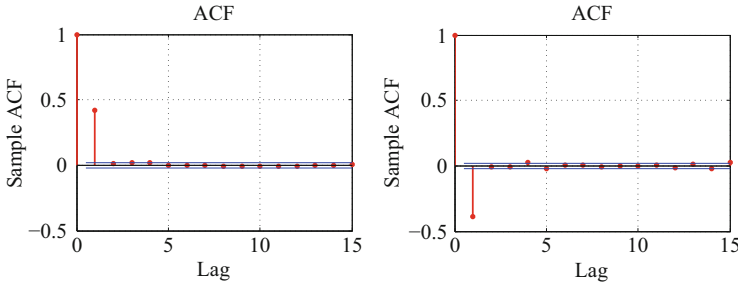
Fig. 12.1 ACF of a MA(1) process with $\beta = 0.5$ (*left*) and $\beta = -0.5$ (*right*)   ▣ SFEacfma1

Obviously the process

$$X_t = 1/\beta \varepsilon_{t-1} + \varepsilon_t$$

has the same ACF, and it holds that

$$\rho_1 = \frac{1/\beta}{1 + (1/\beta)^2} = \frac{\beta}{1 + \beta^2}.$$

In other words the process with the parameter β has the same stochastic properties as the process with the parameter $1/\beta$. This identification problem can be countered by requiring that the solutions of the characteristic equation

$$1 + \beta_1 z + \ldots + \beta_q z^q = 0 \tag{12.4}$$

lie outside of the complex unit circle. In this case the linear filter $\beta(L)$ is invertible, i.e. a polynomial $\beta^{-1}(L)$ exists so that $\beta(L)\beta^{-1}(L) = 1$ and $\beta^{-1}(L) = b_0 + b_1 L + b_2 L^2 + \ldots$. Figure 12.2 displays the correlogram of a MA(2) process $X_t = \beta_1 \varepsilon_{t-1} + \beta_2 \varepsilon_{t-2} + \varepsilon_t$ for some collections of parameters.

12.2 Autoregressive Process

The linear autoregressive process of order p, AR(p), is defined as

$$X_t = \nu + \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \varepsilon_t \tag{12.5}$$

Using the definition of the lag-operator L (see Definition 11.13), (12.5) can also be written as
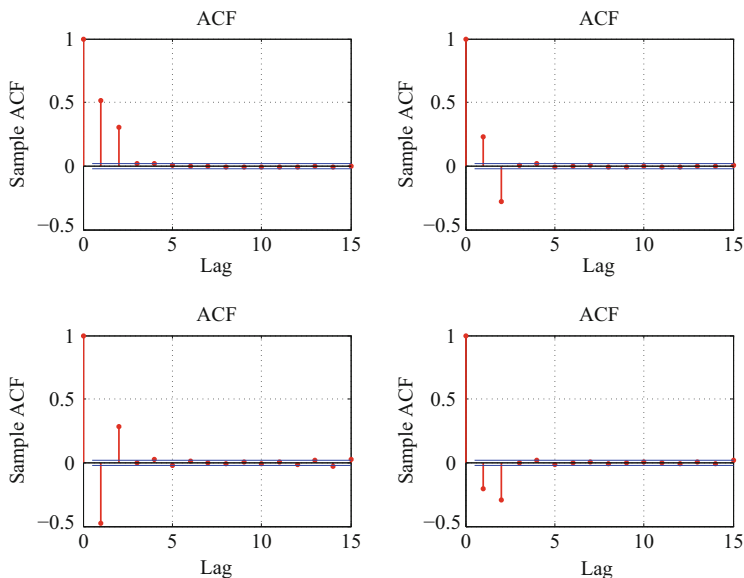
$$\alpha(L)X_t = \nu + \varepsilon_t,$$

Fig. 12.2 ACF of a MA(2) process with ($\beta_1 = 0.5, \beta_2 = 0.4$) (*top left*), ($\beta_1 = 0.5, \beta_2 = -0.4$) (*top right*), ($\beta_1 = -0.5, \beta_2 = 0.4$) (*bottom left*) and ($\beta_1 = -0.5, \beta_2 = -0.4$) (*bottom right*)
                                                                                      Q SFEacfma2

with the lag-polynomial $\alpha(L) = 1 - \alpha_1 L - \ldots - \alpha_p L^p$. The process X_t is stationary if all roots of the characteristic equation

$$\alpha(z) = 1 - \alpha_1 z - \ldots - \alpha_p z^p = 0. \tag{12.6}$$

lie outside of the complex unit circle, that is, if for all z with $|z| \le 1$ it holds that

$$\alpha(z) \ne 0. \tag{12.7}$$

In this case there is an inverted filter $\alpha^{-1}(L)$ for the linear filter $\alpha(L)$, such that the following holds,

$$\alpha(L)\alpha^{-1}(L) = 1$$

and

$$\alpha^{-1}(L) = a_0 + a_1 L + a_2 L^2 + \ldots = \sum_{i=0}^{\infty} a_i L^i.$$

The process (12.5) can also be written under the condition (12.7)

$$X_t = \alpha^{-1}(1)v + \alpha^{-1}(L)\varepsilon_t$$

$$= \sum_{i=0}^{\infty} a_i \nu + \sum_{i=0}^{\infty} a_i L^i \varepsilon_t,$$

as a MA(∞) process.

A simple way to find and invert the autocovariance of an AR(p) process with given parameters is to use the Yule–Walker equations. They are derived from the definition of an AR(p) process in (12.5) by multiplying by $X_{t-\tau}$ and taking expectations.

$$\mathsf{E}[X_t X_{t-\tau}] = \alpha_1 \mathsf{E}[X_{t-1} X_{t-\tau}] + \ldots + \alpha_p \mathsf{E}[X_{t-p} X_{t-\tau}]. \tag{12.8}$$

Since $\mathsf{E}[X_t X_{t-\tau}]$ is the definition of the autocovariance function γ_τ for $\nu = 0$, it can even be written simpler for $\tau = 1, \ldots, p$

$$\begin{aligned}
\alpha_1 \gamma_0 &+ \alpha_2 \gamma_1 &+ \cdots + \alpha_p \gamma_{p-1} &= \gamma_1 \\
\alpha_1 \gamma_1 &+ \alpha_2 \gamma_0 &+ \cdots + \alpha_p \gamma_{p-2} &= \gamma_2 \\
&\cdots\cdots\cdots\cdots\cdots \\
\alpha_1 \gamma_{p-1} &+ \alpha_2 \gamma_{p-2} + \cdots + &\alpha_p \gamma_0 &= \gamma_p
\end{aligned} \tag{12.9}$$

or by dividing by the variance γ_0

$$\rho = R\alpha \tag{12.10}$$

with $\rho = (\rho_1 \rho_2 \ldots \rho_p)^\top$, $\alpha = (\alpha_1 \alpha_2 \ldots \alpha_p)^\top$ and the $p \times p$-autocovariance matrix

$$R = \begin{pmatrix}
1 & \rho_1 & \cdots & \rho_{p-1} \\
\rho_1 & 1 & \cdots & \rho_{p-2} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{p-1} & \rho_{p-2} & \cdots & 1
\end{pmatrix}.$$

Yule–Walker equations are useful in determining the ACF for given parameters or, vice versa, in determining the estimated parameters for the given (empirical) autocorrelation.

Example 12.1 (AR(1)) The AR(1) process from Example 11.1 with $\nu = 0$ has the characteristic equation $1 - \alpha z = 0$. The explicit solution $z = 1/\alpha$ and $|z| > 1$ occurs exactly when $|\alpha| < 1$. The inverse filter of $\alpha(L) = 1 - \alpha L$ is thus $\alpha^{-1}(L) = \sum_{i=0}^{\infty} \alpha^i L^i$ and the MA(∞) representation of the AR(1) process is

$$X_t = \sum_{i=0}^{\infty} \alpha^i \varepsilon_{t-i}.$$

The ACF of the AR(1) process is $\rho_\tau = \alpha^\tau$. For $\alpha > 0$ all autocorrelations are positive, and for $\alpha < 0$ they alternate between positive and negative, see Fig. 12.3.
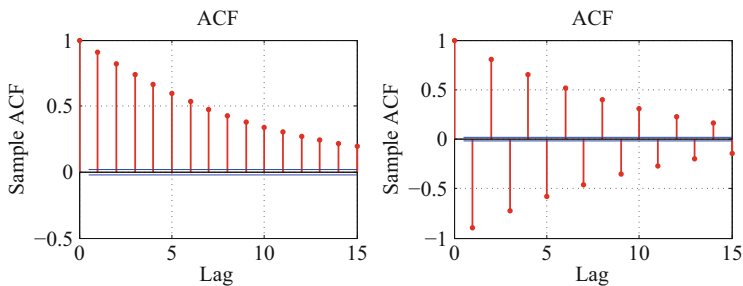
Fig. 12.3 ACF of an AR(1) process with $\alpha = 0.9$ (*left*) and $\alpha = -0.9$ (*right*)   SFEacfar1

Example 12.2 (AR(2)) The AR(2) process with $\nu = 0$,

$$X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \varepsilon_t$$

is stationary when given the roots z_1 and z_2 of the quadratic equation

$$1 - \alpha_1 z - \alpha_2 z^2 = 0,$$

it holds that $|z_1| > 1$ and $|z_2| > 1$. We obtain solutions as

$$z_{1,2} = -\frac{\alpha_1}{2\alpha_2} \pm \sqrt{\frac{\alpha_1^2}{4\alpha_2^2} + \frac{1}{\alpha_2}}$$

and $z_1 z_2 = -1/\alpha_2$. Due to $|z_1| > 1$ and $|z_2| > 1$ it holds that $|z_1 z_2| = 1/|\alpha_2| > 1$ and

$$|\alpha_2| < 1. \tag{12.11}$$

From the Yule–Walker equations in the case of an AR(2) process

$$\rho_1 = \alpha_1 + \alpha_2 \rho_1 \tag{12.12}$$

$$\rho_2 = \alpha_1 \rho_1 + \alpha_2 \tag{12.13}$$

it follows that $\rho_1 = \alpha_1/(1 - \alpha_2)$. The case $\rho_1 = \pm 1$ is excluded because a root would lie on the unit circle (at 1 or -1). Thus for a stationary AR(2) process it must hold that

$$|\rho_1| = |\alpha_1/(1 - \alpha_2)| < 1,$$

from which, together with (12.11), we obtain the "stationarity triangle"

$$\alpha_1 + \alpha_2 < 1 \tag{12.14}$$

$$\alpha_2 - \alpha_1 < 1 \tag{12.15}$$

i.e. the region in which the AR(2) process is stationary.
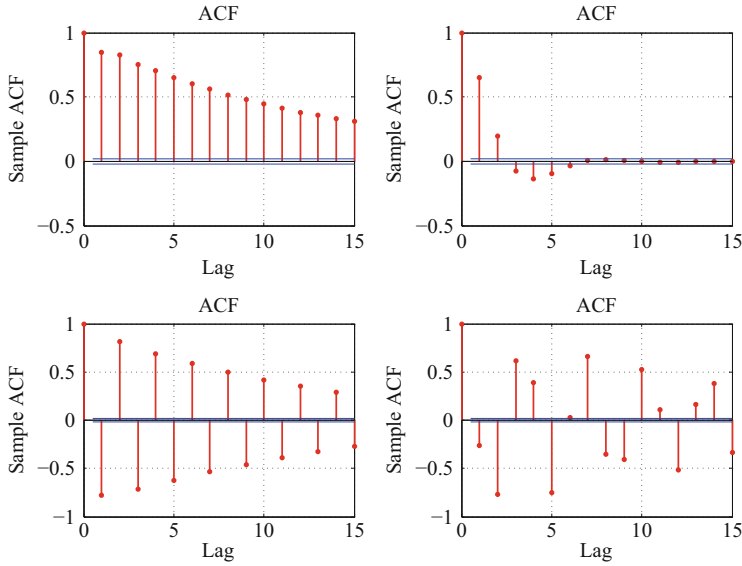
Fig. 12.4 ACF of a AR(2) process with $(\alpha_1 = 0.5, \alpha_2 = 0.4)$ (*top left*), $(\alpha_1 = 0.9, \alpha_2 = -0.4)$ (*top right*), $(\alpha_1 = -0.4, \alpha_2 = 0.5)$ (*bottom left*) and $(\alpha_1 = -0.5, \alpha_2 = -0.9)$ (*bottom right*)

SFEacfar2

The ACF of the AR(2) process is recursively given with (12.12), (12.13) and $\rho_\tau = \alpha_1 \rho_{\tau-1} + \alpha_2 \rho_{\tau-2}$ for $\tau > 2$. Figure 12.4 displays typical patterns.

12.3 ARMA Models

The ARMA(p, q) model is defined as

$$X_t = \nu + \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \beta_1 \varepsilon_{t-1} + \ldots + \beta_q \varepsilon_{t-q} + \varepsilon_t, \qquad (12.16)$$

or as

$$\alpha(L)X_t = \nu + \beta(L)\varepsilon_t$$

with the moving average lag-polynomial $\beta(L) = 1 + \beta_1 L + \ldots + \beta_q L^q$ and the autoregressive lag-polynomial $\alpha(L) = 1 - \alpha_1 L - \ldots - \alpha_p L^p$. In order that the process (12.16) can have explicit parameterization, it is required that the characteristic polynomials $\alpha(z)$ and $\beta(z)$ do not have any common roots. The process (12.16) is stationary when all the roots of the characteristic equation (12.6) lie outside of the unit circle. In this case (12.16) has the MA(∞) representation

$$X_t = \alpha^{-1}(L)\beta(L)\varepsilon_t.$$

The process X_t in (12.16) is invertible when all the roots of the characteristic equation (12.4) lie outside of the unit circle. In this case (12.16) can be written as

$$\beta^{-1}(L)\alpha(L)X_t = \varepsilon_t,$$

that is an AR(∞) process. Thus we can approximate every stationary, invertible ARMA(p,q) process with a pure AR or MA process of sufficiently large order. Going in the other direction, an ARMA(p,q) process offers the possibility of parsimonious parameterization.

12.4   Partial Autocorrelation

For a given stochastic process one is often interested in the connection between two random variables of a process at different points in time. One way to measure a linear relationship is with the ACF, i.e. the correlation between these two variables. Another way to measure the connection between X_t and $X_{t+\tau}$ is to filter out of X_t and $X_{t+\tau}$ the linear influence of the random variables that lie in between, $X_{t+1}, \ldots, X_{t+\tau-1}$, and then calculate the correlation of the transformed random variables. This is called the *partial* autocorrelation.

Definition 12.3 (Partial Autocorrelation) The partial autocorrelation of k-th order is defined as

$$\phi_{kk} = \mathsf{Corr}(X_t - \mathcal{P}(X_t \mid X_{t+1}, \ldots, X_{t+k-1}),$$
$$X_{t+k} - \mathcal{P}(X_{t+k} \mid X_{t+1}, \ldots, X_{t+k-1})) \qquad (12.17)$$

where $\mathcal{P}(W \mid Z)$ is the best linear projection of W on Z, i.e. $\mathcal{P}(W \mid Z) = \Sigma_{WZ}\Sigma_{ZZ}^{-1}Z$ with $\Sigma_{ZZ} = \mathsf{Var}(Z)$ as the covariance matrix of the regressors and $\Sigma_{WZ} = \mathsf{Cov}(W, Z)$ as the matrix of covariances between W and Z.

The "best linear projection" is understood in the sense of minimizing the mean squared error.

An equivalent definition is the solution to the system of equations

$$P_k \phi_k = \rho_{(k)}$$

with

$$P_k = \begin{pmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \cdots & 1 \end{pmatrix}$$

$\phi_k = (\phi_{k1}, \ldots, \phi_{kk})^\top$ and $\rho_{(k)} = (\rho_1, \ldots, \rho_k)^\top$. These are the Yule–Walker equations for an AR(k) process. The last coefficient, ϕ_{kk}, is the partial autocorrelation of order k, as defined above. Since only this coefficient is of interest in this context, the system of equations can be solved for ϕ_{kk} using the Cramer-Rule. We get

$$\phi_{kk} = \frac{|P_k^*|}{|P_k|}$$

where P_k^* is equal to the matrix P_k, in which the k-th column is replaced with $\rho_{(k)}$. Here $|.|$ indicates the determinant. Since this can be applied to various orders k, in the end we obtain a *partial autocorrelation function* (PACF). The PACF can be graphically displayed for a given stochastic process, similar to the ACF as a function of order k. This is called the partial autocorrelogram.

From the definition of PACF it immediately follows that there is no difference between PACF and ACF of order 1:

$$\phi_{11} = \rho_1.$$

For order 2 we have

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & \rho_2 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} \tag{12.18}$$

Example 12.3 (AR(1)) The AR(1) process $X_t = \alpha X_{t-1} + \varepsilon_t$ has the ACF $\rho_\tau = \alpha^\tau$. For the PACF we have $\phi_{11} = \rho_1 = \alpha$ and

$$\phi_{22} = \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2} = \frac{\alpha^2 - \alpha^2}{1 - \alpha^2} = 0,$$

and $\phi_{kk} = 0$ for all $k > 1$. This is plausible since the last coefficient of an AR(k) model for this process is zero for all $k > 1$. For $k = 2$ we illustrate the equivalence with Definition 12.3: From $X_t = \alpha X_{t-1} + \varepsilon_t$ we directly obtain $\mathcal{P}(X_{t+2}|X_{t+1}) = \alpha X_{t+1}$ with

$$\alpha = \frac{\text{Cov}(X_{t+2}, X_{t+1})}{\text{Var}(X_{t+1})}.$$

From the "backward regression" $X_t = \alpha' X_{t+1} + \eta_t$ with white noise η_t it further follows that $\mathcal{P}(X_t|X_{t+1}) = \alpha' X_{t+1}$ with

$$\alpha' = \frac{\text{Cov}(X_t, X_{t+1})}{\text{Var}(X_{t+1})}.$$

For $|\alpha| < 1$ the process is covariance-stationary and it holds that $\mathsf{Cov}(X_{t+2}, X_{t+1}) = \mathsf{Cov}(X_t, X_{t+1}) = \gamma_1$ and $\alpha = \alpha' = \rho_1$. We obtain

$$\mathsf{Cov}\{X_t - \mathcal{P}(X_t|X_{t+1}), X_{t+2} - \mathcal{P}(X_{t+2}|X_{t+1})\}$$
$$= \mathsf{Cov}(X_t - \rho_1 X_{t+1}, X_{t+2} - \rho_1 X_{t+1})$$
$$= \mathsf{E}[(X_t - \rho_1 X_{t+1})(X_{t+2} - \rho_1 X_{t+1})]$$
$$= \gamma_2 - 2\rho_1\gamma_1 + \rho_1^2\gamma_0$$

and

$$\mathsf{Var}\{X_{t+2} - \mathcal{P}(X_{t+2}|X_{t+1})\} = \mathsf{E}[(X_{t+2} - \rho_1 X_{t+1})^2]$$
$$= \gamma_0(1 + \rho_1^2) - 2\rho_1\gamma_1$$
$$= \mathsf{E}[(X_t - \rho_1 X_{t+1})^2]$$
$$= \mathsf{Var}[X_t - \mathcal{P}(X_t|X_{t+1})].$$

With this we get for the partial autocorrelation of 2nd order

$$\phi_{22} = \mathsf{Corr}\{X_t - \mathcal{P}(X_t|X_{t+1}), X_{t+2} - \mathcal{P}(X_{t+2}|X_{t+1})\}$$
$$= \frac{\mathsf{Cov}\{X_t - \mathcal{P}(X_t|X_{t+1}), X_{t+2} - \mathcal{P}(X_{t+2}|X_{t+1})\}}{\sqrt{\mathsf{Var}\{X_{t+2} - \mathcal{P}(X_{t+2}|X_{t+1})\}}\sqrt{\mathsf{Var}(X_t - \mathcal{P}(X_t|X_{t+1}))}}$$
$$= \frac{\gamma_2 - 2\rho_1\gamma_1 + \rho_1^2\gamma_0}{\gamma_0(1 + \rho_1^2) - 2\gamma_1\rho_1}$$
$$= \frac{\rho_2 - \rho_1^2}{1 - \rho_1^2}$$

which corresponds to the results in (12.18). For the AR(1) process it holds that $\rho_2 = \rho_1^2$ and thus $\phi_{22} = 0$.

It holds in general for AR(p) processes that $\phi_{kk} = 0$ for all $k > p$. In Fig. 12.5 the PACF of an AR(2) process is displayed using the parameters as in Fig. 12.4.

Example 12.4 (MA(1)) For a MA(1) process $X_t = \beta\varepsilon_{t-1} + \varepsilon_t$ with $\mathsf{Var}(\varepsilon_t) = \sigma^2$ it holds that $\gamma_0 = \sigma^2(1 + \beta^2)$, $\rho_1 = \beta/(1 + \beta^2)$ and $\rho_k = 0$ for all $k > 1$. For the partial autocorrelations we obtain $\phi_{11} = \rho_1$ and

$$\phi_{22} = \frac{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 0 \end{vmatrix}}{\begin{vmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{vmatrix}} = -\frac{\rho_1^2}{1 - \rho_1^2} \tag{12.19}$$
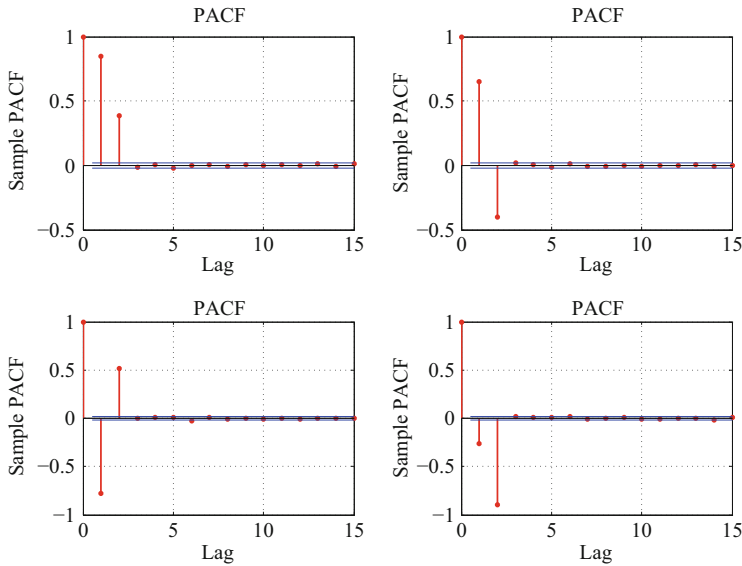
Fig. 12.5 PACF of an AR(2) process with $(\alpha_1 = 0.5, \alpha_2 = 0.4)$ (*top left*), $(\alpha_1 = 0.9, \alpha_2 = -0.4)$ (*top right*), $(\alpha_1 = -0.4, \alpha_2 = 0.5)$ (*bottom left*) and $(\alpha_1 = -0.5, \alpha_2 = -0.9)$ (*bottom right*)
Q SFEpacfar2

For a MA(1) process it strictly holds that $\phi_{22} < 0$. If one were to continue the calculation with $k > 2$, one could determine that the partial autocorrelations will not reach zero.

Figure 12.6 shows the PACF of a MA(2) process. In general for a MA(q) process it holds that the PACF does not decay, in contrast to the autoregressive process. Compare the PACF to the ACF in Fig. 12.2. This is thus a possible criterion for the specification of a linear model.

12.5   Estimation of Moments

Below we assume a stationary stochastic process X_t, i.e. $\mathsf{E}[X_t] = \mu$ and $\mathsf{Cov}(X_t, X_{t+\tau}) = \gamma_\tau$. In the previous sections, we have assumed that we knew the process, and thus, that the moment generating function was also known. In practice, one observes only a realization of the process, $X_1, \ldots, X_n$, and thus there is the problem of estimating the moment generating function.
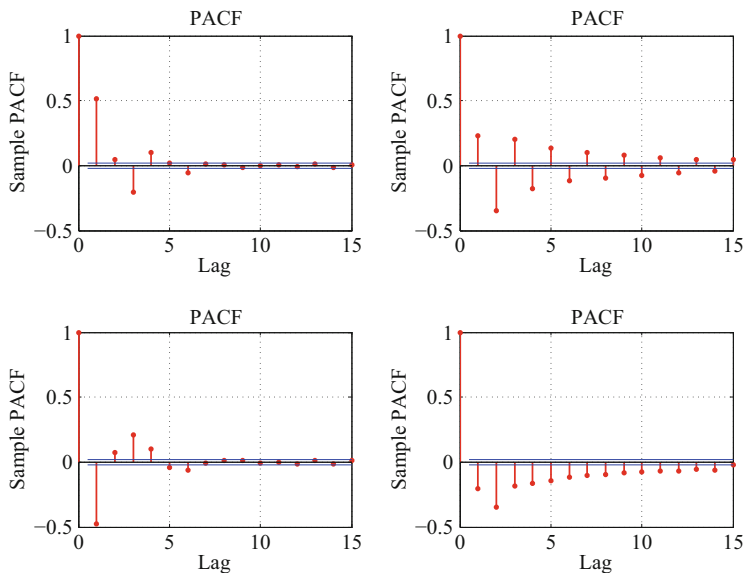
Fig. 12.6 PACF of a MA(2) process with $(\beta_1 = 0.5, \beta_2 = 0.4)$ (*top left*), $(\beta_1 = 0.5, \beta_2 = -0.4)$ (*top right*), $(\beta_1 = -0.5, \beta_2 = 0.4)$ (*bottom left*) and $(\beta_1 = -0.5, \beta_2 = -0.4)$ (*bottom right*)

Q SFEpacfma2

12.5.1   Estimation of the Mean Function

The parameter $\mu = \mathsf{E}[X_t]$ can be estimated with the simple arithmetic sample mean:

$$\bar{X}_n = n^{-1} \sum_{i=1}^{n} X_i. \tag{12.20}$$

The estimator $\bar{X}_n$ is unbiased since it holds that $\mathsf{E}[\bar{X}_n] = \mu$, and its variance is

$$\mathsf{Var}(\bar{X}_n) = \mathsf{Var}(n^{-1} \sum_{i=1}^{n} X_i)$$

$$= n^{-2} \sum_{t=1}^{n} \sum_{s=1}^{n} \mathsf{Cov}(X_t, X_s)$$

$$= n^{-2} \sum_{t=1}^{n} \sum_{s=1}^{n} \gamma_{t-s}$$

$$= n^{-1} \sum_{\tau=-(n-1)}^{n-1} \frac{n - |\tau|}{n} \gamma_\tau$$

When the autocovariance function γ_τ is absolutely summable, it holds that $\mathsf{Var}(\bar{X}_n) < \infty$ and $\lim_{n \to \infty} \mathsf{Var}(\bar{X}_n) = 0$. The estimator X is then also a consistent estimator for μ. In many cases there are more efficient estimators which take advantage of the correlation structure of the process.

The asymptotic variance

$$\lim_{n \to \infty} n \mathsf{Var}(\bar{X}_n) = \gamma_0 + 2 \sum_{\tau=1}^{\infty} \gamma_\tau$$

is denoted as $f(0)$, since this is exactly the spectral density at frequency zero. Under the absolute summability of γ_τ the following asymptotic distribution for the estimator holds:

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathrm{N}(0, f(0)). \tag{12.21}$$

12.5.2   Estimation of the Covariance Function

A possible estimator of the covariance function γ_τ is

$$\hat{\gamma}_{\tau,n} = n^{-1} \sum_{t=1}^{n-\tau} (X_t - \bar{X}_n)(X_{t+\tau} - \bar{X}_n) \tag{12.22}$$

with the mean estimator $\bar{X}_n$ from (12.20). Instead of dividing by n in (12.22) one could also divide by $n - \tau$, although the estimator would then have less favorable properties. The estimator $\hat{\gamma}_{\tau,n}$ is no longer unbiased, since the following can be shown.

$$\mathsf{E}[\hat{\gamma}_{\tau,n}] = \left(1 - \frac{\tau}{n}\right) \gamma_\tau - \left(1 - \frac{\tau}{n}\right) \mathsf{Var}(\bar{X}_n) + \mathcal{O}(n^{-2}).$$

Positive autocovariances are in general underestimated with $\hat{\gamma}_{\tau,n}$. Asymptotically $\hat{\gamma}_{\tau,n}$ is nevertheless unbiased: $\lim_{n \to \infty} \mathsf{E}[\hat{\gamma}_{\tau,n}] = \gamma_\tau$. For the variance when terms of higher order are ignored it holds that:

$$\mathsf{Var}(\hat{\gamma}_{\tau,n}) = n^{-1} \sum_{j=-\infty}^{\infty} (\gamma_j^2 + \gamma_{j-\tau}\gamma_{j+\tau}) + \mathcal{O}(n^{-1}) = n^{-1}\sigma_{\tau,\infty}^2 + \mathcal{O}(n^{-1})$$

and since $\lim_{n \to \infty} \text{Var}(\hat{\gamma}_{\tau,n}) = 0$ holds, $\hat{\gamma}_{\tau,n}$ is a consistent estimator for γ_τ. Furthermore, it can be shown that the covariance estimator behaves asymptotically like a normally distributed random variable:

$$\sqrt{n}(\hat{\gamma}_{\tau,n} - \gamma_\tau) \xrightarrow{\mathcal{L}} N(0, \sigma_{\tau,\infty}^2).$$

12.5.3   Estimation of the ACF

An obvious estimator for the ACF ρ_τ is

$$\hat{\rho}_{\tau,n} = \frac{\hat{\gamma}_{\tau,n}}{\hat{\gamma}_{0,n}}. \tag{12.23}$$

Once again we have a bias of order n^{-1}, i.e.

$$E(\hat{\rho}_{\tau,n}) = \rho_\tau + \mathcal{O}(n^{-1})$$

and $\hat{\rho}_{\tau,n}$ is asymptotically unbiased. For the variance it holds that

$$\text{Var}(\hat{\rho}_{\tau,n}) = n^{-1}\Sigma_{\rho,\tau\tau} + \mathcal{O}(n^{-2}).$$

The estimator $\hat{\rho}_{\tau,n}$ is consistent, since $\lim_{n \to \infty} \text{Var}(\hat{\rho}_{\tau,n}) = 0$. For the asymptotic distribution of the vector $\hat{\rho}_{(k),n} = (\hat{\rho}_{1,n}, \ldots, \hat{\rho}_{k,n})^\top$ it can be shown that

$$\sqrt{n}(\hat{\rho}_{(k),n} - \rho_{(k)}) \xrightarrow{\mathcal{L}} N(0, \Sigma_\rho)$$

with the covariance matrix Σ_ρ with the typical element

$$\Sigma_{\rho,kl} = \sum_{j=-\infty}^{\infty} \rho_j \rho_{j+k+l} + \sum_{j=-\infty}^{\infty} \rho_j \rho_{j+k-l}$$

$$+ 2\rho_k \rho_l \sum_{j=-\infty}^{\infty} \rho_j^2 - 2\rho_l \sum_{j=-\infty}^{\infty} \rho_j \rho_{j+k} - 2\rho_k \sum_{j=-\infty}^{\infty} \rho_j \rho_{j+l}.$$

In particular for the asymptotic variance of $\sqrt{n}(\hat{\rho}_{\tau,n} - \rho_\tau)$, it holds that

$$\Sigma_{\rho,\tau\tau} = \sum_{j=-\infty}^{\infty} \rho_j \rho_{j+2\tau} + \sum_{j=-\infty}^{\infty} \rho_j^2$$

$$+ 2\rho_\tau^2 \sum_{j=-\infty}^{\infty} \rho_j^2 - 4\rho_\tau \sum_{j=-\infty}^{\infty} \rho_j \rho_{j+\tau}.$$

Example 12.5 (MA(q)) For the MA(q) process in (12.1) we know that $\rho_\tau = 0$ for all $\tau > q$. Thus the asymptotic variance can be simplified from $\sqrt{n}(\hat{\rho}_{\tau,n} - \rho_\tau)$ for $\tau > q$ to

$$\Sigma_{\rho,\tau\tau} = 1 + 2 \sum_{i=1}^{q} \rho_i^2.$$

Example 12.6 (White Noise) If X_t is white noise, it holds that

$$\mathsf{E}(\hat{\rho}_{\tau,n}) = -n^{-1} + \mathcal{O}(n^{-2})$$

and

$$\mathsf{Var}(\hat{\rho}_{\tau,n}) = n^{-1} + \mathcal{O}(n^{-2})$$

for $\tau \neq 0$. The asymptotic covariance matrix of $\sqrt{n}(\hat{\rho}_{(k),n} - \rho_{(k)})$ is the identity matrix. Using this we can build approximately 95 % confidence intervals for the ACF: $[-n^{-1} \pm \frac{2}{\sqrt{n}}]$.

12.6   Portmanteau Statistics

With the help of the knowledge about the asymptotic distribution of the autocorrelations we can derive a statistic to test the hypothesis of white noise. One can either test the original series X_t or the residuals of an ARMA(p,q) process. The number of estimated parameters is in the first case $k = 0$ and in the second case $k = p + q$.

Under the null hypothesis it holds for every m

$$\rho_1 = 0, \ldots, \rho_m = 0.$$

The alternative hypothesis is accordingly, that at least one ρ_i, $1 \leq i \leq m$ is not equal to zero. Under the null hypothesis $\sqrt{n}\hat{\rho}_{\tau,n}$ is asymptotically standard normally distributed. The statistic

$$Q_m = n \sum_{j=1}^{m} \hat{\rho}_{j,n}^2$$

has an asymptotic χ^2 distribution with $m - k$ degrees of freedom. One would reject the null hypothesis at a significance level of α, as long as $Q_m > \chi^2_{m-k;\alpha}$, the $(1-\alpha)$-quantile of the Chi-squared distribution with $m - k$ degrees of freedom.

Studies show that Q_m in small samples poorly approximates the asymptotic distribution. This results from the fact that $\hat{\rho}_{\tau,n}$ is a biased estimator for ρ_τ. The bias is stronger for small τ, and thus an asymptotically equivalent statistic can be defined as

$$Q_m^* = n(n+2) \sum_{j=1}^{m} \frac{1}{n-j} \hat{\rho}_{j,n}^2$$

which weights the empirical autocorrelations of smaller order less than those of larger order. The modified Portmanteau statistic Q_m^* is therefore in small samples frequently closer to the asymptotic χ^2 distribution. For large n, both statistics performs equally well.

12.7   Estimation of AR(p) Models

A simple way to estimate the parameters of the autoregressive model

$$X_t = \alpha_1 X_{t-1} + \ldots + \alpha_p X_{t-p} + \varepsilon_t$$

with $\mathsf{Var}(\varepsilon_t) = \sigma^2$ is to use the Yule–Walker equations from (12.10), where the theoretical autocorrelation is replaced with the empirical:

$$\begin{pmatrix} 1 & \hat{\rho}_1 & \cdots & \hat{\rho}_{p-1} \\ \hat{\rho}_1 & 1 & \cdots & \hat{\rho}_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_{p-1} & \hat{\rho}_{p-2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_p \end{pmatrix} = \begin{pmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \vdots \\ \hat{\rho}_p \end{pmatrix}.$$

Solving for $\hat{\alpha}$ gives the Yule–Walker estimator. It is consistent and has an asymptotic normal distribution with covariance matrix $\sigma^2 \Gamma^{-1}$,

$$\Gamma = \begin{pmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \cdots & \gamma_0 \end{pmatrix}. \tag{12.24}$$

The Yule–Walker estimators are asymptotically equivalent to other estimators such as the least squares estimator, in the special case of normally distributed ε_t, and the maximum likelihood (ML) estimator for the normally distributed X_t. In this case, these estimators are also asymptotically efficient.

12.8   Estimation of MA(q) and ARMA(p, q) Models

As soon as *moving average* coefficients are included in an estimated model, the estimation turns out to be more difficult. Consider the example of a simple MA(1) model:

$$X_t = \beta \varepsilon_{t-1} + \varepsilon_t \tag{12.25}$$

with $|\beta| < 1$ and $\mathsf{Var}(\varepsilon_t) = \sigma^2$. A simple estimator for the parameter β is obtained from the Yule–Walker equations $\gamma_0 = \sigma^2(1 + \beta^2)$ and $\gamma_1 = \beta \sigma^2$. By dividing both equations we get $\rho_1 = \gamma_1/\gamma_0 = \beta/(1 + \beta^2)$ and the solution to the quadratic equation is:

$$\beta = \frac{1}{2\rho_1} \pm \sqrt{\frac{1}{4\rho_1^2} - 1}. \tag{12.26}$$

The Yule–Walker estimator replaces in (12.26) the theoretical autocorrelation of first order ρ_1 with the estimator $\hat{\rho}_1$. The estimator is quite simple, but has the disadvantage that it is asymptotically inefficient.

The least squares estimator leads to non-linear systems of equations that can only be solved with iterative numerical algorithms. Using the example of a MA(1) process (12.25) this is illustrated: The LS estimator is defined by

$$\hat{\beta} = \arg\min_{\beta} \sum_{t=2}^{n} \varepsilon_t^2 = \arg\min_{\beta} \sum_{t=2}^{n} (X_t - \beta \varepsilon_{t-1})^2 \tag{12.27}$$

Given that ε_t is not observed, one must turn to the AR(∞) representation of the MA(1) process in order to find the solution, i.e.

$$\varepsilon_t = X_t + \sum_{k=1}^{\infty} (-\beta)^k X_{t-k}. \tag{12.28}$$

Given $X_1, \ldots, X_n$, (12.28) can be approximated by

$$\varepsilon_t = X_t + \sum_{k=1}^{t-1} (-\beta)^k X_{t-k}.$$

Solving the first order conditions

$$\frac{\partial}{\partial \beta} \sum_{t=2}^{n} \varepsilon_t^2 = 0,$$

we obtain a non-linear equation for β, which cannot be explicitly solved. For the minimization problem (12.27) one usually implements numerical optimization methods. The least squares estimator is asymptotically efficient and has asymptotically the same properties as the ML estimator.

In the following we assume a stationary and invertible ARMA(p, q) process with the AR(∞) representation

$$X_t = \sum_{j=1}^{\infty} \pi_j X_{t-j} + \varepsilon_t.$$

Maximum likelihood estimation alludes to the distribution assumptions

$$\varepsilon_t \sim N(0, \sigma^2),$$

under which $X = (X_1, \ldots, X_n)^\top$ have multivariate normal distributions with a density

$$p(x \mid \theta) = (2\pi\sigma^2)^{-n/2} |\Gamma|^{-1/2} \exp\left(-\frac{1}{2\sigma^2} x^\top \Gamma^{-1} x\right)$$

with covariance matrix Γ, which is given in (12.24), and the parameter vector

$$\theta = (\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q; \sigma^2)^\top.$$

The likelihood function L is then a density function interpreted as a function of the parameter vector θ for given observations, i.e. $L(\theta \mid x) = p(x \mid \theta)$. One chooses the respective parameter vector that maximizes the likelihood function for the given observations, i.e. the ML estimator is defined by

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta \mid x).$$

Under the assumption of the normal distribution the logarithm of the likelihood function

$$\log L(\theta \mid x) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log |\Gamma| - \frac{1}{2\sigma^2} x^\top \Gamma^{-1} x \qquad (12.29)$$

takes on a simple form without changing the maximizer $\hat{\theta}$. The log-likelihood function (12.29) is also called the *exact* log-likelihood function. One notices that, in particular, the calculation of the inverse and the determinant of the ($n \times n$) matrix Γ is quite involved for long time series. Therefore one often forms an approximation to the exact likelihood, which is good for long time series. One possibility is use the *conditional* distribution $p(X_t \mid X_{t-1}, \ldots, X_1; \theta)$:

$$L(\theta \mid x) = \prod_{t=1}^{n} p(X_t \mid X_{t-1}, \ldots, X_1; \theta)$$

Under the assumption of normal distributions the conditional distributions are normal with an expected value

$$\mathsf{E}[X_t \mid X_{t-1}, \ldots, X_1]$$

and variance

$$\mathsf{Var}(X_t \mid X_{t-1}, \ldots, X_1).$$

The larger t is, the better the approximation of

$$\mathsf{E}[X_t \mid X_{t-1}, \ldots, X_1, \ldots] = \sum_{j=1}^{\infty} \pi_j X_{t-j}$$

by $\sum_{j=1}^{t-1} \pi_j X_{t-j}$ becomes. The conditional log-likelihood function

$$\log L^b(\theta \mid x) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{t=1}^{n} \left(X_t - \sum_{j=1}^{t-1} \pi_j X_{t-j} \right)^2$$

$$(12.30)$$

can be calculated from the data $X_1, \ldots, X_n$ and optimized with respect to the parameter θ. As an initial value for the numerical optimization algorithm the Yule–Walker estimators, for example, can be used (except in specific cases of asymptotic inefficiency).

To compare the exact and the conditional likelihood estimators consider a MA(1) process (12.25) with $\beta = 0.5$ and $\varepsilon_t \sim N(0, 1)$. The matrix Γ is band diagonal with elements $1 + \beta^2$ on the main diagonal and β on diagonals both above and below it. Two realizations of the process with $n = 10$ and $n = 20$ are shown in Fig. 12.7. Since the process has only one parameter, one can simply search in the region $(-1, 1)$. This is shown for both estimators in Figs. 12.8 ($n = 10$) and 12.9 ($n = 20$). For the process with $n = 10$ one still sees a clear discrepancy between both likelihood functions, which for $n = 20$ can be ignored. Both estimators are in this case quite close to the true parameter 0.5.

Under some technical assumptions the ML estimators are consistent, asymptotically efficient and have an asymptotic normal distribution:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} N(0, J^{-1})$$
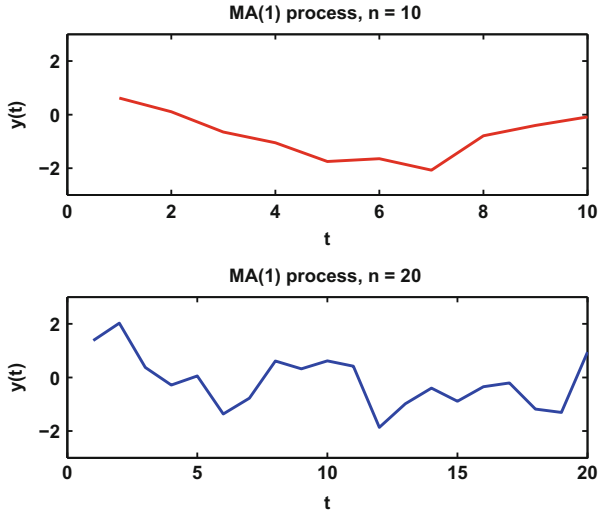
with the Fisher Information matrix

Fig. 12.7 Two realizations of a MA(1) process with $\beta = 0.5$, $\varepsilon_t \sim N(0, 1)$, $n = 10$ (*above*) and $n = 20$ (*below*)   **Q** SFEplotma1
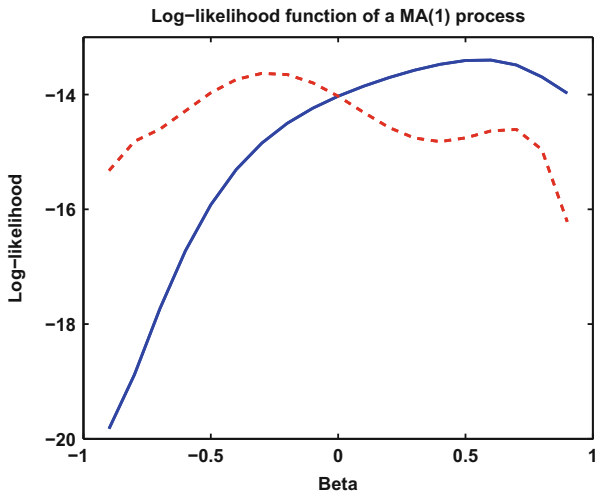

Fig. 12.8 Exact (*solid*) and conditional (*dashed*) likelihood functions for the MA(1) process from Fig. 12.7 with $n = 10$. The true parameter is $\beta = 0.5$ **Q** SFElikma1

$$J = \mathsf{E}\left[-\frac{\partial^2 \log L(\theta, x)}{\partial \theta \, \partial \theta^\top} \right]. \tag{12.31}$$

For the optimization of the likelihood function one frequently uses numerical methods. The necessary condition for a maximum is
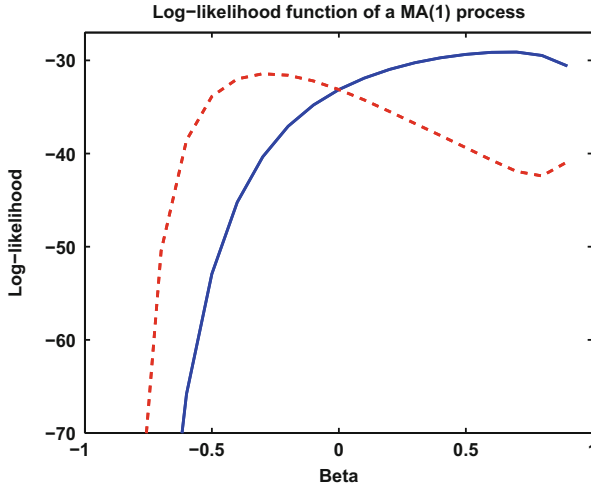
Fig. 12.9 Exact (*solid*) and conditional (*dashed*) likelihood functions for the MA(1) process from Fig. 12.7 with $n = 20$. The true parameter is $\beta = 0.5$ Q SFElikma1

$$\operatorname{grad} l^b(\theta) = 0$$

with $l^b = \log L(\theta \mid x)$. By choosing an initial value θ_0 (for example, the Yule–Walker estimator), and the Taylor approximation

$$\operatorname{grad} l^b(\theta) \approx \operatorname{grad} l^b(\theta_0) + \operatorname{Hess} l^b(\theta_0)(\theta - \theta_0)$$

one obtains the following relation:

$$\theta = \theta_0 - \operatorname{Hess} l^b(\theta_0)^{-1} \operatorname{grad} l^b(\theta_0).$$

Since generally one does not immediately hit the maximizing parameter, one builds the iteration

$$\theta_{j+1} = \theta_j - \operatorname{Hess} l^b(\theta_j)^{-1} \operatorname{grad} l^b(\theta_j)$$

with $j = 1, 2, \ldots$ until a convergence is reached, i.e. $\theta_{j+1} \approx \theta_j$. Often it is easier to use the expectation of the Hessian matrix, that is, the information matrix from (12.31):

$$\theta_{j+1} = \theta_j + J(\theta_j)^{-1} \operatorname{grad} l^b(\theta_j). \tag{12.32}$$

The notation $J(\theta_j)$ here means that (12.31) is evaluated at θ_j. The iteration (12.32) is called the *score*-algorithm or Fisher scoring.

12.9   Recommended Literature

Comprehensive textbooks on classic time series analysis are, for example, Schlittgen and Streitberg (1995), Brockwell and Davis (1991), Gouriéroux and Monfort (1996). In addition, classic books are Box and Jenkins (1976) and Hamilton (1994).

12.10   Exercises

Exercise 12.1 *Which condition do time series have to fulfil in order to be fitted by the ARIMA model? And what does the letter "I" in the word "ARIMA" mean?*

Exercise 12.2 *Suppose that the stationary process X_t has an autocovariance function is given by γ_τ. Find the autocorrelation function (ACF) (in terms of γ_τ) of the (stationary) process Y_t defined as $Y_t = X_t - X_{t-1}$.*

Exercise 12.3 *Calculate the ACF of the MA(1) process $X_t = -0.5\varepsilon_{t-1} + \varepsilon_t$.*

Exercise 12.4 *Find the ACF of the second order moving average process MA(2) defined as*

$$X_t = \varepsilon_t + 0.5\varepsilon_{t-1} - 0.2\varepsilon_{t-2}$$

where ε_t denotes white noise.

Exercise 12.5 *Let*

$$X_t = \sum_{k=0}^{m} \frac{1}{m+1}\varepsilon_{t-k}$$

be the m-th order moving average process MA(m). Show that the autocorrelation function ACF of this process is given by

$$\rho_\tau = \begin{cases} (m+1-k)/(m+1) & \text{if} \quad k = 0, 1, \ldots, m \\ 0 & \text{if} \quad k > m \end{cases}$$

Exercise 12.6

1. *What is meant by saying that a linear process is stationary? How can we evaluate whether a process is stationary?*
2. *What is meant by saying that a linear process is invertible? How can we evaluate whether a process is invertible?*
3. *For each of the following models express the model in terms of the shift operator L acting on ε_t and determine whether the model is stationary or/and invertible or not.*

   (i) $X_t = 0.2X_{t-1} + \varepsilon_t$

(ii) $X_t = \varepsilon_t - 1.5\varepsilon_t + 0.3\varepsilon_{t-2}$

(iii) $X_t = 0.4X_{t-1} + \varepsilon_t - 1.5\varepsilon_{t-1} + 0.3\varepsilon_{t-2}$

Exercise 12.7 *Calculate the partial autocorrelations of first and second order of the AR(1) process $X_t = 0.5X_{t-1} + \varepsilon_t$ by using Yule–Walker equations.*

Exercise 12.8 *Which methods could we use to estimate the parameters in an ARIMA model? What are the advantage and drawback of each of them?*

Exercise 12.9 *Could you give some statistical tests to assess the adequacy of a fitted ARIMA model?*

Exercise 12.10 *What characteristics would one expect of a realization of the MA(1) process $Y_t = 1 + \varepsilon_t + 0.8\varepsilon_{t-1}$? How would these characteristics differ from the those of a realization of the process? $Y_t' = 1 + \varepsilon_t' - 0.8\varepsilon_{t-1}'$*

Exercise 12.11 *Calculate the covariances γ_k for MA(3), the moving average of order 3. Determine the ACF for this process. Plot the ACF for the MA(3) process:*

$$Y_t = 1 + \varepsilon_t + 0.8\varepsilon_{t-1} - 0.5\varepsilon_{t-2} + 0.3\varepsilon_{t-3}.$$

Exercise 12.12 *Derive the ACF for the ARMA(2,1) process:*

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t - \theta_1\varepsilon_{t-1}$$

that is, determine ρ_1, ρ_2, etc., in term of ϕ_1, ϕ_2 and θ_1.

Exercise 12.13 *Derive expressions for the one-, two-, three-period forecast, $\hat{Y}_t(1)$, $\hat{Y}_t(2)$ and $\hat{Y}_t(3)$, for the second-order autoregressive process AR(2). What are the variances of the errors for these forecasts?*

Exercise 12.14 *Suppose an ARMA(0,2) model has been estimated from a time series generated by an ARMA(1,2) process. How would the diagnostic test indicate that the model has been misspecified?*

Exercise 12.15 *Which of the following processes are covariance stationary? Explain your answer for every process.*

(a) $X_t = \varepsilon_t - \varepsilon_{t-1} + 2\varepsilon_{t-2}$

(b) $X_t = 2 + \frac{1}{2}X_{t-1} + \varepsilon_t$

(c) $X_t = 4 - 1.3X_{t-1} + 0.8X_{t-2} + \varepsilon_t$

(d) $X_t = 4 - 1.3X_{t-1} - 0.8X_{t-2} + \varepsilon_t$

Exercise 12.16 *Find the ACF for the following processes:*

(a) *A white noise process with $E(X_t) = 0$, $Var(X_t) = \sigma^2 \; \forall t$*

(b) $X_t = \varepsilon_t - \varepsilon_{t-1}$

(c) *For the MA(1) process defined as $X_t = \varepsilon_t - \theta_1\varepsilon_{t-1}$. Show that you cannot identify an MA(1) process uniquely from the autocorrelation by comparing the results using θ_1 with those if you replace θ_1 by θ_1^{-1}.*

Exercise 12.17 *Consider a first order AR(1) process without drift:*

$$X_t = \alpha X_{t-1} + \varepsilon_t, \quad |\alpha| < 1$$

(a) Find the mean and the variance
(b) Show that for the variance to be finite, $|\alpha|$ must be less than 1.
(c) Find the ACF assuming that the process is stationary.

Exercise 12.18 *Let X_t be a stationary AR(p) process with mean 0, i.e.*

$$X_t = \sum_{i=1}^{p} \alpha_i X_{t-i} + \varepsilon_t$$

Show that the process Y_t defined as

$$Y_t = \sum_{j=0}^{q} \beta_j X_{t-j}$$

where $\beta_0 = 1$ can be written as an ARMA(p, q) process.

Exercise 12.19

1. *Let $X(t)$ be a stochastic process in $\mathbb{R}^3$ defined by the vectorial Ornstein–Uhlenbeck equation $dX(t) = AC + Sdt + e_k\sigma(t)dB(t)$ with e_k being the k-th vector in $\mathbb{R}^3$ with $k = 1, 2, 3$. Further $\sigma(t)$ is a real valued square integrable function and A is the 3×3 matrix.*

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -\alpha_3 & -\alpha_2 & -\alpha_1 \end{pmatrix}$$

   We suppose that $\alpha_k, k = 1, 2, 3$ are constant. Show that by iterating the finite difference approximations of the time dynamics of the CAR(3) process, we can get the time discrete version for $t = 0, 1, 2$. i.e. $X_1(t + 3) \approx (3 - \alpha_1)X_1(t + 2) + (2\alpha_1 - \alpha_2 - 3)X_1(t + 1) + (1 + \alpha_2 - \alpha_3 - \alpha_1)X_1(t) + \sigma(t)d(t)$
2. *The stationary condition of the CAR(3) model says that the eigenvalues of the matrix A need to have negative real parts. Supposing that $\beta_1 = 0.41, \beta_2 = -0.2, \beta_3 = 0.07$ and using the results from the previous question, verify that the stationarity condition holds.*

Exercise 12.20 *Consider the data, COCACOLA.txt containing daily prices (p_t) of the Coca-Cola company from January 2002 to November 2004.*

1. *Display the graph of the time series.*
2. *Plot the ACF of the daily price series up to 100 lags and describe the nature of the decay.*

3. *Test for stationarity of* (p_t) *by any suitable procedure.*
4. *Plot the rate of returns* r_t *using,* $r_t = \frac{(p_t - p_{t-1})}{p_{t-1}}$ *and* $r_t = \log p_t - \log p_{t-1}$.
   Comment on the return pattern.
5. *Model the return rate* r_t *as a GARCH(1,1) process.*

Exercise 12.21 *Given a first order autoregressive model for a series* X_t *with a GARCH$(1, 1)$ process of the error term,*

$$x_t = \theta_0 + \theta_1 x_{t-1} + \varepsilon_t, \tag{12.33}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2 \tag{12.34}$$

where σ_t^2 *is the conditional variance.*

1. *Explain how the GARCH model is a generalization of the ARCH model.*
2. *Discuss what the model in Eq. (12.34) implies for the process of the squared errors,* ε_t^2.
3. *Explain what happens to the model if* $\alpha_1 + \beta_1 = 1$ *and discuss the implication.*

Chapter 13
Time Series with Stochastic Volatility

In the previous chapters we have already discussed that volatility plays an important role in modelling financial systems and time series. Unlike the term structure, volatility is unobservable and thus must be estimated from the data.

Reliable estimations and forecasts of volatility are important for large credit institutes where volatility is directly used to measure risk. The risk premium, for example, is often specified as a function of volatility, and it is interesting to find an appropriate model for volatility. The capability of macroeconomic factors to forecast volatility has already been examined in the literature. Although macroeconomic factors have some forecasting capabilities, the most important factor seems to be the lagged endogenous return. As a result, recent studies are mainly concentrated on time series models.

Stocks, exchange rates, interest rates and other financial time series have *stylized facts* that are different from other time series. A good candidate for modelling of financial time series should represent the properties of stochastic processes. Neither the classic linear AR or ARMA processes nor the non-linear generalizations can fulfil this task. In this chapter we will describe the most popular volatility class of models: the ARCH (*autoregressive conditional heteroscedasticity*) model that can replicate these stylized facts appropriately.

Stylised Fact 1 *Time series of share prices X_t and other basic financial instruments are not stationary time series and possess a local trend at the least.*

Similar to the ARIMA model in Chap. 12, we transform the original data by first taking differences to get a stationary time series. Here we consider the log return (see Definition 11.15) instead of the original share prices. We simply call it *return* in this chapter. One could consider the simple return R_t as well (see Definition 11.14).

Stylised Fact 2 *Returns r_t have a leptokurtic distribution. The empirically estimated kurtosis is mostly greater than 3.*

We discussed the properties of the return's distribution in Sects. 3.3 and 11.2. The leptokurtosis can be illustrated in a comparison of the density of a normal distribution and a kernel estimator of the adjusted data (see Fig. 15.4). We can see in Theorem 13.3 that an ARCH process has a kurtosis greater than 3 even if the innovation of the process itself is normally distributed.

Stylised Fact 3 *The return process is white noise (Definition 11.8) since the sample autocorrelation $\hat{\rho}_{\tau,n}, k \neq 0$ (12.23) is not significantly different from 0. Furthermore the white noise is not independent since the sample autocorrelations of squared and absolute returns are clearly greater than 0.*

ARCH models possess the characteristic (Theorem 13.1) that we have already described in Sect. 11.2. A stronger condition than pairwise uncorrelation of returns is that returns are unpredictable, which is connected to the no-arbitrage condition. As in Sect. 11.3 $\mathcal{F}_t$ denotes the information set at time t. The best prediction of return r_{t+1} at day t for day $t + 1$ is the conditional expectation $r_{t+1|t} = \mathsf{E}[r_{t+1}|\mathcal{F}_t]$ (Theorem 11.1) based on the information set $\mathcal{F}_t$. The time series of the return is considered unpredictable if

$$r_{t+1|t} = \mathsf{E}[r_{t+1}|\mathcal{F}_t] = \mathsf{E}[r_{t+1}],$$

i.e. the best prediction of the next return is simply its unconditional mean. The information set $\mathcal{F}_t$ gives no hints of predicting future prices. ARCH processes are automatically unpredictable (Definition 13.1).

An unpredictable time series is always white noise because the autocorrelation is equal to 0. It is even possible that a linear prediction is better than the expectation estimated only by the unpredictable time series (see the proof of Theorem 13.1). The condition of unpredictability is actually stronger than pairwise uncorrelation. A predictable white noise is, for example $\varepsilon_t = \eta_t + \gamma \eta_{t-1} \eta_{t-2}$, where η_t is independent white noise with the expectation of 0. This bilinear process has vanishing autocorrelations but $\mathsf{E}[\varepsilon_{t+1}|\mathcal{F}_t] = \gamma \eta_t \eta_{t-1} \neq 0 = \mathsf{E}[\varepsilon_{t+1}]$.

If the returns were predictable we could develop a trading strategy based on the resulting predictions of price, which would give us a positive profit. The existence of a stochastic arbitrage probability obviously contradicts the assumption of a perfect financial market (Sect. 2.1).

Stylised Fact 4 *Volatility tends to form clusters: after a large (small) price change (positive or negative) a large (small) price change tends to occur. This effect is called volatility clustering.*

We will consider the properties of financial time series in more detail in the following section. According to stylized fact 4, the squared returns are positively correlated. Thus returns are conditionally heteroscedastic, i.e.

$$\mathsf{Var}[r_{t+1}|\mathcal{F}_t] \neq \mathsf{Var}[r_{t+1}].$$

The returns r_t are not independent but their variability depends on the recent changes of price.

13.1   ARCH and GARCH Models

Following the introduction of ARCH models in the 1980s there were enormous theoretical and practical developments in financial econometrics. It became clear that ARCH models could efficiently and quite easily represent the typical empirical findings in financial time series, e.g. the conditional heteroscedasticity. In particular after the collapse of the Bretton Woods system and the implementation of flexible exchange rates in the 1970s ARCH models became increasingly used by researchers and practitioners.

In addition a far-reaching agreement was formed that returns cannot be regarded as i.i.d. and can, at most, be regarded as being uncorrelated. This argument holds at least for financial time series of relatively high frequency, for example for daily data. In Fig. 13.1 we show a normally distributed white noise, a GARCH(1,1) process in Fig. 13.2, and the DAX and FTSE 100 daily index returns (1998–2007) in Fig. 13.3. It can be seen from the figure that the GARCH process is obviously more appropriate for modelling stock returns than the white noise process.
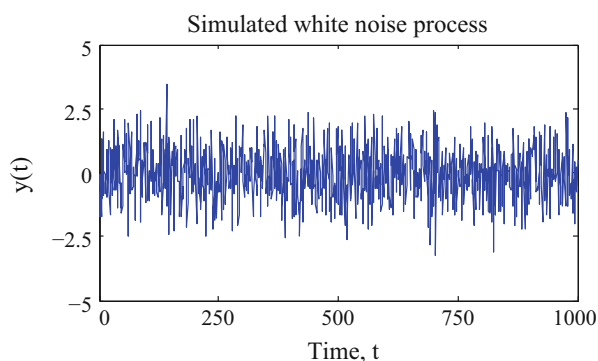

Fig. 13.1 Time series of a normally distributed white noise process   Q SFEtimewn
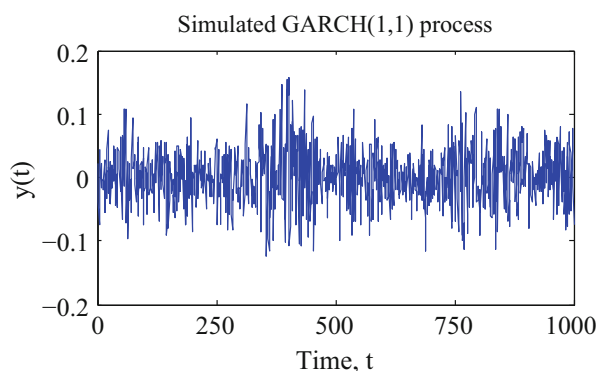

Fig. 13.2 Time series of a GARCH(1,1) process ($\alpha = 0.15$, $\beta = 0.80$)   Q SFEtimegarch

DAX returns
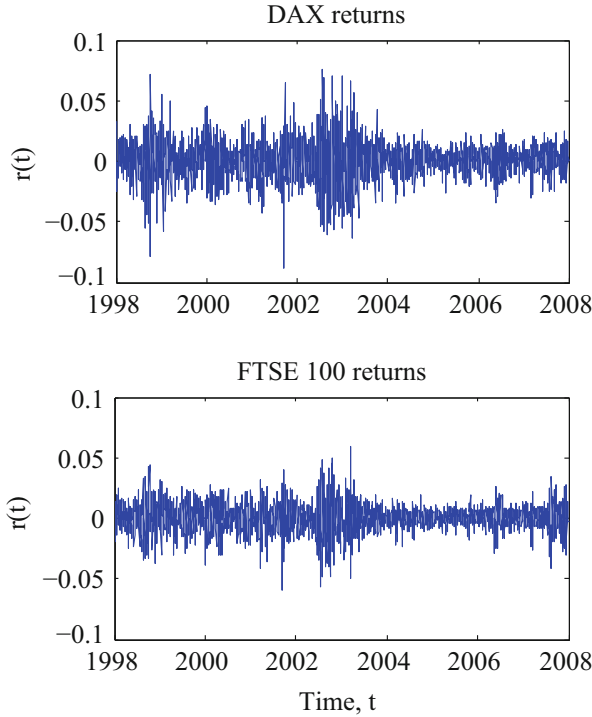

FTSE 100 returns

Fig. 13.3   DAX and FTSE 100 daily returns from 1998 to 2007   **Q** SFEtimeret

However, the ARCH model is only the starting point of the empirical study and relies on a wide range of specification tests. Some practically relevant disadvantages of the ARCH model have been discovered recently, for example, the definition and modelling of the persistence of shocks and the problem of modelling asymmetries. Thus a large number of extensions of the standard ARCH model have been suggested. We will discuss them in detail later.

Let X_t be a discrete stochastic process and from Definition 11.15 $r_t = \log X_t / X_{t-1}$ the relative increase or the *return* of the process X_t. If the returns are almost independent and identically distributed, then X_t follows a geometric random walk. It is assumed in ARCH models that the returns depend on past information with a specific form.

As mentioned previously $\mathcal{F}_t$ denotes the information set at time t, which encompasses X_t and all the past realizations of the process X_t. This means in a general model

$$r_t = \mu_t + \varepsilon_t \tag{13.1}$$

with $\mathsf{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0$. Here μ_t can represent the risk premium which results from the econometric models and is time dependent. The stochastic error term ε_t is no longer independent but centred and uncorrelated. In ARCH models the conditional variance of ε_t is a linear function of the lagged squared error terms.

13.1.1   ARCH(1): Definition and Properties

The ARCH model of order 1, ARCH(1), is defined as follows:

Definition 13.1 (ARCH(1)) The process ε_t, $t \in \mathbb{Z}$, is ARCH(1), if $\mathsf{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0$,

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 \tag{13.2}$$

with $\omega > 0$, $\alpha \geq 0$ and

- $\mathsf{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t^2$ and $Z_t = \varepsilon_t/\sigma_t$ is i.i.d. (strong ARCH)
- $\mathsf{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t^2$ (semi-strong ARCH),
- $\mathcal{P}(\varepsilon_t^2 \mid 1, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \ldots) = \sigma_t^2$ (weak ARCH),

where $\mathcal{P}$ is the best linear projection described in Sect. 12.4. Obviously a strong ARCH(1) process is also semi-strong and a semi-strong also weak. On the other hand the conditional variance of a weak ARCH(1) process can be non-linear (unequal to σ_t^2). In this case it cannot be a semi-strong ARCH process.

Setting $Z_t = \varepsilon_t/\sigma_t$, it holds for the semi-strong and the strong ARCH models that $\mathsf{E}[Z_t] = 0$ and $\mathsf{Var}(Z_t) = 1$. In strong ARCH models Z_t is i.i.d. so that no dependence can be modelled in higher moments than the second moment. It is frequently assumed that Z_t is normally distributed, which means ε_t is conditionally normally distributed:

$$\varepsilon_t \mid \mathcal{F}_{t-1} \sim \mathsf{N}(0, \sigma_t^2). \tag{13.3}$$

Under (13.3) the difference between the strong and the semi-strong ARCH models disappears.

Originally only strong and semi-strong ARCH models are discussed in the literature. Weak ARCH models are important because they are closed under temporal aggregation. If, for example, daily returns follow a weak ARCH process, then the weekly and monthly returns are also weak ARCH with corresponding parameter adjustments. This phenomenon holds in general for strong and semi-strong models.

According to Definition 13.1 the process ε_t is a martingale difference and is therefore white noise.

Theorem 13.1 *Assume that the process ε_t is a weak ARCH(1) process with* $\mathsf{Var}(\varepsilon_t) = \sigma^2 < \infty$. *Then it follows that ε_t is white noise.*

Proof From $\mathsf{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0$ it follows that $\mathsf{E}[\varepsilon_t] = 0$ and $\mathsf{Cov}(\varepsilon_t, \varepsilon_{t-k}) = \mathsf{E}[\varepsilon_t \varepsilon_{t-k}] = \mathsf{E}[\mathsf{E}(\varepsilon_t \varepsilon_{t-k} \mid \mathcal{F}_{t-1})] = \mathsf{E}[\varepsilon_{t-k} \, \mathsf{E}(\varepsilon_t \mid \mathcal{F}_{t-1})] = 0.$ □

Note that ε_t is not an independent white noise.

Theorem 13.2 (Unconditional Variance of the ARCH(1)) *Assume the process ε_t is a semi-strong ARCH(1) process with* $\mathsf{Var}(\varepsilon_t) = \sigma^2 < \infty$. *Then it holds that*

$$\sigma^2 = \frac{\omega}{1 - \alpha}.$$

Proof $\sigma^2 = \mathsf{E}[\varepsilon_t^2] = \mathsf{E}[\mathsf{E}(\varepsilon_t^2 \mid \mathcal{F}_{t-1})] = \mathsf{E}[\sigma_t^2] = \omega + \alpha \, \mathsf{E}[\varepsilon_{t-1}^2] = \omega + \alpha \sigma^2.$ It holds then $\sigma^2 = \omega/(1 - \alpha)$ when $\alpha < 1$. □

$\alpha < 1$ is the necessary and sufficient condition for a weak stationarity of a semi-strong process.

If the innovation $Z_t = \varepsilon_t/\sigma_t$ is symmetrically distributed around zero, then all odd moments of ε_t are equal to zero. Under the assumption of normal distribution (13.3) the conditions for the existence of higher even moments can be derived.

Theorem 13.3 (Fourth Moment) *Let ε_t be a strong ARCH(1) process, $Z_t \sim N(0, 1)$ and $\mathsf{E}[\varepsilon_t^4] = c < \infty$. Then*

1.

$$\mathsf{E}[\varepsilon_t^4] = \frac{3\omega^2}{(1 - \alpha)^2} \frac{1 - \alpha^2}{1 - 3\alpha^2}$$

   with $3\alpha^2 < 1$.
2. *the unconditional distribution of ε_t is leptokurtic.*

Proof 1. $c = \mathsf{E}[\varepsilon_t^4] = \mathsf{E}[\mathsf{E}(\varepsilon_t^4 \mid \mathcal{F}_{t-1})] = \mathsf{E}[\sigma_t^4 \, \mathsf{E}(Z_t^4 \mid \mathcal{F}_{t-1})] = \mathsf{E}[Z_t^4] \, \mathsf{E}[(\omega + \alpha_1 \varepsilon_{t-1}^2)^2] = 3(\omega^2 + 2\omega\alpha \, \mathsf{E}[\varepsilon_{t-1}^2] + \alpha^2 \, \mathsf{E}[\varepsilon_{t-1}^4]).$ Since $\mathsf{E}[\varepsilon_{t-1}^2] = \omega/(1 - \alpha)$ and $\mathsf{E}[\varepsilon_{t-1}^4] = c$, after rearranging the claim follows.
2.

$$\mathrm{Kurt}(\varepsilon_t) = \frac{\mathsf{E}[\varepsilon_t^4]}{\mathsf{E}[\varepsilon_t^2]^2} = 3 \frac{1 - \alpha^2}{1 - 3\alpha^2} \geq 3. \qquad \square$$

For the boundary case $\alpha = 0$ and the normally distributed innovations $\mathrm{Kurt}(\varepsilon_t) = 3$, while for $\alpha > 0$ it holds that $\mathrm{Kurt}(\varepsilon_t) > 3$. The unconditional distribution is also leptokurtic under conditional heteroscedasticity, i.e. the curvature is high in the middle of the distribution and the tails are fatter than those of a normal distribution, which is frequently observed in financial markets.

The thickness of the tails and thus the existence of moments depend on the parameters of the ARCH models. The variance of the ARCH(1) process is finite when $\alpha < 1$ (Theorem 13.2), while the fourth moment in the case of normally distributed error terms exists when $3\alpha^2 < 1$ (Theorem 13.3). As early as the 1960s Mandelbrot questioned the existence of the variance of several financial time series. Frequently empirical distributions have so fat tails that one cannot conclude a finite variance. In order to make empirical conclusions on the degree of the tail's thickness of the unconditional distribution, one can assume, for example, that the distribution is a Pareto type, i.e. for large x:

$$P(x) = P(X_t > x) \sim kx^{-a}$$

for $a > 0$. When $a > c$, it holds that $\mathsf{E}[|X_t|^c] < \infty$. The question is, how can we estimate the tail index a? A simple method follows from the conclusion that for large x the log function $P(x)$ is linear, i.e.

$$\log P(x) \approx \log k - a \log x. \tag{13.4}$$

Therefore we can build the order statistics $X_{(1)} > X_{(2)} > \ldots > X_{(n)}$ and estimate the probability $P(x)$ for $x = X_{(i)}$ using the relative frequency

$$\frac{\#\{t; X_t \geq X_{(i)}\}}{n} = \frac{i}{n}.$$

In (13.4) $P(X_{(i)})$ is replaced with the estimator i/n:

$$\log \frac{i}{n} \approx \log k - a \log X_{(i)}, \tag{13.5}$$

from which a can be estimated from the regression of i/n on $X_{(i)}, i = 1, \ldots, n$, using the least squares method. In general only a small part of the data will be used for the regression, since the linear approximation of $\log P(x)$ is only appropriate in the tail. Thus only the largest order statistics are used to estimate the regression (13.5). Figure 13.4 shows the regression (13.5) for the DAX and FTSE 100 returns from 1998 to 2007 with $m = 10$, i.e. we choose the 10 largest observations. The slope of the least squares (LS) lines are 7.46 and 6.17 for the DAX and FTSE 100 returns, respectively. Thus, both return processes have finite first four moments.

Hill (1975) has suggested an estimator using the maximum likelihood method:

$$\hat{a} = \left(\frac{1}{m-1} \sum_{i=1}^{m} \log X_{(i)} - \log X_{(m)} \right)^{-1}, \tag{13.6}$$

where m is the number of observations taken from the tail and used in the estimation. How to choose m obviously raises a problem. When m is too large,

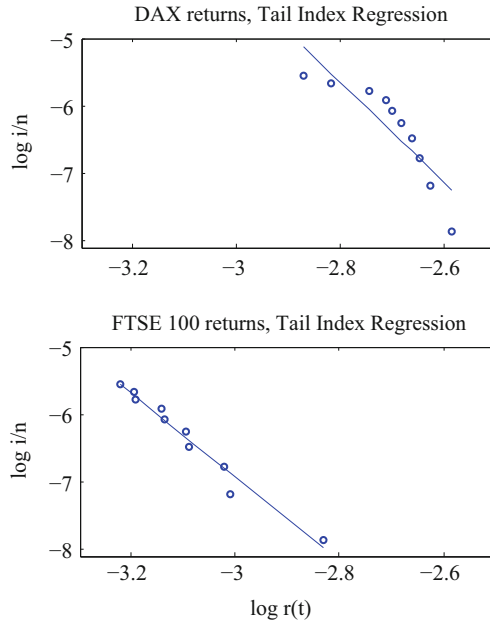**Fig. 13.4** The right side of the logged empirical distribution of the DAX and FTSE 100 daily returns from 1998 to 2007   Q SFEtail

the approximation of the distribution is no longer good; when m is too small, the bias and the variance of the estimator could increase. A simple rule of thumb says that m/n should be around 0.5 or 1 %. Clearly one requires a large amount of data in order to estimate a well. As an example we can consider again the DAX and FTSE 100 indices, as well as 20 German and 20 British stocks, a total of 2,608 observations per index or stock. The results of the ordinary least squares estimator and the Hill estimator based on the analysed return processes with $m = 10$ and $m = 20$ are given in Tables 13.1 and 13.2, for the Frankfurt and London stock exchanges, respectively. In almost every case the estimators are larger than 2, which indicates the existence of variances. Variance may not exist for Bayer and British American Tobacco. The third moment may not exist in a few additional cases in both markets.

Theorem 13.4 (Representation of an ARCH(1) Process) *Let ε_t be a strong ARCH(1) process with $\mathsf{Var}(\varepsilon_t) = \sigma^2 < \infty$. It holds that*

$$\varepsilon_t^2 = \omega \sum_{k=0}^{\infty} \alpha^k \prod_{j=0}^{k} Z_{t-j}^2$$

and the sum converges in L_1.

Proof Through the recursive substitution of $\varepsilon_s^2 = \sigma_s^2 Z_s^2$ and $\sigma_s^2 = \omega + \alpha \varepsilon_{s-1}^2$. The convergence follows from

Table 13.1 Least Square (LS) and Hill estimators of the tail index a for returns of the DAX index and selected 20 German stocks with m observations used for the estimation from 1998 to 2007   ⊙ SFElshill

	LS		Hill	
m	10	25	10	25
DAX	7.46	4.34	5.44	3.49
ADIDAS	4.22	4.79	5.51	5.24
ALLIANZ	4.09	3.35	3.50	3.29
BASF	3.17	3.79	4.13	5.11
BAYER	1.37	2.08	3.05	3.77
BMW	4.75	4.61	5.12	4.67
COMMERZBANK	2.84	3.76	5.35	4.23
CONTINENTAL	4.66	3.89	6.85	3.71
DEUTSCHE BANK	4.17	4.14	7.30	4.05
DEUTSCHE TELEKOM	4.12	4.88	6.43	5.71
EON	4.98	4.58	5.27	4.23
HENKEL	3.80	4.24	4.13	4.42
LINDE	4.25	4.75	5.79	4.87
LUFTHANSA	2.99	3.71	4.27	3.96
MAN	4.43	5.19	5.57	6.45
MUENCHENER RUECK	3.45	3.19	3.85	2.84
RWE	2.75	3.21	3.79	4.06
SAP	2.54	3.08	3.36	3.85
SIEMENS	3.16	3.82	4.02	4.86
THYSSENKRUPP	4.84	5.75	6.42	7.03
VOLKSWAGEN	7.89	4.81	7.41	4.60

$$\mathsf{E}[\varepsilon_t^2 - \omega \sum_{k=0}^{m} \alpha^k \prod_{j=0}^{k} Z_{t-j}^2] = \alpha^{m+1} \mathsf{E}[\varepsilon_{t-m-1}^2 \prod_{j=0}^{m} Z_{t-j}^2]$$

$$= \alpha^{m+1} \mathsf{E}[\varepsilon_{t-m-1}^2] \longrightarrow 0$$

for $m \longrightarrow \infty$, since Z_t is independent with $\mathsf{E}(Z_t^2) = 1$.   □

Theorem 13.5 *Let ε_t be a stationary strong ARCH(1) process with $\mathsf{E}(\varepsilon_t^4) = c < \infty$ and $Z_t \sim N(0,1)$. It holds that*

1.

$$\varepsilon_t^2 = \omega \sum_{k=0}^{\infty} \alpha^k \prod_{j=0}^{k} Z_{t-j}^2$$

   and the sum converges in L_2.
2. *$\eta_t = \sigma_t^2(Z_t^2 - 1)$ is white noise.*
3. *ε_t^2 is an AR(1) process with $\varepsilon_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \eta_t$.*

Table 13.2 Least Square (LS) and Hill estimators of the tail index a for returns of the FTSE 100 index and selected 20 British stocks with m observations used for the estimation from 1998 to 2007   🔍 SFElshill

	LS		Hill	
m	10	25	10	25
FTSE 100	6.17	4.74	6.93	5.19
ASTRAZENECA	3.90	3.43	4.12	3.70
BARCLAYS	7.13	6.14	6.13	6.09
BG GROUP	5.44	4.41	4.37	4.90
BHP BILLITON	3.41	3.40	3.08	3.84
BP	3.55	4.39	4.78	5.09
BRITISH AMERICAN TOBACCO	1.75	2.21	3.12	2.44
BT GROUP	6.19	4.62	6.45	3.48
DIAGEO	4.90	3.90	4.96	3.66
GLAXO SMITH KLINE	2.07	2.60	3.04	3.52
HBOS	2.26	2.81	3.70	3.61
HSBC	4.30	4.39	5.86	3.89
LLOYDS TSB GROUP	4.99	4.42	7.35	3.36
NATIONAL GRID	2.73	3.11	3.49	3.79
RIO TINTO	2.46	2.99	5.15	3.44
ROYAL BANK OF SCOT	6.41	4.89	7.36	4.45
ROYAL DUTCH SHELL A	4.52	5.40	6.78	5.61
STANDARD CHARTERED	3.88	4.07	5.27	4.61
TESCO	4.38	4.45	5.22	4.30
UNILEVER	6.50	3.86	7.21	3.64
VODAFONE GROUP	4.76	5.00	5.96	5.54

Proof 1. As in Theorem 13.4. The convergence is L_2 follows from

$$\mathsf{E}[(\varepsilon_t^2 - \omega \sum_{k=0}^{m} \alpha^k \prod_{j=0}^{k} Z_{t-j}^2)^2] = \mathsf{E}[(\alpha^{m+1} \varepsilon_{t-m-1}^2 \prod_{j=0}^{m} Z_{t-j}^2)^2]$$

$$= \alpha^{2(m+1)} 3^{m+1} \mathsf{E}[\varepsilon_{t-m-1}^4]$$

$$= \alpha^{2(m+1)} 3^{m+1} c \longrightarrow 0$$

for $m \longrightarrow \infty$, since $3\alpha^2 < 1$ due to the assumption that $\mathsf{E}(\varepsilon_t^4)$ is finite and since Z_t is independent with $\mathrm{Kurt}(Z_t) = 3$.

2.(a) $\mathsf{E}[\eta_t] = \mathsf{E}[\sigma_t^2] \mathsf{E}[Z_t^2 - 1] = 0$

(b)

$$\mathrm{Var}(\eta_t) = \mathsf{E}[\sigma_t^4] \mathsf{E}[(Z_t^2 - 1)^2] = 2 \mathsf{E}[(\omega + \alpha \varepsilon_{t-1}^2)^2]$$

$$= 2(\omega^2 + 2\alpha\omega \mathsf{E}[\varepsilon_{t-1}^2] + \alpha^2 \mathsf{E}[\varepsilon_{t-1}^4]) = \text{const.}$$

is independent of t.

(c)

$$\begin{aligned}
\mathsf{Cov}(\eta_t, \eta_{t+s}) &= \mathsf{E}[\sigma_t^2(Z_t^2 - 1)\sigma_{t+s}^2(Z_{t+s}^2 - 1)] \\
&= \mathsf{E}[\sigma_t^2(Z_t^2 - 1)\sigma_{t+s}^2]\,\mathsf{E}[(Z_{t+s}^2 - 1)] \\
&= 0 \qquad \text{for } s \neq 0.
\end{aligned}$$

3. It follows from rearranging: $\varepsilon_t^2 = \sigma_t^2 Z_t^2 = \sigma_t^2 + \sigma_t^2(Z_t^2 - 1) = \omega + \alpha\varepsilon_{t-1}^2 + \eta_t$.
$\square$

Remark 13.1 Nelson (1990a) shows that the strong ARCH(1) process ε_t is strictly stationary when $\mathsf{E}[\log(\alpha Z_t^2)] < 0$. If, for example, $Z_t \sim N(0, 1)$, then the condition for strict stationarity is $\alpha < 3.5622$, which is weaker than the condition for covariance-stationarity with $\alpha < 1$ due to the assumption that the variance is finite.

The dynamics of the volatility process in the case of ARCH(1) is essentially determined by the parameter α. In Theorem 13.5 it was shown that the square of an ARCH(1) process follows an AR(1) process. The correlation structure of the empirical squared observations of returns is frequently more complicated than a simple AR(1) process. In Sect. 13.1.3 we will consider an ARCH model of order q with $q > 1$, which allows a more flexible modelling of the correlation structure.

The volatility is a function of the past squared observations in ARCH models in a narrow sense. In the more general GARCH models (Sect. 13.1.5) it may depend, in addition, on the past squared volatilities. These models belong to the large group of unpredictable time series with *stochastic volatility*. In the strong form, they have $\varepsilon_t = \sigma_t Z_t$ where σ_t is $\mathcal{F}_{t-1}$-measurable, i.e. the volatility σ_t depends only on the information to the time point $t - 1$ and the i.i.d. innovations Z_t with $\mathsf{E}[Z_t] = 0, \mathsf{Var}(Z_t) = 1$. For such a time series it holds that $\mathsf{E}[\varepsilon_t | \mathcal{F}_{t-1}] = 0, \mathsf{Var}(\varepsilon_t | \mathcal{F}_{t-1}) = \sigma_t^2$, i.e. ε_t is unpredictable and except in the special case that $\sigma_t \stackrel{\text{def}}{=} \text{const.}$, conditionally heteroscedastic. The stylized facts 2–4 are only fulfilled under certain qualitative assumptions. For example, in order to produce a volatility cluster, σ_t must tend to be large when the squared observations or volatilities of the recent past observations are large. The generalizations of the ARCH models observed in this section fulfil the corresponding conditions.

Remark 13.2 At first glance stochastic volatility models in discrete time deliver a different approach in modelling the financial data compared with diffusion processes, on which the Black–Scholes model and its generalization are based (Sect. 5.4). Nelson (1990b) has however shown that ARCH and also the more general GARCH processes converge in the limit to a diffusion process in continuous time when the difference of the time points of the successive observations goes against zero.

This result is often used reversely in order to estimate the parameter of financial models in the continuous time where one approximates the corresponding diffusion processes through discrete GARCH time series and estimates its parameter. Nelson (1990b) shows only the convergence of GARCH processes against diffusion

processes in a weak sense (convergence on the distribution). A recent work by Wang (2002) however, shows that the approximation does not hold in a stronger sense, in particular the likelihood process is not asymptotically equivalent. In this sense the maximum likelihood estimators for the discrete time series do not converge against the parameters of the diffusion limit process.

13.1.2 Estimation of ARCH(1) Models

Theorem 13.5 says that an ARCH(1) process can be represented as an AR(1) process in X_t^2. A simple Yule–Walker estimator uses this property:

$$\hat{\alpha}^{(0)} = \frac{\sum_{t=2}^{n}(\varepsilon_t^2 - \hat{\omega}^{(0)})(\varepsilon_{t-1}^2 - \hat{\omega}^{(0)})}{\sum_{t=2}^{n}(\varepsilon_t^2 - \hat{\omega}^{(0)})^2}$$

with $\hat{\omega}^{(0)} = n^{-1} \sum_{t=1}^{n} \varepsilon_t^2$. Since the distribution of ε_t^2 is naturally not normal, the Yule–Walker estimator is inefficient. However it can be used as an initial value for iterative estimation methods.

The estimation of ARCH models is normally carried out using the maximum likelihood (ML) method. Assuming that the returns ε_t have a conditionally normal distribution, we have:

$$p(\varepsilon_t \mid \mathcal{F}_{t-1}) = \frac{1}{\sqrt{2\pi}\sigma_t} \exp\left\{-\frac{1}{2}\frac{\varepsilon_t^2}{\sigma_t^2}\right\}, \tag{13.7}$$

The log-likelihood function $l(\omega, \alpha)$ can be written as a function of the parameters ω and α:

$$l(\omega, \alpha) = \sum_{t=2}^{n} l_t(\omega, \alpha) + \log p_\varepsilon(\varepsilon_1) \tag{13.8}$$

$$= \sum_{t=2}^{n} \log p(\varepsilon_t \mid \mathcal{F}_{t-1}) + \log p_\varepsilon(\varepsilon_1)$$

$$= -\frac{n-1}{2}\log(2\pi) - \frac{1}{2}\sum_{t=2}^{n}\log(\omega + \alpha\varepsilon_{t-1}^2)$$

$$-\frac{1}{2}\sum_{t=2}^{n}\frac{\varepsilon_t^2}{\omega + \alpha\varepsilon_{t-1}^2} + \log p_\varepsilon(\varepsilon_1),$$

where p_ε is the stationary marginal density of ε_t. A problem is that the analytical expression for p_ε is unknown in ARCH models thus (13.8) cannot be calculated. In the conditional likelihood function $l^b = \log p(\varepsilon_n, \ldots, \varepsilon_2 \mid \varepsilon_1)$ the expression $\log p_\varepsilon(\varepsilon_1)$ disappears:

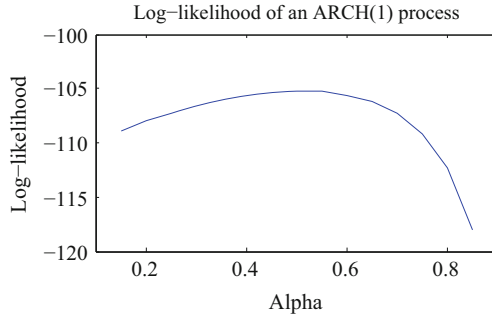**Fig. 13.5** Conditional log-likelihood function of a generated ARCH(1) process with $n = 100$. The true parameter is $\alpha = 0.50$ SFElikarch1

$$l^b(\omega, \alpha) = \sum_{t=2}^{n} l_t(\omega, \alpha) \tag{13.9}$$

$$= \sum_{t=2}^{n} \log p(\varepsilon_t \mid \mathcal{F}_{t-1})$$

$$= -\frac{n-1}{2} \log(2\pi) - 1/2 \sum_{t=2}^{n} \log(\omega + \alpha \varepsilon_{t-1}^2) - 1/2 \sum_{t=2}^{n} \frac{\varepsilon_t^2}{\omega + \alpha \varepsilon_{t-1}^2}.$$

For large n the difference $l - l^b$ is negligible.

Figure 13.5 shows the conditional likelihood of a generated ARCH(1) process with $n = 100$. The parameter ω is chosen so that the unconditional variance is constant everywhere, i.e. with a variance of σ^2, $\omega = (1 - \alpha)\sigma^2$. The optimization of the likelihood of an ARCH(1) model can be found by analysing the graph. Most often, we would like to know the precision of the estimator as well. Essentially it is determined by the second derivative of the likelihood at the optimization point by the asymptotic properties of the ML estimator (see Sect. 13.1.6). Furthermore one has to use numerical methods such as the score algorithm introduced in Sect. 12.8 to estimate the parameters of the models with a larger order. In this case the first and second partial derivatives of the likelihood must be calculated.

With the ARCH(1) model these are:

$$\frac{\partial l_t^b}{\partial \omega} = \frac{1}{2\sigma_t^2} \left(\frac{\varepsilon_t^2}{\sigma_t^2} - 1 \right) \tag{13.10}$$

$$\frac{\partial l_t^b}{\partial \alpha} = \frac{1}{2\sigma_t^2} \varepsilon_{t-1}^2 \left(\frac{\varepsilon_t^2}{\sigma_t^2} - 1 \right) \tag{13.11}$$

$$\frac{\partial^2 l_t^b}{\partial \omega^2} = -\frac{1}{2\sigma_t^4} \left(2 \frac{\varepsilon_t^2}{\sigma_t^2} - 1 \right) \tag{13.12}$$

$$\frac{\partial^2 l_t^b}{\partial \omega^2} = -\frac{1}{2\sigma_t^4}\left(2\frac{\varepsilon_t^2}{\sigma_t^2} - 1\right) \tag{13.13}$$

$$\frac{\partial^2 l_t^b}{\partial \alpha^2} = -\frac{1}{2\sigma_t^4}\varepsilon_{t-1}^4\left(2\frac{\varepsilon_t^2}{\sigma_t^2} - 1\right) \tag{13.14}$$

$$\frac{\partial^2 l_t^b}{\partial \omega \partial \alpha} = -\frac{1}{2\sigma_t^4}\varepsilon_{t-1}^2\left(2\frac{\varepsilon_t^2}{\sigma_t^2} - 1\right). \tag{13.15}$$

The first order conditions are $\sum_{t=2}^{n} \partial l_t^b/\partial \omega = 0$ and $\sum_{t=2}^{n} \partial l_t^b/\partial \alpha = 0$. The expected value of the second derivative has to be calculated for the same algorithm. It is assumed that $\mathsf{E}[Z_t^2] = \mathsf{E}[(\varepsilon_t/\sigma_t)^2] = 1$, so that the expression in the parentheses $(2\varepsilon_t^2/\sigma_t^2 - 1)$ has an expected value of one. From this it follows that:

$$\mathsf{E}\left[\frac{\partial^2 l_t^b}{\partial \omega^2}\right] = -\frac{1}{2}\mathsf{E}\left[\frac{1}{\sigma_t^4}\right].$$

The expectation of σ_t^{-4} is consistently estimated by $(n-1)^{-1}\sum_{t=2}^{n}(\omega + \alpha\varepsilon_{t-1}^2)^{-2}$, so that for the estimator of the expected value of the second derivative we have:

$$\hat{\mathsf{E}}\frac{\partial^2 l_t^b}{\partial \omega^2} = -\frac{1}{2(n-1)}\sum_{t=2}^{n}\frac{1}{\sigma_t^4}.$$

Similarly the expected value of the second derivative with respect to α follows with:

$$\mathsf{E}\left[\frac{\partial^2 l_t^b}{\partial \alpha^2}\right] = -\frac{1}{2}\mathsf{E}\left[\frac{\varepsilon_{t-1}^4}{\sigma_t^4}\right]$$

and the estimator is:

$$\hat{\mathsf{E}}\frac{\partial^2 l_t^b}{\partial \alpha^2} = -\frac{1}{2(n-1)}\sum_{t=2}^{n}\frac{\varepsilon_{t-1}^4}{\sigma_t^4}.$$

Theorem 13.6 *Given $Z_t \sim N(0,1)$, it holds that:*

$$\mathsf{E}\left[\left(\frac{\partial l_t^b}{\partial \omega}\right)^2\right] = -\mathsf{E}\left[\frac{\partial^2 l_t^b}{\partial \omega^2}\right]$$

Proof This follows immediately from $\mathsf{E}\left[\left(\frac{\partial l_t^b}{\partial \omega}\right)^2\right] = \mathsf{E}\left[\frac{1}{4\sigma_t^4}(Z_t^4 - 2Z_t^2 + 1)\right] =$ $\mathsf{E}\left[\frac{1}{4\sigma_t^4}\right](3 - 2 + 1)$. $\qquad\qquad\square$

Obviously Theorem 13.6 also holds for the parameter α in place of ω. In addition it essentially holds for more general models, for example the estimation of GARCH

models in Sect. 13.1.6. In more complicated models one can replace the second derivative with the square of the first derivative, which is easier to calculate. It is assumed, however, that the likelihood function is correctly specified, i.e. the true distribution of the error terms is normal.

Under the two conditions

1. $\mathsf{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathsf{E}[Z_t^2 \mid \mathcal{F}_{t-1}] = 1$
2. $\mathsf{E}[\log(\alpha Z_t^2) \mid \mathcal{F}_{t-1}] < 0$ (strict stationarity)

and under certain technical conditions, the ML estimators are consistent. If $\mathsf{E}[Z_t^4 \mid \mathcal{F}_{t-1}] < \infty$ and $\omega > 0, \alpha > 0$ hold in addition, then $\hat{\theta} = (\hat{\omega}, \hat{\alpha})^\top$ is asymptotically normally distributed:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathrm{N}(0, J^{-1}IJ^{-1}) \tag{13.16}$$

with

$$I = \mathsf{E}\left(\frac{\partial l_t(\theta)}{\partial \theta} \frac{\partial l_t(\theta)}{\partial \theta^\top}\right)$$

and

$$J = -\mathsf{E}\left(\frac{\partial^2 l_t(\theta)}{\partial \theta \partial \theta^\top}\right).$$

If the true distribution of Z_t is normal, then $I = J$ and the asymptotic covariance matrix is simplified to J^{-1}, i.e. the inverse of the Fisher Information matrix. If the true distribution is instead leptokurtic, then the maximum of (13.9) is still consistent, but no longer efficient. In this case the ML method is interpreted as the "Quasi Maximum Likelihood" (QML) method.

In a Monte Carlo simulation study in Shephard (1996) 1000 ARCH(1) processes with $\omega = 0.2$ and $\alpha = 0.9$ were generated and the parameters were estimated using QML. The results are given in Table 13.3. Obviously with the moderate sample sizes ($n = 500$) the bias is negligible. The variance, however, is still so large that a

Table 13.3 Monte Carlo simulation results for QML estimates of the parameter $\alpha = 0.9$ from an ARCH(1) model with $k = 1,000$ replications

n	$k^{-1} \sum_{j=1}^{k} \hat{\alpha}_j$	$\sqrt{k^{-1} \sum_{j=1}^{k} (\hat{\alpha}_j - \alpha)^2}$	$\#(\alpha_j \geq 1)$ (%)
100	0.852	0.257	27
250	0.884	0.164	24
500	0.893	0.107	15
1,000	0.898	0.081	10

The last column gives the proportion of the estimator that is larger than 1 (according to Shephard (1996))

relatively large proportion (10 %) of the estimators are larger than one, which would imply covariance non-stationarity. This, in turn, has a considerable influence on the volatility prediction.

13.1.3 ARCH(q): Definition and Properties

The definition of an ARCH(1) model will be extended for the case that $q > 1$ lags, on which the conditional variance depends.

Definition 13.2 (ARCH(q)) The process (ε_t), $t \in \mathbb{Z}$, is ARCH(q), when $\mathsf{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0$,

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \ldots + \alpha_q \varepsilon_{t-q}^2 \qquad (13.17)$$

with $\omega > 0$, $\alpha_1 \geq 0, \ldots, \alpha_q \geq 0$ and

- $\mathsf{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t^2$ and $Z_t = \varepsilon_t / \sigma_t$ is i.i.d. (strong ARCH)
- $\mathsf{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t^2$ (semi-strong ARCH), or
- $\mathcal{P}(\varepsilon_t^2 \mid 1, \varepsilon_{t-1}, \varepsilon_{t-2}, \ldots, \varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \ldots) = \sigma_t^2$ (weak ARCH)

The conditional variance σ_t^2 in an ARCH(q) model is also a linear function of the q squared lags.

Theorem 13.7 *Let ε_t be a semi-strong ARCH(q) process with $\mathsf{Var}(\varepsilon_t) = \sigma^2 < \infty$. Then*

$$\sigma^2 = \frac{\omega}{1 - \alpha_1 - \ldots - \alpha_q}$$

with $\alpha_1 + \cdots + \alpha_q < 1$.

Proof As in Theorem 13.2.   □

If instead $\alpha_1 + \cdots + \alpha_q \geq 1$, then the unconditional variance does not exist and the process is not covariance-stationary.

Theorem 13.8 (Representation of an ARCH(q) Process) *Let ε_t be a (semi-) strong ARCH(q) process with $\mathsf{E}[\varepsilon_t^4] = c < \infty$. Then*

1. *$\eta_t = \sigma_t^2(Z_t^2 - 1)$ is white noise.*
2. *ε_t^2 is an AR(q) process with $\varepsilon_t^2 = \omega + \sum_{i=1}^q \alpha_i \varepsilon_{t-i}^2 + \eta_t$.*

Proof As in Theorem 13.5.   □

It is problematic with the ARCH(q) model that for some applications a larger order q must be used, since large lags only lose their influence on the volatility slowly. It is suggested as an empirical rule of thumb that a minimum order of $q = 14$ should be used. The disadvantage of a large order is that many parameters have to

be estimated under restrictions. The restrictions can be categorized as conditions for stationarity and the strictly positive parameters. If efficient estimation methods are to be used, for example, the maximum likelihood method, the estimation of large dimensional parameter spaces can be numerically quite complicated to obtain.

One possibility of reducing the number of parameters while including a long history is to assume linearly decreasing weights on the lags, i.e.

$$\sigma_t^2 = \omega + \alpha \sum_{i=1}^{q} w_i \varepsilon_{t-i}^2,$$

with

$$w_i = \frac{2(q + 1 - i)}{q(q + 1)},$$

so that only two parameters need to be estimated. In Sect. 13.1.5 we describe a generalized ARCH model, which on the one hand, has a parsimonious parameterization, and on the other hand a flexible lag structure.

13.1.4   Estimation of an ARCH(q) Model

For the general ARCH(q) model from (13.17) the conditional likelihood is

$$l^b(\theta) = \sum_{t=q+1}^{n} l_t(\theta)$$

$$= -\frac{n-1}{2} \log(2\pi) - 1/2 \sum_{t=2}^{n} \log \sigma_t^2 - 1/2 \sum_{t=q+1}^{n} \frac{\varepsilon_t^q + 1}{\sigma_t^2} \quad (13.18)$$

with the parameter vector $\theta = (\omega, \alpha_1, \ldots, \alpha_q)^\top$. Although one can find the optimum of ARCH(1) models by analysing the graph such as Fig. 13.5, it is complicated and impractical for a high dimensional parameter space. The maximization of (13.18) with respect to θ is a non-linear optimization problem, which can be solved numerically. The *score algorithm* is used empirically not only in ARMA models (see Sect. 12.8) but also in ARCH models. In order to implement this approach the first and second derivatives of the (conditional) likelihood with respect to the parameters need to be formed. For the ARCH(q) model the first derivative is:

$$\frac{\partial l_t^b}{\partial \theta} = \frac{1}{2\sigma_t^2} \frac{\partial \sigma_t^2}{\partial \theta} \left(\frac{\varepsilon_t^2}{\sigma_t^2} - 1 \right) \quad (13.19)$$

with

$$\frac{\partial \sigma_t^2}{\partial \theta} = (1, \varepsilon_{t-1}^2, \dots, \varepsilon_{t-q}^2)^\top.$$

The first order condition is $\sum_{t=q+1}^{n} \partial l_t / \partial \theta = 0$. For the second derivative and the asymptotic properties of the QML estimator see Sect. 13.1.6.

13.1.5   Generalized ARCH (GARCH)

The ARCH(q) model can be generalized by extending it with the autoregressive terms of the volatility.

Definition 13.3 (GARCH(p,q)) The process (ε_t), $t \in \mathbb{Z}$, is GARCH(p,q), if $\mathsf{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0$,

$$\sigma_t^2 = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^2, \tag{13.20}$$

and

- $\mathsf{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t^2$ and $Z_t = \varepsilon_t / \sigma_t$ is i.i.d. (strong GARCH)
- $\mathsf{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma_t^2$ (semi-strong GARCH), or
- $\mathcal{P}(\varepsilon_t^2 \mid 1, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-1}^2, \varepsilon_{t-2}^2, \dots) = \sigma_t^2$ (weak GARCH).

The sufficient but not necessary conditions for

$$\sigma_t^2 > 0 \quad a.s., \qquad (\; \mathsf{P}[\sigma_t^2 > 0] = 1 \;) \tag{13.21}$$

are $\omega > 0$, $\alpha_i \geq 0$, $i = 1, \dots, q$ and $\beta_j \geq 0$, $j = 1, \dots, p$. In the case of the GARCH(1,2) model

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \beta_1 \sigma_{t-1}^2$$

$$= \frac{\omega}{1-\beta} + \alpha_1 \sum_{j=0}^{\infty} \beta_1^j \varepsilon_{t-j-1}^2 + \alpha_2 \sum_{j=0}^{\infty} \beta_1^j \varepsilon_{t-j-2}^2$$

$$= \frac{\omega}{1-\beta} + \alpha_1 \varepsilon_{t-1}^2 + (\alpha_1 \beta_1 + \alpha_2) \sum_{j=0}^{\infty} \beta_1^j \varepsilon_{t-j-2}^2$$

with $0 \leq \beta_1 < 1$. $\omega > 0$, $\alpha_1 \geq 0$ and $\alpha_1 \beta_1 + \alpha_2 \geq 0$ are necessary and sufficient conditions for (13.21) assuming that the sum $\sum_{j=0}^{\infty} \beta_1^j \varepsilon_{t-j-2}^2$ converges.

Theorem 13.9 (Representation of a GARCH(p, q) Process) *Let ε_t be a (semi-)strong GARCH(p, q) process with $\mathsf{E}[\varepsilon_t^4] = c < \infty$. Then*

1. $\eta_t = \sigma_t^2(Z_t^2 - 1)$ *is white noise.*
2. ε_t^2 *is an ARMA(m, p) process with*

$$\varepsilon_t^2 = \omega + \sum_{i=1}^{m} \gamma_i \varepsilon_{t-i}^2 - \sum_{j=1}^{p} \beta_j \eta_{t-j} + \eta_t, \tag{13.22}$$

with $m = \max(p, q)$, $\gamma_i = \alpha_i + \beta_i$. $\alpha_i = 0$ when $i > q$, and $\beta_i = 0$ when $i > p$.

Proof As in Theorem 13.5.   □

If ε_t follows a GARCH process, then from Theorem 13.9 we can see that ε_t^2 follows an ARMA model with conditional heteroscedastic error terms η_t. As we know, if all the roots of the polynomial $(1 - \beta_1 z - \ldots - \beta_p z^p)$ lie outside the unit circle, then the ARMA process (13.22) is invertible and can be written as an AR(∞) process. Moveover it follows from Theorem 13.8 that the GARCH(p, q) model can be represented as an ARCH(∞) model. Thus one can deduce analogous conclusions from the ARMA models in determining the order (p, q) of the model. There are however essential differences in the definition of the persistence of shocks.

Theorem 13.10 (Unconditional Variance of a GARCH(p, q) Process) *Let ε_t be a semi-strong GARCH(p, q) process with $\mathsf{Var}(\varepsilon_t) = \sigma^2 < \infty$. Then*

$$\sigma^2 = \frac{\omega}{1 - \sum_{i=1}^{q} \alpha_i - \sum_{j=1}^{p} \beta_j},$$

with $\sum_{i=1}^{q} \alpha_i + \sum_{j=1}^{p} \beta_j < 1$.

Proof As in Theorem 13.2.   □

General conditions for the existence of higher moments of the GARCH(p, q) models are given in He and Teräsvirta (1999). For the smaller order models and under the assumption of distribution we can derive:

Theorem 13.11 (Fourth Moment of a GARCH(1,1) Process) *Let ε_t be a (semi-)strong GARCH(1,1) process with $\mathsf{Var}(\varepsilon_t) = \sigma^2 < \infty$ and $Z_t \sim N(0, 1)$. Then $\mathsf{E}[\varepsilon_t^4] < \infty$ holds if and only if $3\alpha_1^2 + 2\alpha_1\beta_1 + \beta_1^2 < 1$. The Kurtosis $Kurt(\varepsilon_t)$ is given as*

$$Kurt[\varepsilon_t] = \frac{\mathsf{E}[\varepsilon_t^4]}{\left(\mathsf{E}[\varepsilon_t^2]\right)^2} = 3 + \frac{6\alpha_1^2}{1 - \beta_1^2 - 2\alpha_1\beta_1 - 3\alpha_1^2}. \tag{13.23}$$

Proof It can be proved that $\mathsf{E}[\varepsilon_t^4] = 3\,\mathsf{E}[(\omega + \alpha_1\varepsilon_{t-1}^2 + \beta_1\sigma_{t-1}^2)^2]$ and the stationarity of ε_t.   □

The function (13.23) is illustrated in Fig. 13.6 for all $\alpha_1 > 0$, $Kurt[\varepsilon_t] > 3$, i.e. the distribution of ε_t is leptokurtic. We can observe that the kurtosis equals 3 only
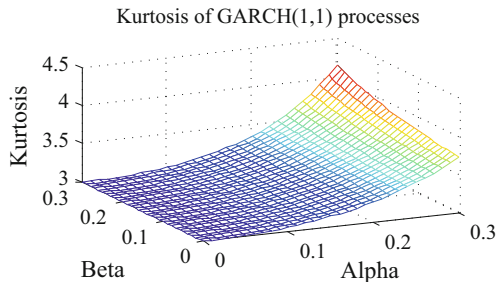
Fig. 13.6   Kurtosis of a GARCH(1,1) process according to (13.23) for different parameters β_1 and α_1 **SFEkurgarch**

in the case of the boundary value $\alpha_1 = 0$ where the conditional heteroscedasticity disappears and a Gaussian white noise takes place. In addition it can be seen in the figure that the kurtosis increases in β_1 slowly for a given α_1. On the contrary it increases in α_1 much faster for a given β_1.

Remark 13.3 Nelson (1990a) shows that the strong GARCH(1,1) process X_t is strictly stationary when $\mathsf{E}[\log(\alpha_1 Z_t^2 + \beta_1)] < 0$. If $Z_t \sim \mathrm{N}(0, 1)$, then the conditions for strict stationarity are weaker than those for covariance-stationarity: $\alpha_1 + \beta_1 < 1$.

In practical applications it is frequently shown that models with smaller order sufficiently describe the data. In most cases GARCH(1,1) is sufficient.

A substantial disadvantage of the standard ARCH and GARCH models exists since they cannot model asymmetries of the volatility with respect to the sign of past shocks. This results from the squared form of the lagged shocks in (13.17) and (13.20). Therefore they have an effect on the level but no effect on the sign. In other words, bad news (identified by a negative sign) has the same influence on the volatility as good news (positive sign) if the absolute values are the same. Empirically it is observed that bad news has a larger effect on the volatility than good news. In Sects. 13.2 and 15.1 we will take a closer look at the extensions of the standard models which can be used to calculate these observations.

13.1.6   Estimation of GARCH(p, q) Models

Based on the ARMA representation of GARCH processes (see Theorem 13.9) Yule–Walker estimators $\tilde{\theta}$ are considered once again. These estimators are, as can be shown, consistent and asymptotically normally distributed, $\sqrt{n}(\tilde{\theta}-\theta) \xrightarrow{\mathcal{L}} \mathrm{N}(0, \tilde{\Sigma})$. However in the case of GARCH models they are not efficient in the sense that the matrix $\tilde{\Sigma} - J^{-1}IJ^{-1}$ is positively definite, where $J^{-1}IJ^{-1}$ is the asymptotic covariance matrix of the QML estimator, see (13.26). In the literature there are several experiments on the efficiency of the Yule–Walker and QML estimators in finite samples, see Sect. 13.6. In most cases maximum likelihood methods are chosen in order to get the efficiency.
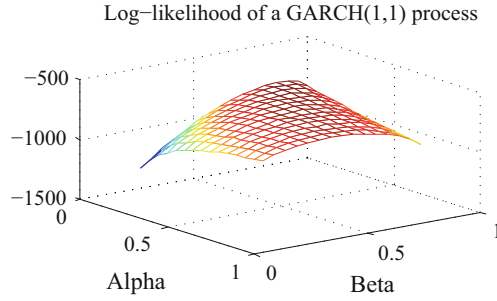
Log−likelihood of a GARCH(1,1) process

Fig. 13.7   Log-likelihood function of a generated GARCH(1,1) process with $n = 500$. The true parameters are $\omega = 0.10$, $\alpha = 0.10$ and $\beta = 0.80$ Q SFElikgarch
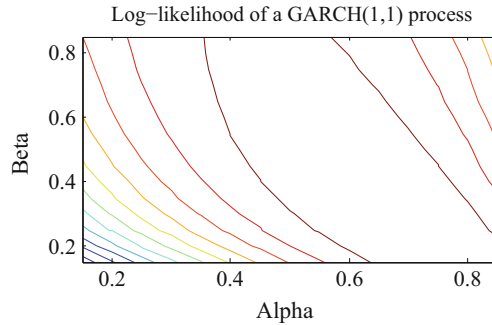

Fig. 13.8   Contour plot of the likelihood function of a generated GARCH(1,1) process with $n = 500$. The true parameters are $\omega = 0.10$, $\alpha = 0.10$ and $\beta = 0.80$ Q SFElikgarch

The likelihood function of the general GARCH(p, q) model (13.20) is identical to (13.18) with the extended parameter vector $\theta = (\omega, \alpha_1, \ldots, \alpha_q, \beta_1, \ldots, \beta_p)^\top$. Figure 13.7 displays the likelihood function of a generated GARCH(1,1) process with $\omega = 0.1$, $\alpha = 0.1$, $\beta = 0.8$ and $n = 500$. The parameter ω was chosen so that the unconditional variance is constant everywhere, i.e. with a variance of σ^2, $\omega = (1 - \alpha - \beta)\sigma^2$. As one can see, the function is flat on the right, close to the optimum, thus the estimation will be relatively imprecise, i.e. it will have a larger variance. In addition, Fig. 13.8 displays the contour plot of the likelihood function.

The first partial derivatives of (13.18) are

$$\frac{\partial l_t}{\partial \theta} = \frac{1}{2\sigma_t^2} \frac{\partial \sigma_t^2}{\partial \theta} \left(\frac{\varepsilon_t^2}{\sigma_t^2} - 1 \right) \tag{13.24}$$

with

$$\frac{\partial \sigma_t^2}{\partial \theta} = \vartheta_t + \sum_{j=1}^{p} \frac{\partial \sigma_{t-j}^2}{\partial \theta}.$$

and $\vartheta_t = (1, \varepsilon_{t-1}^2, \ldots, \varepsilon_{t-q}^2, \sigma_{t-1}^2, \ldots, \sigma_{t-p}^2)^\top$. The first order conditions are $\sum_{t=q+1}^n \partial l_t / \partial \theta = 0$. The matrix of the second derivatives takes the following form:

$$
\frac{\partial^2 l_t(\theta)}{\partial\theta\,\partial\theta^\top} = \frac{1}{2\sigma_t^4} \frac{\partial\sigma_t^2}{\partial\theta} \frac{\partial\sigma_t^2}{\partial\theta^\top} - \frac{1}{2\sigma_t^2} \frac{\partial^2\sigma_t^2(\theta)}{\partial\theta\,\partial\theta^\top}
$$

$$
- \frac{\varepsilon_t^2}{\sigma_t^6} \frac{\partial\sigma_t^2}{\partial\theta} \frac{\partial\sigma_t^2}{\partial\theta^\top} + \frac{\varepsilon_t^2}{2\sigma_t^4} \frac{\partial^2\sigma_t^2(\theta)}{\partial\theta\,\partial\theta^\top}   \tag{13.25}
$$

Under the conditions

1. $\mathsf{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathsf{E}[Z_t^2 \mid \mathcal{F}_{t-1}] = 1$,
2. strict stationarity of ε_t

and under some technical conditions the ML estimator is consistent. If in addition it holds that $\mathsf{E}[Z_t^4 \mid \mathcal{F}_{t-1}] < \infty$, then $\hat\theta$ is asymptotically normally distributed:

$$
\sqrt{n}(\hat\theta - \theta) \xrightarrow{\mathcal{L}} \mathrm{N}_{p+q+1}(0, J^{-1}IJ^{-1})   \tag{13.26}
$$

with

$$
I = \mathsf{E}\left(\frac{\partial l_t(\theta)}{\partial\theta} \frac{\partial l_t(\theta)}{\partial\theta^\top} \right)
$$

and

$$
J = -\mathsf{E}\left(\frac{\partial^2 l_t(\theta)}{\partial\theta\,\partial\theta^T} \right).
$$

Theorem 13.12 (Equivalence of I and J) *If $Z_t \sim N(0,1)$, then it holds that $I = J$.*

Proof Building the expectations of (13.25) one obtains

$$
\mathsf{E}\left[\frac{\partial^2 l_t(\theta)}{\partial\theta\,\partial\theta^\top} \right] = -\mathsf{E}\,\frac{1}{2\sigma_t^4} \frac{\partial\sigma_t^2}{\partial\theta} \frac{\partial\sigma_t^2}{\partial\theta^\top}.
$$

For I we have

$$
\mathsf{E}\left[\frac{\partial l_t(\theta)}{\partial\theta} \frac{\partial l_t(\theta)}{\partial\theta^T} \right] = \mathsf{E}\left[\frac{1}{4\sigma_t^4} \frac{\partial\sigma_t^2}{\partial\theta} \frac{\partial\sigma_t^2}{\partial\theta^\top} (Z_t^4 - 2Z_t^2 + 1) \right]   \tag{13.27}
$$

$$
= \mathsf{E}\left[\frac{1}{4\sigma_t^4} \frac{\partial\sigma_t^2}{\partial\theta} \frac{\partial\sigma_t^2}{\partial\theta^\top} \right] \{\mathrm{Kurt}(Z_t \mid \mathcal{F}_{t-1}) - 1\}
$$

From the assumption $Z_t \sim N(0,1)$ it follows that $\mathrm{Kurt}(Z_t \mid \mathcal{F}_{t-1}) = 3$ and thus the claim. $\square$

If the distribution of Z_t is specified correctly, then $I = J$ and the asymptotic variance can be simplified to J^{-1}, i.e. the inverse of the Fisher Information matrix. If this is not the case and it is instead leptokurtic, for example, the maximum of (13.9) is still consistent but no longer efficient. In this case the ML method is interpreted as the QML method.

Consistent estimators for the matrices I and J can be obtained by replacing the expectation with the simple average.

13.2 Extensions of the GARCH Model

Standard GARCH models assume that positive and negative error terms have a symmetric effect on volatility. In other words, good and bad news have the same effect on the volatility in this model. In practice this assumption is frequently violated, in particular by stock returns, in that the volatility increases more after bad news than after good news. This so called *Leverage Effect* appears firstly in Black (1976), who noted that:

> a drop in the value of the firm will cause a negative return on its stock, and will usually increase the leverage of the stock. [...] That rise in the debt-equity ratio will surely mean a rise in the volatility of the stock.

A very simple but plausible explanation for the leverage effect: Negative returns imply a larger proportion of debt through a reduced market value of the firm, which leads to a higher volatility. The risk, i.e. the volatility reacts first to larger changes of the market value, nevertheless it is empirically shown that there is a high volatility after smaller changes. On the other hand, Black said nothing about the effect of positive returns on the volatility. Although the positive returns cause smaller increases, they do cause an increase in the volatility. From an empirical point of view the volatility reacts asymmetrically to the sign of the shocks and therefore a number of parameterized extensions of the standard GARCH model have been suggested recently. Below we will discuss two of the most important ones: the exponential GARCH (EGARCH) and the threshold GARCH (TGARCH) model.

13.2.1 Exponential GARCH

Let Z_t further denote a series of $i.i.d.$ standardized random variables with expectation 0 and variance 1. The general EGARCH model is given by Nelson (1991):

$$\log \sigma_t^2 = \omega_t + \sum_{k=1}^{\infty} \beta_k g(Z_{t-k}), \qquad (13.28)$$

where ω_t, β_k are deterministic coefficients and

$$g(Z_t) = \theta Z_t + \gamma(|Z_t| - E|Z_t|). \tag{13.29}$$

It can be directly seen that $\mathsf{E}\left[g(Z_t)\right] = 0$.

The EGARCH model in (13.28) shows some differences from the standard GARCH model:

- Volatility of the EGARCH model, which is measured by the conditional variance σ_t^2, is an explicit multiplicative function of lagged innovations. On the contrary, volatility of the standard GARCH model is an additive function of the lagged error terms ε_t, which causes a complicated functional dependency on the innovations; for example the ARCH(1) in Theorem 13.4.
- Volatility can react asymmetrically to the good and the bad news.
- For the general distributions of Z_t the parameter restrictions for strong and covariance-stationarity coincide.
- The parameters in (13.28) and (13.29) are not restricted to positive values.

The function $g(\cdot)$ in (13.29) is piecewise linear. It contains two parameters which define the "size effect" and the "sign effect" of the shocks on volatility. The first is a typical ARCH effect while the second is an asymmetrical effect, for example, the leverage effect. The term $\gamma(|Z_t| - E|Z_t|)$ determines the size effect and the term θZ_t determines the sign effect. The parameter γ is thus typically positive and θ is negative.

To estimate EGARCH models instead of the general MA(∞) representation (13.28), an ARMA(p, q) model is applied, i.e.

$$\Delta(L) \log \sigma_t^2 = \omega + \Psi(L)g(Z_t), \tag{13.30}$$

with lag-polynomial $\Delta(L)$ and $\Psi(L)$ of order p and q, respectively.

EGARCH models benefit from no parameter restrictions, thus the possible instabilities of optimization routines are reduced. On the other hand the theoretical properties of QML estimators of EGARCH models are not clarified to a great extent.

Let $\omega_t = \omega = 0$ and $\sum_{k=1}^{\infty} \beta_k^2 < \infty$. Then σ_t^2 is strictly stationary and ergodic, see Theorem 2.1 in Nelson (1991). Furthermore, under these conditions the unconditional variance exists when Z_t has a *generalized error distribution* (GED) with parameter $\zeta > 1$, which determines the thickness of the tails, see Theorem 2.2 in Nelson (1991). The GED is leptokurtic when $\zeta < 2$.

The normal distribution is a special case of the GED ($\zeta = 2$). In addition, Nelson gives complicated formulas for the unconditional moments. One problem is that under other leptokurtic distributions such as the Student-t, the unconditional variance does not exist. The reason is that exponential growth of the conditional variance changes with the level of the shocks, which leads to the explosion of the

unconditional variance when the probability for extreme shocks is sufficiently large. Therefore the existence of the unconditional moments depends on the choice of the distribution of the innovations, which is an undesirable property of the EGARCH models. In empirical studies it has been found that EGARCH often overweighs the effects of larger shocks on volatility and thus results in poorer fits than standard GARCH models, see the empirical studies of Engle and Ng (1993).

13.2.2   Threshold ARCH Models

The idea of the Threshold ARCH (TARCH) models is to divide the distribution of the innovations into disjoint intervals and then approximate a piecewise linear function for the conditional standard deviation, see Zakoian (1991), and the conditional variance respectively, see Glosten, Jagannathan and Runkle (1993). If there are only two intervals, the division is normally at zero, i.e. the influence of positive and negative innovations on the volatility is differentiated. In this case the TARCH model of order q can be written as:

$$\sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^\delta + \sum_{i=1}^{q} \alpha_i^- \varepsilon_{t-i}^\delta \mathbf{1}(\varepsilon_{t-i} < 0), \tag{13.31}$$

with the indicator function $\mathbf{1}(\cdot)$ and $\delta = 1$ as in Zakoian (1991) or $\delta = 2$ as in Glosten et al. (1993).

Rabemananjara and Zakoian (1993) extend this model by including the lagged conditional standard deviations (variance respectively) as a regressor, which is known as the TGARCH model. They also give conditions for covariance-stationarity in their study.

Instead of a piecewise linear function Gouriéroux and Monfort (1992) use a stepwise function (piecewise constant) as a model for the volatility. Let $A_j, j = 1, \ldots, J$ be a partition of the distribution of the innovation. Then a *qualitative threshold ARCH* model (QTARCH) of order 1 is given by:

$$y_t = \sum_{j=1}^{J} m_j \mathbf{1}(y_{t-1} \in A_j) + \sum_{j=1}^{J} s_j \mathbf{1}(y_{t-1} \in A_j) Z_t, \tag{13.32}$$

where m_j and s_j are scalars. In (13.32) conditional expectations and conditional standard deviations are modelled as stepwise functions. One notices that (13.32) is a homogenous Markov Chain of order one. Models of higher order can be easily derived. Gouriéroux and Monfort (1992) give a detailed discussion of the statistical properties of (13.32). Moreover the consistency and asymptotic normal distribution of the QML estimators are also discussed by them.

The threshold models are identified by an abrupt transition between two regimes when the generating innovation crosses a threshold value. If a smooth transition is preferred in the model, then the indicator function in (13.31) can be replaced with the desired continuous function, which tends to zero if the values are close to ε_{t-i} and tends to one for the values further away. Frequently, the logistic function is chosen. The model is thus:

$$\sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^\delta + \sum_{i=1}^{q} \alpha_i^- \varepsilon_{t-i}^\delta F_\gamma(\varepsilon_{t-i}) \tag{13.33}$$

$$F_\gamma(u) = \frac{1}{1 + \exp(-\gamma u)}, \tag{13.34}$$

with the parameter $\gamma > 0$. The indicator function is a limiting case of the logistic function for $\gamma \longrightarrow \infty$.

Finally, another model class needs to be mentioned, which is very general and can replicate the asymmetries: the Asymmetric Power ARCH (APARCH) model from Ding, Granger and Engle (1993),

$$\sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i (|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i})^\delta + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^\delta, \tag{13.35}$$

where $\delta > 0$ is a parameter to be estimated. However, note that the EGARCH model is not included in this model class, a direct test between GARCH and EGARCH models is thus impossible. A very general ARCH model, the *augmented GARCH* model from Duan (1997), also includes the EGARCH model.

13.2.3   Risk and Returns

In finance theory the relationship between risk and returns plays an important role. Many theoretical models such as CAPM imply a linear relationship between the expected returns of a market portfolio and the variance. If the risk (i.e. the variance) is not constant over time, then the conditional expectation of the market returns is a linear function of the conditional variance. The idea from Engle, Lilien and Robins (1987) was consequently used to estimate the conditional variances in GARCH and then the estimations used in the conditional expectations' estimation. This is the so-called *ARCH in Mean* (ARCH-M) model.

Let y_t be a covariance-stationary return process of a broad market index and σ_t^2 be the conditional variance specified in an ARCH(1) model. The ARCH-M model is

$$y_t = \nu + \lambda g(\sigma_t^2) + \varepsilon_t \tag{13.36}$$
$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2,$$

where $g(\cdot)$ is a known parameter function. The CAPM implies that $g(x) = x$, although in most empirical applications the square root or the logarithm is used. The expression $\lambda g(\sigma_t^2)$ can be interpreted as the risk premium.

13.2.4   Estimation Results for DAX and FTSE 100 Returns

We apply various GARCH models discussed previously to the returns of the DAX and FTSE 100 index. We consider daily returns from 1998 to 2007 (2,607 observations). The returns indicate a clear autocorrelation of first order. There are various possibilities to model these autocorrelations. The two most important models, which we have discussed, are the AR models and the ARCH-M models. The latter can be interpreted economically, i.e. a time dependent risk premium implying an autocorrelation of the returns. This economic motivation is missing for the AR models: the AR term cannot be interpreted as the risk premium, since it can be negative, which contradicts the usual assumption of a risk averse agent. However, the AR models frequently offer a better fit to the data than the ARCH-M model. Therefore, we consider the following model:

$$y_t = \mu_t + \sigma_t Z_t$$

with $\mu_t = \nu + \phi y_{t-1}$, an AR(1) model, and with σ_t specified as a GARCH, TGARCH or EGARCH model.

The estimation results are given in Table 13.4. They can be summarized as follows:

- The parameter ν is significantly positive for almost all AR models, which is reflected in the long-run increasing trend of the stock price. For FTSE 100 returns, the estimated parameter ν when using the TGARCH or EGARCH is not significant.
- The ARCH effects are very pronounced, i.e. the parameter α in the GARCH model is significant.
- There is a high persistence of shocks in the volatility. This persistence is measured in the GARCH case by the sum of α and β and is in each case close to 1.
- A leverage effect exists: the corresponding parameters in the TGARCH and the EGARCH models have the appropriate signs. In the TGARCH case the t-statistic for α^- is 8.89 and 9.95, respectively, and in the EGARCH case the t-statistic for θ is -10.74 and -12.48, respectively. Negative shocks increase the volatility more than positive shocks on both markets.
- The TGARCH and the EGARCH models have a priority for asymmetry, since they have a better fit to the data when the same number of parameters is considered.

Table 13.4 Estimation results of various GARCH models, applied to DAX and FTSE 100 returns from 1998 to 2007

	DAX			FTSE 100		
	GARCH	TGARCH	EGARCH	GARCH	TGARCH	EGARCH
$\hat{\nu}$	7.57E−04	4.25E−04	4.30E−04	3.63E−04	5.90E−05	−1.66E−05
	(3.338)	(1.902)	(1.955)	(2.217)	(0.347)	(−0.101)
$\hat{\phi}$	−0.021	−0.017	−0.019	−0.027	−0.022	−0.017
	(−0.938)	(−0.790)	(−0.876)	(−1.263)	(−1.076)	(−0.856)
$\hat{\omega}$	1.73E−06	2.05E−06	−0.139	1.26E−06	1.19E−06	−0.126
	(4.565)	(5.837)	(−6.768)	(3.916)	(5.462)	(−6.321)
$\hat{\alpha}$	0.080	0.022		0.097	0.006	
	(10.048)	(3.299)		(10.051)	(0.633)	
$\hat{\alpha}^-$		0.093			0.112	
		(8.888)			(9.946)	
$\hat{\beta}$	0.912	0.919	0.984	0.895	0.926	0.986
	(112.2)	(118.6)	(417.8)	(88.1)	(99.2)	(451.3)
$\hat{\gamma}$			0.135			0.097
			(10.089)			(6.777)
$\hat{\theta}$			−0.071			−0.100
			(−10.737)			(−12.484)
logL	7624.90	7649.24	7643.97	8387.09	8414.19	8422.94

Parenthesis show the t-statistic based on the QML asymptotic standard error   Q SFEgarchest

13.3   Shortfalls of GARCH

The purpose of this section is to examine some shortfalls uncovered in previous academic literature, primarily the inability of the GARCH model to serve as a true data generating process as well as an accurate forecasting tool. Furthermore, this section provides an explanation of these shortfalls which include the overestimation of the unconditional variance and the existence of the Integrated GARCH (IGARCH) effect. Section 13.3.1 discusses analysis of returns of the S&P 500 stock market index from Starica (2003). Section 13.3.2 presents original analysis of DAX and FTSE 100 returns for comparison.

13.3.1   Recent Challenges to GARCH Models

One of the main sources criticizing the ability of GARCH models to accurately model financial time series is the paper by Starica (2003) which discusses how well the endogenous dynamics created by the GARCH(1,1) model replicate the actual dynamics of main financial indices. Specifically, Starica tests the performance of the GARCH model while describing and forecasting the dynamics of returns of the S&P 500 stock market index.

Modelling S&P 500 Data

Fitting a conditionally normal GARCH(1,1) process $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$ to the S&P 500 daily logarithmic returns (first differences of the logarithms of daily closing prices) from 16 May 1995 to 29 April 2003, a range containing exactly 2,000 observations, yields the following equation with estimated parameters:

$$\sigma_t^2 = 1.4264 \times 10^{-6} + 0.0897 \varepsilon_{t-1}^2 + 0.9061 \sigma_{t-1}^2 \qquad (13.37)$$

Since $\hat{\alpha} + \hat{\beta} = 0.9958$, a value very close to 1, the integrated GARCH (IGARCH) effect is present.

IGARCH Effect

When estimating the parameters in GARCH models, it is observed that for shorter samples, the estimated parameters α and β sum up to values significantly different from 1 while for longer samples, their sum approaches 1. The combination of these phenomenon is called the integrated GARCH (IGARCH) effect of return data.

   Stationarity of the data is one of the basic assumptions of GARCH models. The IGARCH effect occurs due to non-stationarity of the data as a result of structural breaks, i.e. changes in the unconditional mean or variance. For a GARCH(1,1) process, when $\alpha + \beta$ approaches 1, this signals a presence of an IGARCH effect. The larger the absolute difference in the variance (the greater the degree of non-stationarity) of sub-samples having differing unconditional variances, the closer $\alpha + \beta$ gets to 1.

   Structural breaks in unconditional variance resulting in IGARCH effects pose problems when using GARCH models to estimate volatilities. IGARCH is likely to cause an explosion of the estimated unconditional variance of the GARCH processes.

Next-Day Volatility Forecasting

In practice, one of main applications of the GARCH model is to provide next-day volatility forecasts. When estimated innovations, or residuals, defined as

$$\hat{Z}_t = \varepsilon_t / \hat{\sigma}_t, \quad \hat{\sigma}_t^2 = \hat{\omega} + \hat{\alpha} \varepsilon_{t-1}^2 + \hat{\beta} \hat{\sigma}_{t-1}^2, \quad t = 1, \dots, n \qquad (13.38)$$

are almost independent, this is taken as evidence of accurate next-day volatility forecasts.

   GARCH(1,1) volatility can be found using previously estimated parameters applied to the GARCH(1,1) model. However, for the S&P 500 data, the following infinite moving-average approximation also provides an accurate estimate.

$$\hat{\sigma}_t^2 = \hat{\omega} + \hat{\alpha}\varepsilon_{t-1}^2 + \hat{\beta}\hat{\sigma}_{t-1}^2$$

$$= \hat{\omega} + \hat{\alpha}\varepsilon_{t-1}^2 + \hat{\beta}(\hat{\omega} + \hat{\alpha}\varepsilon_{t-2}^2 + \hat{\beta}\hat{\sigma}_{t-2}^2)$$

$$= \hat{\omega}(1 + \hat{\beta}) + \hat{\alpha}(\varepsilon_{t-1}^2 + \hat{\beta}\varepsilon_{t-2}^2) + \hat{\beta}^2\hat{\sigma}_{t-2}^2$$

$$\approx \hat{\omega}(1 + \hat{\beta}) + (1 - \hat{\beta})(\varepsilon_{t-1}^2 + \hat{\beta}\varepsilon_{t-2}^2) + \hat{\beta}^2\hat{\sigma}_{t-2}^2 \qquad (13.39)$$

$$\approx \dots$$

$$\approx \hat{\omega}(1 + \hat{\beta} + \hat{\beta}^2 + \dots) + (1 - \hat{\beta})(\varepsilon_{t-1}^2 + \hat{\beta}\varepsilon_{t-2}^2 + \hat{\beta}^2\varepsilon_{t-3}^2 \dots)$$

$$\approx \frac{\hat{\omega}}{1 - \hat{\beta}} + \frac{\varepsilon_{t-1}^2 + \hat{\beta}\varepsilon_{t-2}^2 + \hat{\beta}^2\varepsilon_{t-3}^2 + \dots}{1 + \hat{\beta} + \hat{\beta}^2 + \dots} \qquad (13.40)$$

$$\approx \frac{\varepsilon_{t-1}^2 + \hat{\beta}\varepsilon_{t-2}^2 + \hat{\beta}^2\varepsilon_{t-3}^2 + \dots}{1 + \hat{\beta} + \hat{\beta}^2 + \dots} \qquad (13.41)$$

where (13.39) substitutes $1 - \hat{\beta}$ for $\hat{\alpha}$ since $\hat{\alpha} + \hat{\beta} \approx 0.995$, (13.40) uses the fact that $1 + \hat{\beta} + \hat{\beta}^2 + \dots = 1/(1 - \hat{\beta})$ while (13.41) neglects the constant $\hat{\omega}/(1 - \hat{\beta})$, since $\hat{\omega}$ is very small.

The approximation (13.41) is in fact an exponential smoother, the exponentially weighted moving average (EWMA) applied to ε_t^2. Forecasting with EWMA is optimal in the mean-square error sense for the state space model

$$x_t = \mu_t + \eta_t \qquad (13.42)$$

$$\mu_t = \mu_{t-1} + \nu_t \qquad (13.43)$$

where

- η_t and ν_t are i.i.d. sequences
- $\mathsf{E}[\eta] = 0$ and $\mathsf{E}[\nu] = 0$.

Assuming

$$r_t = \sigma_t Z_t, \qquad \mathsf{E}[Z_t] = 0, \qquad \mathsf{E}[Z_t^2] = 1, \qquad (13.44)$$

the GARCH(1,1) model can be written in the form of (13.42):

$$r_t^2 = \sigma_t^2 + \sigma_t^2(Z_t^2 - 1) = \sigma_t^2 + \eta_t \qquad \text{with} \qquad \mathsf{E}[\eta_t] = 0, \qquad (13.45)$$

where σ_t is a deterministic, smooth function of time. Here, we assume the unconditional variance is time-varying and that the returns are independent but not i.i.d. Equation (13.43) incorporates the uncertainty about the form of the model.

Similarly, the non-parametric regression model shares a closely-related setup, sharing the state model representation and ability to incorporate uncertainty about

the form of the model. In this case, the uncertainty is handled by modelling the signal μ as a deterministic function of time.

From non-parametric regression literature, the time-varying unconditional variance is estimated as:

$$\hat{\sigma}_{t;h}^2 = \sum_{k=1}^{n} W_k(t;h) r_k^2 \tag{13.46}$$

where h is the bandwidth, K a kernel and

$$W_k(t;h) = K\left(\frac{t-k}{h}\right) \Big/ \sum_{k=1}^{n} K\left(\frac{t-k}{h}\right). \tag{13.47}$$

Together, these two equations are from Nadaraya–Watson (zero-degree local polynomial) kernel estimate of σ_t^2.

Empirical tests of the S&P 500 data from 16 May 1995 to 29 April 2003 find that the paths of the GARCH(1,1) estimate $\hat{\sigma}_t^2$ and the non-parametric estimate (13.46) closely follow each other. Furthermore, there were no statistically significant differences between GARCH(1,1) and a non-stationary, non-parametric regression approach to next-day volatility forecasting.

However, one possible critique of the non-parametric approach is that it lacks any dynamics. It is assumed that volatility is driven by unstructured, exogenous factors. Since not much is known about these factors, the current volatility level is forecast to be the return volatility for the near future. Near future returns are modelled as i.i.d. with a variance equal to today's estimate.

In contrast, the GARCH(1,1) model contains an endogenous mechanism for the volatility process which is capable of foreseeing future developments in the movements of prices. Therefore, it can be argued that the predictability power of the GARCH(1,1) model makes it a preferable choice.

Modelling Performance

As previously mentioned, the GARCH(1,1) process is assumed to be a realistic data generating process for financial returns. According to this assumption, given a sample of data with parameters estimated on a reasonable long sub-sample, these parameters should provide a good model for the entire sample. However, when using the estimated parameters from the S&P 500 returns from 16 May 1995 to 29 April 2003 to describe the data of the entire 11,727 observation sample from March 1957 to October 2003, the GARCH(1,1) model's simulated samples differ vastly from the real data. In fact, the GARCH(1,1) model greatly overestimates the variance for the long series of S&P 500 returns. In 25,000 simulated samples, GARCH(1,1) produced an interval of variances [0.00014, 0.01526] while the true variance of the returns is 0.00008, see Starica (2003).

The inaccurate results of these simulations require a closer look at the estimated unconditional variance. Under the assumption of weak stationarity ($\alpha + \beta < 1$), the unconditional variance of the GARCH(1,1) model is

$$\sigma^2 = \omega/(1 - \alpha - \beta). \tag{13.48}$$

Using this formula for the sub-sample of S&P 500 returns from 16 May 1995 to 29 April 2003, the standard deviation of the GARCH(1,1) estimated unconditional variance ranges from 1.5 to 5 times larger than the sample's true standard deviation. Hence, the GARCH(1,1) model's inability to produce accurate unconditional point estimates has implications in its data forecasting performance.

Forecasting Performance: S&P 500 Short Sample

Under the assumption that data follows a GARCH(1,1) process with $\alpha + \beta < 1$, the p-step ahead variance forecast is given by:

$$\sigma_{t+p}^{2,GARCH} = \mathsf{E}[r_{t+p}^2] = \sigma^2 + (\alpha + \beta)^p(\sigma_t^2 - \sigma^2), \tag{13.49}$$

where σ^2 is the unconditional variance defined in (13.48). Since $\alpha + \beta < 1$, for large values of p, the second term on the right-hand side of the equation goes to zero so that the forecast $\sigma_{t+p}^{2,GARCH}$ is equal to the unconditional variance σ^2. The minimum Mean Square Error (MSE) forecast for the variance of the next p aggregated returns is:

$$\overline{\sigma}_{t,p}^{2,GARCH} = \mathsf{E}[r_{t+1} + \ldots + r_{t+p}]^2 = \sigma_{t+1}^{2,GARCH} + \ldots + \sigma_{t+p}^{2,GARCH}. \tag{13.50}$$

As a benchmark (BM) for volatility forecasting, a simple non-stationary model (13.44) is used. As previously discussed, since this model does not specify any dynamics for the variance, future observations are modelled as i.i.d. with constant variance $\hat{\sigma}_t^2$, which is an estimate of σ_t^2. Here, the sample variance of the previous year of returns will be used to estimate σ_t^2:

$$\hat{\sigma}_t^2 = \frac{1}{250} \sum_{i=1}^{250} r_{t-i+1}^2. \tag{13.51}$$

Therefore, the forecast for $\mathsf{E}[r_{t+p}^2]$ is

$$\sigma_{t+p}^{2,BM} = \hat{\sigma}_t^2. \tag{13.52}$$

The forecast for the variance of the next p aggregated returns, $\mathsf{E}[r_{t+1} + \ldots + r_{t+p}]^2$, is

$$\overline{\sigma}_{t,p}^{2,BM} = p\,\hat{\sigma}_t^2 = \frac{p}{250}\sum_{i=1}^{250} r_{t-i+1}^2. \tag{13.53}$$

In order to evaluate forecasting performance, realized volatility $\overline{r}_{t,p}$ in the interval $[t+1, t+p]$ is defined by

$$\overline{r}_{t,p}^2 = \sum_{i=1}^{p} r_{t+i}^2. \tag{13.54}$$

The following formula is preferred for calculating an accurate measure of the MSE

$$\sum_{t=1}^{n}(r_{t+p}^2 - \sigma_{t+p}^{2,*})^2 \tag{13.55}$$

where * represents BM or GARCH.

By initially estimating the GARCH(1,1) model on the first number of data points and then continuously re-estimating using a specified interval of time, simultaneously estimating σ_t^2 using (13.51), (13.50) and (13.53) can be used to forecast volatility for the year to come ($p = 1, \ldots, 250$). When forecasting the GARCH(1,1) volatility of S&P returns between 16 May 1999 and 29 April 2003, the effects of the over-estimation of conditional variance shown above are evident as GARCH(1,1) provides poor longer horizon forecasts for this period. In fact, the simple non-parametric, non-stationary model produced better forecasting results. Moreover, MSE analysis shows an increasing $\text{MSE}^{GARCH}/\text{MSE}^{BM}$ ratio over the forecast horizon. These results lead one to wonder if GARCH(1,1) always over-estimates unconditional variance or if the S&P 500 sample used here is unique in some way. The following sections investigate this issue as the GARCH(1,1) model is applied to a longer sample of S&P 500 data.

Forecasting Performance: S&P 500 Long Sample

At this point, it is important to examine the options that one has when using the stationary GARCH(1,1) model. These choices depend on the working assumptions made about the data to be modelled. If the data is assumed to be stationary, a GARCH(1,1) process would be used as a true data generating process for the entire data set. In this case, it is best to use as much data as possible in order to draw any statistical conclusions or to make any forecasts.

However, if the data is assumed to be non-stationary, then the GARCH (1,1) process might be used as a local approximation for a true data generating process. Given this condition, the parameters of the model should be periodically re-estimated on a moving window. Only data from the recent past should be used to calibrate the model to be used for forecasting. Making any statistical statements

based on long samples should be avoided. In the situation of non-stationary data, it is unclear if the GARCH(1,1) model can produce accurate long-horizon forecasts.

Hypothesis: Returns are Stationary

Under the working assumption of stationary data, GARCH(1,1) estimated parameters should be unchanging if they are estimated on an increasing sample size or on a window that moves through the data. When testing a hypothesis of stationarity, if significant changes in the estimated parameters outside of the confidence interval are detected, then the hypothesis should be rejected and the model should not be considered to be a true data generating process.

For this hypothesis testing, Starica uses a much longer sample of S&P data from 4 March 1957 to 9 October 2003, re-estimating the data every 50 observations, both on a moving window of 2000 past observations and on an increasing sample containing all past observations. However, the hypothesis of stationarity was strongly rejected as the estimated parameters changed significantly over time and did a poor job of remaining in their confidence intervals. A similar analysis was performed on returns from the Dow Jones Industrial Index for the same period. The hypothesis was also rejected for the Dow Jones series. From this evidence, it can be concluded that the GARCH(1,1) process is not a data generating process for the returns of the S&P 500 index between 4 March 1957 and 9 October 2003.

The significant changes in the value of the estimated parameters suggests that the data is not stationary, particularly that the unconditional variance might be time-varying. This issue will be examined as a hypothesis of non-stationarity next.

Hypothesis: Returns are Non-stationary

The assumption that long range S&P 500 data may be non-stationary due to a time-varying unconditional variance might explain the poor point estimates described above. Under closer examination of the sum $\alpha + \beta$, estimated from a moving window of 2000 S&P 500 returns from 1957 to 2003, IGARCH effects are present in the period from 1997 to 2003 as $\alpha + \beta < 1$ for this period. An IGARCH effect is likely to cause an explosion of estimated unconditional variance when fitting GARCH(1,1) processes to samples that end in this time period, see Eq. (13.50).

In fact, a comparison of the GARCH(1,1) and sample standard deviations reveals a good agreement between the two estimates for the entire period except from 1997 to 2003 when the IGARCH effect becomes strongly statistically significant. Analysis of the Dow Jones Industrial Index shows similar findings, see Starica (2003).

The strong IGARCH effect in the late 1990s can be attributed to a sharp change in unconditional variance. In the paper by Mikosch and Starica (2004), it is proved, theoretically and empirically, that sharp changes in unconditional variance can cause the IGARCH effect. Similarly, non-parametric modelling of the unconditional

standard deviation of both the S&P 500 and Dow Jones Industrial Index also reveal that the level of volatility triples from 5 % in 1993–1994 to 15 % in 2000–2003.

Implications on Long Horizon Volatility Forecasting

As discussed above, the coefficients of a GARCH(1,1) model can change significantly over time. Moreover, these parameters vary greatly when looking at the S&P 500 and Dow Jones Industrial Index, both of which are major financial indices.

In order to take a closer look at this issue, one can look at a comparison of the forecasting performances of GARCH(1,1) versus a simple approach based on historical volatility, using a similar approach as described above, see Starica (2003). From analysis of MSE^{GARCH} and MSE^{BM}, it is evident that for shorter horizons (60 days ahead for the locally estimated model and 40 days ahead for the model using all past observations) $MSE^{GARCH} < MSE^{BM}$, however beyond this short forward-looking horizon, $MSE^{GARCH} > MSE^{BM}$, meaning that for longer horizons, the approach based on historical volatility performed better. The GARCH(1,1) approach becomes significantly more volatile during the late 1990s, the period in which the IGARCH effect is present, than does the historical volatility method. GARCH(1,1)'s poor performance is also seen for the Dow Jones series of returns, leading one to question its applicability in long-term volatility forecasting.

13.3.2 Volatility Forecasting for DAX and FTSE 100 Returns

Based on Starica's example, this section compares GARCH(1,1) next-day volatility forecasting with a non-parametric estimation very similar to the method described in Sect. 13.3.1 for DAX and FTSE daily index returns in the period from 1 January 1998 to 31 December 2007 (2,607 observations). The time series of both processes has already been shown in Fig. 13.3.

After fitting a WN-GARCH(1,1) process to the data, the following equations with estimated parameters were obtained,

$$\sigma_t^2 = 1.7217 \times 10^{-6} + 0.0803\varepsilon_{t-1}^2 + 0.9125\sigma_{t-1}^2 \tag{13.56}$$

$$\sigma_t^2 = 1.2518 \times 10^{-6} + 0.0968\varepsilon_{t-1}^2 + 0.8951\sigma_{t-1}^2, \tag{13.57}$$

for the DAX and FTSE daily index returns, respectively.

This indicates the presence of an IGARCH effect which can most likely be attributed to this DAX and FTSE 100 index return data being non-stationary with a changing and overestimated unconditional variance. However, to confirm this assumption, further analysis of the unconditional variance is necessary.

Figure 13.9 shows the estimated volatility resulting from using the non-parametric model based on historical volatility. This non-parametric estimate was
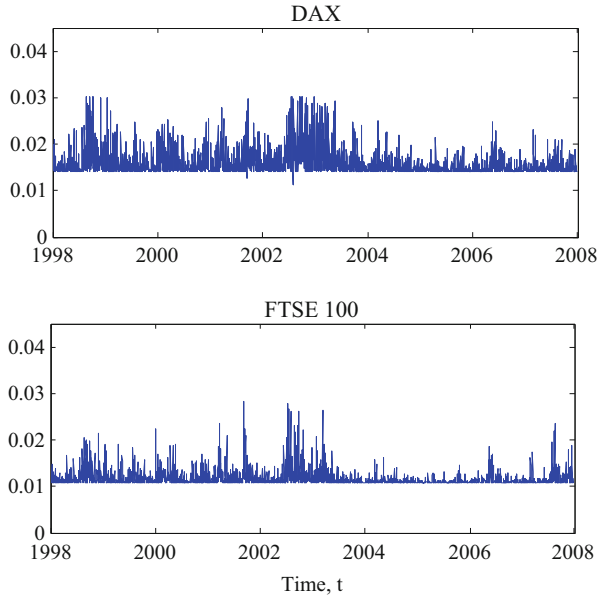
Fig. 13.9 Estimated next-day volatility of DAX and FTSE 100 returns from 1998 to 2007 using a non-parametric model based on historical volatility   **Q** SFEvolnonparest

calculated using a Nadaraya–Watson kernel estimate. The optimal bandwidth was calculated by choosing the value of h that minimises the   final prediction error, see Härdle, Müller, Sperlich and Werwatz (2004). It is evident when comparing the non-parametric estimate in Fig. 13.9 to the GARCH(1,1) estimate in Fig. 13.10, that the paths of these volatility processes are quite similar.

From this analysis of DAX and FTSE 100 index returns, it is observed that the period from 1998 to 2004 was a period with high volatility. This can be seen in the plot of the returns, and it is captured by both the GARCH(1,1) and non-parametric next-day volatility estimates. Moreover, these results are very similar to Starica's study of the S&P 500 returns, demonstrating the correlation between the returns of the securities represented by both indices.

13.4 Multivariate GARCH Models

The generalization of univariate GARCH models to the multivariate case is straight-forward. For the error term ε_t of a d-dimensional time series model we assume that the conditional mean is zero and the conditional covariance matrix is given by the positive definite $(d \times d)$ matrix H_t, i.e.

$$\varepsilon_t = H_t^{1/2}\xi_t \tag{13.58}$$
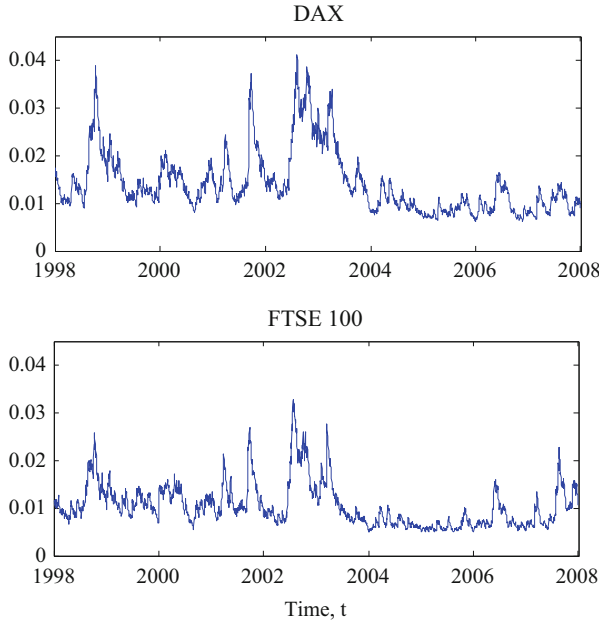
Fig. 13.10 Estimated next-day volatility for DAX and FTSE 100 returns from 1998 to 2007 using a GARCH(1,1) model   ◼ SFEvolgarchest

with i.i.d. innovation vector ξ_t, whose mean is zero and covariance matrix equals the identity matrix I_d. As in the univariate case, H_t depends on lagged error terms ε_{t-i}, $i = 1, \ldots, q$, and on lagged conditional covariance matrices H_{t-i}, $i = 1, \ldots, p$. As we will see shortly, the general case with arbitrary dependencies can lead to very complex structures that may be too difficult to deal with in practice. The reduction of the dimension of the parameter space is therefore often tried. Below, we first discuss a general specification and then a popular restriction, the BEKK model. We will also briefly sketch a model that assumes constant conditional correlations.

13.4.1 The Vec Specification

Let vech($\cdot$) denote the operator that stacks the lower triangular part of a symmetric $d \times d$ matrix into a $d^* = d(d + 1)/2$ dimensional vector. Furthermore we will use the notation $h_t = \text{vech}(H_t)$ and $\eta_t = \text{vech}(\varepsilon_t \varepsilon_t^\top)$. The *Vec specification* of a multivariate GARCH(p, q) model is then given by

$$h_t = \omega + \sum_{i=1}^{q} A_i \eta_{t-i} + \sum_{j=1}^{p} B_j h_{t-j}, \tag{13.59}$$

where A_i and B_j are parameter matrices with each one containing $(d*)^2$ parameters. The vector ω represents constant components of the covariances and contains $d*$ parameters.

For the bivariate case and $p = q = 1$ we can write the model explicitly as

$$\begin{pmatrix} h_{11,t} \\ h_{12,t} \\ h_{22,t} \end{pmatrix} = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \omega_3 \end{pmatrix} + \begin{pmatrix} a_{11} \ a_{12} \ a_{13} \\ a_{21} \ a_{22} \ a_{23} \\ a_{31} \ a_{32} \ a_{33} \end{pmatrix} \begin{pmatrix} \varepsilon_{1,t-1}^2 \\ \varepsilon_{1,t-1}\varepsilon_{2,t-1} \\ \varepsilon_{2,t-1}^2 \end{pmatrix}$$

$$+ \begin{pmatrix} b_{11} \ b_{12} \ b_{13} \\ b_{21} \ b_{22} \ b_{23} \\ b_{31} \ b_{32} \ b_{33} \end{pmatrix} \begin{pmatrix} h_{11,t-1} \\ h_{12,t-1} \\ h_{22,t-1} \end{pmatrix}$$

By rearranging terms, we can write the second order process η_t as a vector autoregressive moving average (VARMA) process of order $(\max(p,q), p)$,

$$\eta_t = \omega + \sum_{i=1}^{\max(p,q)} (A_i + B_i)\eta_{t-i} - \sum_{j=1}^{p} B_j u_{t-j} + u_t, \qquad (13.60)$$

where $u_t = \eta_t - h_t$ is a vector white noise process, i.e. $\mathsf{E}[u_t] = 0$, $\mathsf{E}[u_t u_t^\top] = \Sigma_u$ und $\mathsf{E}[u_t u_s^\top] = 0$, $s \neq t$. In (13.60) we set $A_{q+1} = \ldots = A_p = 0$ if $p > q$ and $B_{p+1} = \ldots = B_q = 0$ if $q > p$. Often the VARMA representation of multivariate GARCH models simplifies the derivation of stochastic properties, as one can refer to known results of the VARMA literature.

In the Vec representation (13.59), the multivariate GARCH(p,q) process ε_t is covariance stationary if, and only if, all eigenvalues of the matrix

$$\sum_{i=1}^{\max(p,q)} (A_i + B_i)$$

are smaller than one in modulus, see Engle and Kroner (1995). In that case, the unconditional covariance matrix is given by

$$\sigma = \text{vech}(\Sigma) = \left(I_{d*} - \sum_{i=1}^{\max(p,q)} (A_i + B_i) \right)^{-1} \omega. \qquad (13.61)$$

In order to illustrate the prediction of volatility, let us consider the following often used GARCH(1,1) model. The optimal prediction with respect to the mean squared prediction error is the conditional expectation of volatility. Due to the law of iterated expectations, the k-step prediction of η_{t+k} is identical to the k-step prediction of h_{t+k}, that is,

$$\mathsf{E}[\eta_{t+k} \mid \mathcal{F}_t] = \mathsf{E}[\mathsf{E}(\eta_{t+k} \mid \mathcal{F}_{t+k-1}) \mid \mathcal{F}_t] = \mathsf{E}[h_{t+k} \mid \mathcal{F}_t].$$

Having information up to time t, the predictions for the next three time periods are given by

$$\mathsf{E}[\eta_{t+1} \mid \mathcal{F}_t] = h_{t+1}$$
$$\mathsf{E}[\eta_{t+2} \mid \mathcal{F}_t] = \omega + (A + B)h_{t+1}$$
$$\mathsf{E}[\eta_{t+3} \mid \mathcal{F}_t] = (I_{d*} + A + B)\omega + (A + B)^2 h_{t+1},$$

and it can be seen that in general, the k-step prediction with $k \geq 2$ is given by

$$\mathsf{E}[\eta_{t+k} \mid \mathcal{F}_t] = \left\{I_{d*} + (A + B) + \ldots + (A + B)^{k-2}\right\} \omega + (A + B)^{k-1} h_{t+1}.$$

This converges to the unconditional covariance matrix $\sigma = (I_{d*} - A - B)^{-1}\omega$ if, and only if, the process is covariance stationary.

In the bivariate case ($d = 2$) and with $p = q = 1$, there are already 21 parameters that characterize the dynamics of volatility. In order to obtain a feasible model for empirical work, one often imposes restrictions on the parameter matrices of the Vec model. Bollerslev, Engle and Wooldridge (1988) propose using diagonal parameter matrices such that the conditional variance of one variable only depends on lagged squared values of the same variable, and the conditional covariances between two variables only depend on lagged values of the cross-products of these variables. This model substantially reduces the number of parameters (in the above case from 21 to 9), but potentially important causalities are excluded.

For parameter estimation the QML method is suitable. The conditional likelihood function for a sample time series of n observations is given by $\log L = \sum_{t=1}^{n} l_t$ with

$$l_t = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\log\{\det(H_t)\} - \frac{1}{2}\varepsilon_t^\top H_t^{-1}\varepsilon_t. \tag{13.62}$$

If the conditional distribution of ε_t is not normal, then (13.62) is interpreted as a quasi likelihood function, which serves merely as a target function in the numerical optimization, but which does not say anything about the true distribution. In the multivariate case, the QML estimator is consistent and asymptotically normal under the main assumptions that the considered process is strictly stationary and ergodic with a finite eighth moment. Writing all parameters in one vector, θ, we obtain the following standard result.

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathrm{N}(0, J^{-1}IJ^{-1}), \tag{13.63}$$

where I is the expectation of outer product of the score vector (i.e. the vector $\partial l_t/\partial\theta$), and J the negative expectation of the Hessian (i.e. the matrix of second derivatives). In the case of a normal distribution, we have $I = J$ and the asymptotic distribution simplifies to

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathrm{N}(0, J^{-1}). \tag{13.64}$$

In other words, these results are completely analogous to the univariate case, but the analytical expressions for I and J become much more complicated. Of course one can also determine I and J numerically, but this can lead to unreliable results, especially for J, in the multivariate case.

In empirical work one often finds that estimated standardized residuals are not normally distributed. In this case the QML function would be mis-specified and provides only consistent, inefficient parameter estimators. Alternatively, one can assume that the true innovation distribution is given by some specific non-normal parametric distribution, but in general this does not guarantee that parameter estimates are consistent in the case that the assumption is wrong.

13.4.2 The BEKK Specification

Engle and Kroner (1995) discuss the following specification of a multivariate GARCH model.

$$H_t = C_0 C_0^\top + \sum_{k=1}^{K} \sum_{i=1}^{q} A_{ki}^\top \varepsilon_{t-i} \varepsilon_{t-i}^\top A_{ki} + \sum_{k=1}^{K} \sum_{j=1}^{p} B_{kj}^\top H_{t-j} B_{kj}. \tag{13.65}$$

In (13.65), C_0 is a lower triangular matrix and A_{ki} and B_{ki} are $d \times d$ parameter matrices. For example, in the bivariate case with $K = 1$, $p = 1$ and $q = 0$, the conditional variance of ε_{1t} can be written as

$$h_{11,t} = c_{11}^2 + a_{11}^2 \varepsilon_{1t}^2 + a_{12}^2 \varepsilon_{2t}^2 + 2a_{11}a_{12}\varepsilon_{1t}\varepsilon_{2t}$$

and the conditional covariance as

$$h_{12,t} = c_{11}c_{21} + a_{11}a_{21}\varepsilon_{1t}^2 + a_{12}a_{22}\varepsilon_{2t}^2 + (a_{12}a_{21} + a_{11}a_{22})\varepsilon_{1t}\varepsilon_{2t}$$

The so-called BEKK specification in (13.65) guarantees under weak assumptions that H_t is positive definite. A sufficient condition for positivity is, for example, that at least one of the matrices C_0 or B_{ki} have full rank and the matrices $H_0, \ldots, H_{1-p}$ are positive definite. The BEKK model allows for dependence of conditional variances of one variable on the lagged values of another variable, so that causalities in variances can be modelled. For the case of diagonal parameter matrices A_{ki} and B_{ki}, the BEKK model is a restricted version of the Vec model with diagonal matrices.

Due to the quadratic form of the BEKK model, the parameters are not identifiable without further restriction. However, simple sign restrictions will give identifiability. For example, in the often used model $K = 1$ and $p = q = 1$, it suffices to assume that the upper left elements of A_{11} and B_{11} are positive. The number of

parameters reduces typically strongly when compared to the Vec model. For the above mentioned case, the number of parameters reduces from 21 to 11.

For each BEKK model there is an equivalent Vec representation, but not vice versa, so that the BEKK model is a special case of the Vec model. To see this, just apply the vech operator to both sides of (13.65) and define $\omega = L_d(C_0 \otimes C_0)^\top D_d \text{vech}(I_d)$, $A_i = \sum_{k=1}^{K} L_d(A_{ki} \otimes A_{ki})^\top D_d$ and $B_j = \sum_{k=1}^{K} L_d(B_{kj} \otimes B_{kj})^\top D_d$. Here $\otimes$ denotes the Kronecker matrix product, and L_d and D_d are the elementary elimination and duplication matrices. Therefore, one can derive the stochastic properties of the BEKK model by those of the Vec model. For the empirical work, the BEKK model will be preferable, because it is much easier to estimate while being sufficiently general.

13.4.3 The CCC Model

Bollerslev (1990) suggested a multivariate GARCH model in which all conditional correlations are constant and the conditional variances are modelled by univariate GARCH models. This so-called CCC model (constant conditional correlation) is not a special version of the Vec model, but belongs to another, non-linear model class. For example, the CCC(1,1) model is given by

$$h_{ii,t} = \omega_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i h_{ii,t-1},$$

$$h_{ij,t} = \rho_{ij} \sqrt{h_{ii,t} h_{jj,t}}$$

for $i, j = 1 \ldots, d$, and ρ_{ij} equal to the constant correlation between ε_{it} and ε_{jt}, which can be estimated separately from the conditional variances. The advantage of the CCC model is in the unrestricted applicability for large systems of time series. On the other hand, the assumption of constant correlation is possibly quite restrictive. For example, in the empirical analysis of financial markets one typically observes increasing correlation in times of crisis or in crash situations.

13.4.4 The DCC Model

Engle (2002) introduced a generalization of the CCC model that allows for time-varying correlations in a straightforward way. Similar to the CCC model the idea is to separate the modelling of volatilities and correlations and render estimation feasible in high dimensions. Again, volatilities $h_{ii,t}$ could be modelled in a first step by univariate GARCH(1,1) models as discussed in the previous section for the CCC

Table 13.5 Descriptive statistics of DAX and FTSE 100 return processes from 1998 to 2007   Q SFEmvol01

	Min.	Max.	Mean	Median	Std. error
DAX	−0.08875	0.07553	0.00025	0.00055	0.00031
FTSE 100	−0.05885	0.05903	0.00009	0.00007	0.00022

model. Rather than assuming that the conditional correlation ρ_{ij} between the i-th and j-th component is constant, it is now the ij-th element of the matrix

$$R_t = \operatorname{diag}(Q_t)^{-1/2} Q_t \operatorname{diag}(Q_t)^{-1/2}$$

with

$$Q_t = \Omega + \gamma v_{t-1} v_{t-1}^\top + \delta Q_{t-1},$$

where $v_t = (\varepsilon_{1,t}/\sqrt{h_{11,t}}, \ldots, \varepsilon_{d,t}/\sqrt{h_{dd,t}})^\top$, Ω is a symmetric positive definite parameter matrix, and γ and δ are positive scalars. Note that by construction, Q_t is positive definite and hence R_t is a correlation matrix, i.e. it is positive definite and all elements on the diagonal are equal to one. A similar specification has been suggested by Tse and Tsui (2002).

13.4.5 An Empirical Illustration

We consider a bivariate index return example, two indices, DAX and FTSE 100. The sample period is from 1998 to 2007, with altogether $n = 2{,}607$ observations. Figure 13.3 shows the time series of returns on both stock indices. Table 13.5 provides some simple descriptive statistics of returns ε_t. Apparently, the empirical mean of both processes is close to zero.

As can be seen in Fig. 13.3, the stock index returns follow a pattern that resembles a GARCH process: there is a clustering of volatilities in both series, and the clusters tend to occur simultaneously. This motivates an application of a bivariate GARCH model.

A first simple method to estimate the parameters of a BEKK model is the BHHH algorithm. This algorithm uses the first derivatives of the QML with respect to the 11 parameters that are contained in C_0, A_{11} and G_{11}, recalling Eq. (13.65). As this is an iterative procedure, the BHHH algorithm needs suitable initial parameters. For the diagonal elements of the matrices A_{11} and B_{11}, values between 0.3 and 0.9 are sensible, because this is the range often obtained in estimations. For the off-diagonal elements there is no rule of thumb, so one can try different starting values or just set them to zero. The starting values for C_0 can be obtained by the starting values for A_{11} and B_{11} using the formula for the unconditional covariance matrix and matching the sample covariance matrix with the theoretical version.

For the bivariate stock index example, we obtain the following estimates:

$$\hat{\theta} = \begin{pmatrix} 0.001669 \\ 0.001028 \\ 0.000786 \\ 0.324305 \\ -0.000001 \\ -0.000001 \\ 0.324306 \\ 0.942145 \\ -0.000007 \\ -0.000001 \\ 0.942140 \end{pmatrix}$$

$$\sum_{t=1}^{n} l_t = -17146.$$

The last expression represents the computed minimum of the negative log likelihood function. The displayed vector contains in the first three components the parameters in C_0, the next four components are the parameters in A_{11} and the last four components are the parameters in B_{11}.

In this example we thus obtain as estimated parameters of the BEKK model:

$$C_0 = 10^{-3} \begin{pmatrix} 1.67 & 1.03 \\ 0.00 & 0.79 \end{pmatrix},$$

$$A_{11} = \begin{pmatrix} 0.32 & -0.00 \\ -0.00 & 0.32 \end{pmatrix}, \; B_{11} = \begin{pmatrix} 0.94 & 0.00 \\ 0.00 & 0.94 \end{pmatrix}. \tag{13.66}$$

Estimates for the conditional covariances are obtained by applying successively the difference equation (13.65), where the empirical covariance matrix

$$\hat{H}_0 = \frac{1}{n} \sum_{t=1}^{n} \varepsilon_t \varepsilon_t^\top$$

of the observations ε_t is taken as initial value.

In Fig. 13.11 estimated conditional variance and covariance processes are compared. The upper and lower plots show the variance of the DAX and FTSE 100 returns and the plot in the middle shows the estimated conditional covariance process. Apart from a very short period at the beginning of the sample, the covariance is positive and of significant magnitude. This confirms our intuition of mutual dependence in stock markets which motivated the use of the bivariate GARCH model.
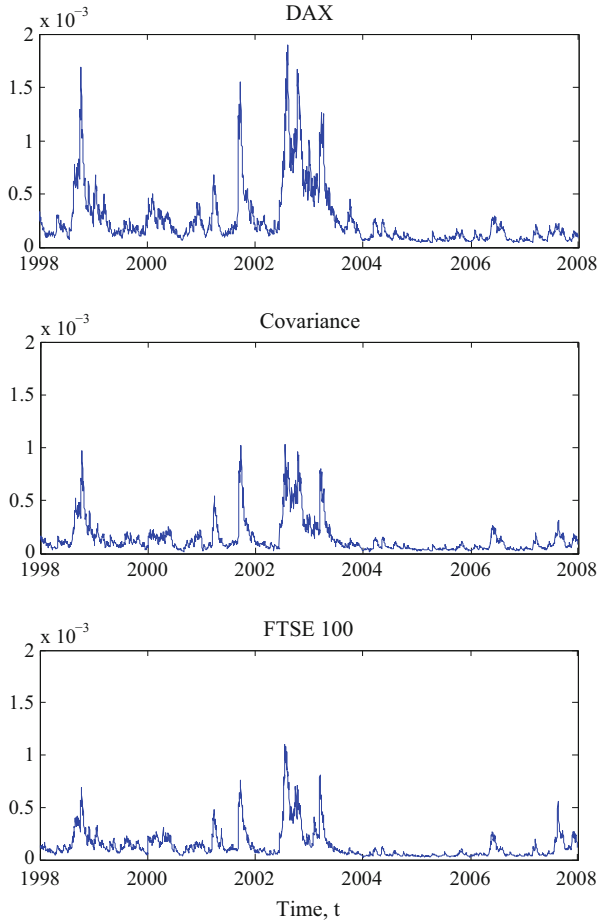
Fig. 13.11 Estimated variance and covariance processes, $\hat{H}_t$ Q SFEmvol02

The estimated parameters can also be used to simulate volatility. This can be done by drawing one realization of a multivariate normal distribution with mean zero and variance $\hat{H}_t$ at every time step. With these realizations one updates $\hat{H}_t$ according to Eq. (13.65). Next, a new realization is obtained by drawing from $N(0, \hat{H}_{t+1})$, and so on. We will now apply this method with $n = 3,000$. The results of the simulation in Fig. 13.12 show similar patterns as in the original process (Fig. 13.11). For a further comparison, we include two independent univariate GARCH processes fitted to the two stock index return series. This corresponds to a bivariate Vec representation with diagonal parameter matrices. Obviously, both methods capture the clustering of volatilities. However, the more general bivariate model also captures a spill over effect, that is, the increased uncertainty in one of the returns due to increased volatility in the other returns. This has an important impact on the amplitude of volatility.
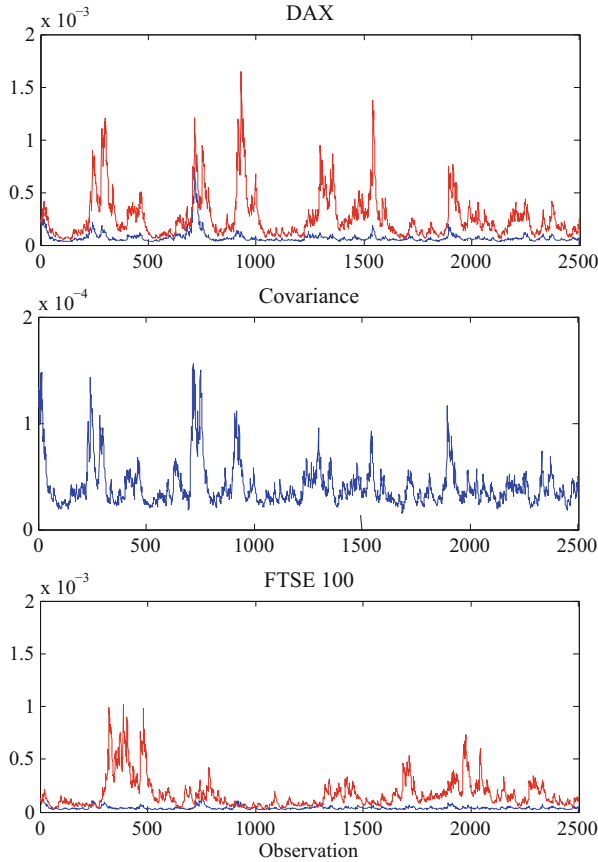
Fig. 13.12 Simulated variance and covariance processes with a bivariate (*blue*) and two univariate (*red*) GARCH processes, $\hat{H}_t$ Q SFEmvol03

13.5   Continuous-Time GARCH Models

Over many years, stochastic volatility models have been investigated mostly in discrete time. However, with modern easy access to voluminous quantities of high-frequency data, a demand for continuous-time models which allow, for instance, for a more natural analysis of possibly irregularly spaced data, has arisen. Moreover, the fact that the theory of option pricing is usually formulated in continuous-time is another motivation to develop appropriate continuous-time stochastic volatility models.

A first attempt to create a continuous-time GARCH model dates back to Nelson (1990b), who, by taking a limit of discrete-time GARCH processes, derived a bivariate diffusion, driven by two independent Brownian motions. This property, to incorporate two independent sources of uncertainty, is, of course, counterintuitive

to the GARCH paradigm. Accordingly, Nelson's continuous-time limit process does not possess the feedback mechanism whereby a large innovation in the mean process produces a burst of higher volatility, which is a distinctive feature of the discrete-time GARCH process. Moreover, the diffusion limit no longer has the heavy tailed distribution of returns needed for realistic modelling of returns in high-frequency financial data.

To overcome these problems, Klüppelberg, Lindner and Maller (2004) suggested an extension of the GARCH concept to continuous-time processes. Their *COGARCH* (continuous-time GARCH) model is based on a single background driving (continuous-time) Lévy process, which preserves the essential features of discrete-time GARCH processes, and is amenable to further analysis, possessing useful Markovian and stationarity properties. For details and properties concerning Lévy processes we refer to Applebaum (2004), Bertoin (1996), and Sato (1999).

13.5.1   COGARCH(1,1): Definition and Properties

The COGARCH(1,1) model is defined as follows:

Definition 13.4 (COGARCH(1,1)) Let $L = (L(t))_{t \geq 0}$ be a background driving Lévy process defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, (\mathcal{F}_t)_{t \geq 0})$ satisfying the "usual hypothesis" (see Protter 2005). Assume that $\mathsf{E}[L(1)] = 0$ and $\mathsf{E}[L^2(1)] = 1$. Given parameters (β, η, φ), with $\beta > 0$, $\eta > 0$, $\varphi \geq 0$, and a square integrable random variable $\sigma(0)$, independent of L, the COGARCH variance process $\sigma^2 = (\sigma^2(t))_{t \geq 0}$ is defined as the almost surely unique solution of the stochastic differential equation

$$\mathrm{d}\sigma^2(t) = \beta \mathrm{d}t - \eta \sigma^2(t-)\mathrm{d}t + \varphi \sigma^2(t-)\mathrm{d}[L, L](t), \quad t > 0. \tag{13.67}$$

The integrated COGARCH(1,1) process $G = (G(t))_{t \geq 0}$ in terms of L and σ is given by

$$G(t) = \int_0^t \sigma(s-)\,\mathrm{d}L(s), \quad t \geq 0. \tag{13.68}$$

The process $[L, L]$ appearing in Eq. (13.67) denotes the bracket process (quadratic variation) of L (cf. Protter 2005). As has been shown in Klüppelberg et al. (2004), the bivariate process $(\sigma(t), G(t))_{t \geq 0}$ is Markovian. Moreover, under a certain integrability condition, the process $\sigma^2(t)$ is strictly stationary (see Klüppelberg et al. 2004). As a consequence, G has stationary increments. Furthermore, Klüppelberg et al. (2004) also provides explicit expressions for the moments and autocorrelation functions of the variance process σ^2 and for the increments $G^{(r)}(t) \stackrel{\text{def}}{=} G(t + r) - G(t)$ of G. These can be used to estimate the COGARCH parameters by the method of moments.

Fasen, Klüppelberg and Lindner (2005) show that the COGARCH model, in general, exhibits regularly varying (heavy) tails, volatility jumps upwards and volatility clusters on high levels, which makes it appropriate for modelling financial time series. More precisely, it can be shown that both the tail of the distribution of the stationary volatility and the tail of the distribution of $G(t)$ are Pareto-like under weak assumptions, cf. Klüppelberg, Lindner and Maller (2006). For more details on the theoretical properties of G and σ^2, we refer to Klüppelberg et al. (2004), Fasen et al. (2005) and Klüppelberg et al. (2006).

13.5.2 Relation Between GARCH and COGARCH

The visual similarity between Eqs. (13.67) and (13.68) is obvious. However, the investigation of the exact relation between GARCH and COGARCH requires quite sophisticated concepts, which allow for a mathematically more precise formulation of "similarity" between different models. In this section we give a short summary of literature assessing the similarity of GARCH and COGARCH.

As one main result, Maller, Müller and Szimayer (2008) showed that the COGARCH model can be considered a continuous-time limit of GARCH models in a strong sense (in probability, in Skorokhod metric). Since this result can be used to develop a pseudo maximum likelihood (PML) estimation procedure for the COGARCH parameters, we briefly sketch the idea of the corresponding approximation.

Starting with a finite interval $[0, T]$, $T > 0$, one takes deterministic sequences $(N_n)_{n \geq 1}$ with $\lim_{n \to \infty} N_n = \infty$ and $0 = t_0(n) < t_1(n) < \ldots < t_{N_n}(n) = T$, and, for each $n = 1, 2, \ldots$, divides $[0, T]$ into N_n subintervals of length $\Delta t_i(n) \stackrel{\text{def}}{=} t_i(n) - t_{i-1}(n)$, for $i = 1, 2, \ldots, N_n$. Now define, for each $n = 1, 2, \ldots$, a discrete time process $(G_{i,n})_{i=1,\ldots,N_n}$ satisfying

$$G_{i,n} = G_{i-1,n} + \sigma_{i-1,n} \sqrt{\Delta t_i(n)} \tilde{\varepsilon}_{i,n}, \qquad (13.69)$$

where $G_{0,n} = G(0) = 0$, and the variance $\sigma_{i,n}^2$ follows the recursion

$$\sigma_{i,n}^2 = \beta \Delta t_i(n) + \left(1 + \varphi \Delta t_i(n) \tilde{\varepsilon}_{i,n}^2\right) e^{-\eta \Delta t_i(n)} \sigma_{i-1,n}^2. \qquad (13.70)$$

Here the innovations $(\tilde{\varepsilon}_{i,n})_{i=1,\ldots,N_n}$, $n = 1, 2, \ldots$, are constructed using a "first jump" approximation to the Lévy process developed by Szimayer and Maller (2007), which divides a compact interval into an increasing number of subintervals and for each subinterval takes the first jump exceeding a certain threshold. Finally, embed the discrete time processes $G_{\cdot,n}$ and $\sigma_{\cdot,n}^2$ into continuous-time versions G_n and σ_n^2 defined by

$$G_n(t) \stackrel{\text{def}}{=} G_{i,n} \quad \text{and} \quad \sigma_n^2(t) \stackrel{\text{def}}{=} \sigma_{i,n}^2, \quad \text{when } t \in [t_{i-1}(n), t_i(n)),\ 0 \leq t \leq T, \qquad (13.71)$$

with $G_n(0) = 0$. The processes G_n and σ_n are in $\mathbb{D}[0, T]$, the space of cádlág real-valued stochastic processes on $[0, T]$. Assuming that $\Delta t(n) \stackrel{\text{def}}{=} \max_{i=1,\dots,N_n} \Delta t_i(n) \to 0$ as $n \to \infty$, Maller et al. (2008) show that the discretised, piecewise constant processes $(G_n, \sigma_n^2)_{n \geq 1}$ defined by (13.71) converge to the COGARCH model (G, σ^2) defined by (13.67) and (13.68) in probability in the Skorokhod metric, as the discrete approximating grid grows finer.

Using a completely different approach, Kallsen and Vesenmayer (2009) derive the infinitesimal generator of the bivariate Markov process representation of the COGARCH model and show that any COGARCH process can be represented as the limit in law of a sequence of GARCH(1,1) processes. Moreover, they give some evidence that Nelson's diffusion limit and the COGARCH model are the only possible continuous-time limits of GARCH. Whereas the diffusion limit in law established by Nelson (1990b) occurs from GARCH by aggregating its innovations, the COGARCH limit arising in Kallsen and Vesenmayer (2009) and Maller et al. (2008) both occur when the innovations are randomly thinned.

Another concept, leading to a very strong measure of similarity between two statistical models, is Le Cam's framework of *statistical equivalence*. As was shown by Wang (2002), the diffusion limit in law established by Nelson (1990b) is not statistically equivalent to the approximating series of GARCH models. A recent paper by Buchmann and Müller (2008) investigates in detail the statistical equivalence between the GARCH and the COGARCH models. They show that, if the corresponding volatilities of the COGARCH process are unobservable, the limit experiment is again not equivalent to GARCH. If, however, in addition, full information about the volatility processes is available, then the limiting COGARCH experiment is equivalent to the approximating sequence of GARCH models.

13.5.3 Estimation of the COGARCH(1,1) Model

Haug et al. (2007) suggested moment estimators for the parameters of the COGARCH process based on equally spaced observations. The resulting estimators are consistent and asymptotically normal. The paper by Müller (2010) uses Bayesian methods to estimate the COGARCH model driven by a compound Poisson process. Based on the GARCH approximation to COGARCH, Maller et al. (2008) develop a straightforward and intuitive PML method. This estimation procedure produces good estimates of the COGARCH parameters β, η and φ, as has been shown in simulation studies. Moreover, the PML method is not restricted to equally spaced data or specific background driving Lévy processes.

As for other ML methods, the aim is to derive a likelihood function for the parameters of interest. To this end we assume that we are given observations $G(t_i)$, $0 = t_0 < t_1 < \dots < t_N = T$, on the integrated COGARCH as defined and parameterized in (13.67) and (13.68). The $\{t_i\}$ are assumed fixed (non-random) time

points. Now set $\Delta t_i \stackrel{\text{def}}{=} t_i - t_{i-1}$ and let $Y_i = G(t_i) - G(t_{i-1})$, $i = 1, \ldots, N$, denote the observed returns. From (13.68) we can write

$$Y_i = \int_{t_{i-1}}^{t_i} \sigma(s-) \mathrm{d}L(s),$$

and, because σ is Markovian, Y_i is conditionally independent of $Y_{i-1}, Y_{i-2}, \ldots$, given the natural filtration of the Lévy process L, with conditional expectation $\mathsf{E}(Y_i | \mathcal{F}_{t_{i-1}}) = 0$. The conditional variance is given by

$$\rho_i^2 \stackrel{\text{def}}{=} \mathsf{E}(Y_i^2 | \mathcal{F}_{t_{i-1}}) = \left(\sigma^2(t_{i-1}) - \frac{\beta}{\eta - \varphi} \right) \left(\frac{e^{(\eta - \varphi)\Delta t_i} - 1}{\eta - \varphi} \right) + \frac{\beta \Delta t_i}{\eta - \varphi}, \qquad (13.72)$$

see Maller et al. (2008). Next, $\sigma^2(t_{i-1})$ in Eq. (13.72) is replaced by the calculable quantity σ_{i-1}^2, which follows the recursive GARCH-type equation

$$\sigma_i^2 = \beta \Delta t_i + e^{-\eta \Delta t_i} \sigma_{i-1}^2 + \varphi e^{-\eta \Delta t_i} Y_i^2.$$

Now assume that the Y_i are conditionally $N(0, \rho_i^2)$ to get a pseudo-log-likelihood function $\mathcal{L}_N$ for $Y_1, Y_2, \ldots, Y_N$:

$$\mathcal{L}_N = \mathcal{L}_N(\beta, \eta, \varphi) = -\frac{1}{2} \sum_{i=1}^{N} \left(\frac{Y_i^2}{\rho_i^2} \right) - \frac{1}{2} \sum_{i=1}^{N} \log(\rho_i^2) - \frac{N}{2} \log(2\pi).$$

Taking as starting value for σ_0^2 the mean of the stationary variance process, $\mathsf{E}\,\sigma^2(0) = \beta/(\eta - \varphi)$, one can maximize the function $\mathcal{L}_N$ to get PML estimates of β, η and φ.

13.5.4   Extensions of the COGARCH Model

A couple of model extensions to the COGARCH model exist which we will briefly also mention in this section. For more details please refer to the appropriate literature.

COGARCH(p, q)

Motivated by the generalization of the GARCH(1,1) to the GARCH(p, q) model, Brockwell, Chadraa and Lindner (2006) introduced the COGARCH(p, q) model. Here the volatility follows a CARMA (continuous-time ARMA) process, which is again driven by a Lévy process. As in the discrete-time case, this model displays a broader range of autocorrelation structures than those of the COGARCH(1,1) process.

ECOGARCH(p,q)

Haug and Czado (2007) introduced an exponential continuous time GARCH
(ECOGARCH) process as analogue to the EGARCH(p, q) models. They investi-
gate stationarity and moments and show an instantaneous leverage effect for the
ECOGARCH(p, q) model. In a subsequent paper, Czado and Haug (2007) derived
a quasi-maximum likelihood estimation procedure for the ECOGARCH(1,1) model,
in the situation when it is driven by a compound Poisson process, assuming normally
distributed jumps.

Multivariate COGARCH

In Stelzer (2010) multivariate COGARCH(1,1) processes are introduced consti-
tuting a dynamical extension of normal mixture models and covering again such
features as dependence of returns (but without autocorrelation), jumps, heavy tailed
distributions, etc. As in the univariate case, the model has only one source of
randomness, a single multivariate Lévy process. The time-varying covariance matrix
is modelled as a stochastic process in the class of positive semi-definite matrices.
The paper analyses the probabilistic properties and gives a sufficient condition
for the existence of a stationary distribution for the stochastic covariance matrix
process, and criteria ensuring the finiteness of moments.

Multivariate ECOGARCH

As for the univariate COGARCH, the multivariate COGARCH can be extended
to a multivariate ECOGARCH model, as is the case in Haug and Stelzer (2010).
Analogously to the papers by Szimayer and Maller (2007) and Maller et al. (2008),
Stelzer (2009) generalizes the first jump approximation of a pure jump Lévy process,
which converges to the Lévy process in the Skorokhod topology in probability,
to a multivariate setting and an infinite time horizon. Applying this result to
multivariate ECOGARCH(1,1) processes, he shows that there is a sequence of
piecewise constant processes determined by multivariate EGARCH(1,1) processes
in discrete time which converge in probability in the Skorokhod topology to the
continuous-time process.

13.6 Recommended Literature

The empirical properties of financial market data, in particular the leptokurtosis
and clustering of volatilities, have been investigated systematically by Mandelbrot
(1963) and Fama (1965). ARCH Models were introduced by Engle (1982) and
generalized to GARCH by Bollerslev (1986). For an excellent overview of ARCH

models we refer to Gouriéroux (1997). An extensive discussion of the estimation of the tail exponent and the Hill estimator can be found in Embrechts, Klüppelberg and Mikosch (1997). A comparison of the efficiency of the Yule–Walker estimator with ML estimators of ARCH and GARCH models is given in Maercker (1997).

The theory of QML estimation was developed by White (1982) and Gouriéroux, Monfort and Trognon (1984). Weiss (1986) applied the theory to ARCH models, Lee and Hansen (1994) and Lumsdaine (1996) to GARCH models. Bollerslev and Wooldridge (1992) considered QML estimation for general conditionally heteroscedastic models. The idea of the *smooth transition* specification stems from Teräsvirta (1994) who used it for AR models. It was applied to GARCH models by Lubrano (1998). Hafner and Herwartz (2000) discussed various methods to test for the significance of the AR parameter under conditional heteroscedasticity. Moreover, they compare empirical results of AR models versus ARCH-M models, applied to several German stock returns.

Starica (2003) examines some shortfalls of GARCH, primarily its inability to function as a true data generating process and accurate forecasting tool as well as GARCH's overestimation of the unconditional variance and the existence of the IGARCH effect.

GARCH(1,1) model can be written in the form of the state space model. A closely related set-up, which shares with the state model representation the ability to incorporate our uncertainty about the form of the model is that of the non-parametric regression, which is discussed in Wand and Jones (1995).

In the multivariate case, Jeantheau (1998) has shown the consistency of the QML estimator, and Comte and Lieberman (2003) derived asymptotic normality. Analytical expressions for the score and Hessian matrices are provided by Hafner and Herwartz (2003). A recent survey of multivariate GARCH models is provided by Bauwens, Laurent and Rombouts (2005). It discusses the various specifications, estimation, inference and applications.

13.7 Exercises

Exercise 13.1 *For the time series of daily DAX and FTSE 100 returns from 1 January 1998 to 31 December 2007, graphically illustrate the following empirical functions:*

(a) autocorrelation function for plain returns,
(b) partial autocorrelation function for plain returns,
(c) autocorrelation function for squared returns, and
(d) autocorrelation function for absolute returns.

In addition, compute the Ljung–Box $\left(Q_m^* \right)$ test statistics, for plain returns, squared returns and absolute returns, as well as the ARCH test statistics for plain returns. Select the number of lags m close to $\log(n)$, where n denotes the sample size, see Tsay (2002).

Are the DAX and FTSE 100 return processes in the period under review:

(a) stationary,
(b) serially uncorrelated,
(c) independent?

Select an appropriate linear time series model for the return processes. Are ARCH and GARCH models appropriate for modelling the volatility processes of the analysed returns?

Exercise 13.2 *For modelling of the volatility processes for the DAX and FTSE 100 returns from* 1 *January* 1998 *to* 31 *December* 2007, *use an ARCH(q) model,* $q = 1, \ldots, 15$. *The return processes should follow the linear time series models discussed in Exercise 13.1. Based on the value of the optimized log-likelihood objective function select an appropriate order q for modelling and provide the return and volatility equations with estimated parameters.*

Furthermore, create a time series of estimated volatility processes in the period under review. Create forecasts of the volatility processes from 1 *January* 2008 *until* 31 *December* 2008 *using the unconditional and the conditional volatility approach.*

Exercise 13.3 *For modelling of the volatility processes for the DAX and FTSE 100 returns from* 1 *January* 1998 *to* 31 *December* 2007, *we use the GARCH(1,1), GARCH(2,1), GARCH(1,2) and GARCH(2,2) model. Compute the value of the log-likelihood objective function and select the best model. Additionally, provide the return and volatility equations with estimated parameters. Create the time series of estimated and forecasted volatility processes based on the unconditional and the conditional volatility for the best model among the estimated ones.*

Exercise 13.4 *Why is a GARCH model sometimes not appropriate to model financial time series? Mention and describe briefly at least one more appropriate specification of the ARCH model.*

Exercise 13.5 *Analyse an appropriately parameterized ARCH(1) process. Show that as a model for a financial time series, this process reasonably captures the following stylized facts:*

(a) heavy tails,
(b) white noise structure,
(c) volatility clustering.

Exercise 13.6 *For an ARCH(1) process, show that*

$$\mathsf{E}[\sigma_{t+s}^2 | \mathcal{F}_{t-1}] = \frac{1-\alpha^s}{1-\alpha}\omega + \alpha^s \sigma_t^2, \quad s \geq 1.$$

Interpret this result.

Exercise 13.7 (Representation of a Strong GARCH(p,q) Process) *Assume that* $p = q$. *Otherwise, we can successively add coefficients* $\alpha_{q+1} = 0$ *or* $\beta_{p+1} = 0$ *until the condition is fulfilled. Let* $0 < \sum_{i=1}^{p}(\alpha_i + \beta_i) < 1$. *Show that*

$$\sigma_t^2 = \omega \sum_{k=0}^{\infty} \sum_{i_1,\dots,i_k=1}^{p} (\alpha_{i_1} Z_{t-i_1}^2 + \beta_{i_1}) \cdots (\alpha_{i_1} Z_{t-i_1-\dots-i_k}^2 + \beta_{i_k}), \qquad (13.73)$$

where Z_t *denotes the innovation* ε_t/σ_t. *In particular, show that the sum on the right-hand side converges (i.e. is finite) almost surely. This result is used in the following exercise.*

Exercise 13.8 (Model Identifiability) *A discussion of a model's identifiability precedes any sound implementation or statistical inference involving the model.*

(a) *Using the representation (13.73), specify a GARCH process that does not admit a unique parametrization.*
(b) *Show that the GARCH(1,1) process*

$$\sigma_t^2 = 1 + \frac{1}{4}\varepsilon_{t-1}^2 + \frac{1}{2}\sigma_{t-1}^2$$

and the GARCH(2,2) process

$$\sigma_t^2 = \frac{5}{4} + \frac{1}{4}\varepsilon_{t-1}^2 + \frac{1}{16}\varepsilon_{t-2}^2 + \frac{1}{4}\sigma_{t-1}^2 + \frac{1}{8}\sigma_{t-2}^2$$

are equivalent, i.e. the given relationships are satisfied by the same process (σ_t^2). *What is noteworthy about the polynomials* $p(x) = \alpha_1 x + \alpha_2 x^2$ *and* $q(x) = 1 - \beta_1 x - \beta_2 x^2$ *for the second process? Recast your observation as a hypothesis concerning the identifiability of a general GARCH(p, q) process.*

Exercise 13.9 (Yule–Walker Estimator) *GARCH models are typically estimated by a numerical implementation of maximum likelihood methods. This procedure has the disadvantage that it does not yield a closed-form estimate and can produce different results depending on the algorithm and its starting value. As an alternative, derive the closed-form Yule–Walker moment estimator of the strong GARCH(1,1) process*

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2.$$

(a) *Express the process as an ARMA(1,1) process in* ε_t^2. *Compute the autocorrelations* $\rho_{\varepsilon^2}(1)$ *and* $\rho_{\varepsilon^2}(2)$. *Express* $\alpha + \beta$ *in terms of* $\rho_{\varepsilon^2}(1)$ *and* $\rho_{\varepsilon^2}(2)$.
(b) *Rewrite* $\rho_{\varepsilon^2}(1)$ *and* $\rho_{\varepsilon^2}(2)$ *as a quadratic equation* $\beta^2 - c\beta - 1 = 0$ *in* β *for an appropriate constant* c *depending on* $\alpha + \beta$.

In practice the autocovariances $\rho_{\varepsilon^2}(s)$ can be estimated by

$$\hat{\rho}_{\varepsilon^2}(s) = \frac{\sum_{t=h+1}^{T} \left(\hat{\varepsilon}_t^2 - \hat{\sigma}^2\right)\left(\hat{\varepsilon}_{t-h}^2 - \hat{\sigma}^2\right)}{\sum_{t=1}^{T} \left(\hat{\varepsilon}_t^2 - \hat{\sigma}^2\right)},$$

where the estimated squared residuals $\hat{\varepsilon}_t^2$ and estimated unconditional variance $\hat{\sigma}^2$ are supplied by a preliminary ARMA(1,1) estimation. By plug-in and by the above calculations, we obtain an estimate for $\alpha + \beta$ and therefore for c, β and α.

Exercise 13.10 *Consider a GARCH(p, q) process. Compute the best one-step forecast $\hat{\varepsilon}_{t+1}$ of ε_{t+1} based on $\mathcal{F}_t$. What is the conditional variance of the forecast error? Provide a non-conditional confidence interval for ε_{t+1} with coverage rate $1 - \alpha$. Interpret the width of this interval.*

Chapter 14
Long Memory Time Series

Empirical studies involving economic variables such as price level, real output and nominal interest rates have been shown to exhibit some degree of persistence. Moreover, findings across several asset markets have revealed a high persistence of volatility shocks and that over sufficiently long periods of time the volatility is typically stationary with "mean reverting" behaviour. Such series are reported to be characterized by distinct, but non-periodic, cyclical patterns and their behaviour is such that current values are not only influenced by immediate past values but values from previous time periods. The terminology associated with a series with such characteristics is "long memory" or "long range dependence". If financial time series exhibit persistence or long-memory, then their unconditional probability distribution may not be normal. This has important implications for many areas in finance, especially asset pricing, option pricing, portfolio allocation and risk management. For example, optimal consumption/savings and portfolio decisions may become sensitive to the investment perspective if stock returns were long range dependent. Moreover, if futures prices exhibit long memory, then co-integrating relationships involving cash prices and futures prices may also exhibit long memory, implying relatively slow adjustments to re-establish equilibrium. Under long-range dependence, the Black–Scholes option pricing solution is fundamentally biased, as the underlying asset is better described by fractional Brownian motion (FBM), a more general stochastic process that nests geometric Brownian motion as a special case.

In this chapter we introduce the concept of long memory, fractional integration, theory and methodology on long memory testing and parameter estimation. Long memory models include the autoregressive fractional integrated moving average (ARFIMA) and a number of GARCH-type models.

14.1   Definition of Long Range Dependence

A long-memory property may be defined in many ways. Let X_t, $t \in \mathbb{Z}$, be a stationary process and ρ_k is the autocorrelation function of X_t. In time domain long range dependence is associated with significant dependence between observations widely separated in time. A usual definition of long range dependence for stationary process X_t is that the autocorrelation ρ_k slowly decays such that

$$\lim_{T \to \infty} \sum_{k=-T}^{T} |\rho(k)| = \infty$$

and therefore not summable.

An alternative one is that long memory can be defined by specifying a hyperbolic decay of autocorrelation ρ_k

$$\rho_k \sim k^{2d-1} L(k) \;\; \text{as } k \to \infty \;\; 0 < d < 0.5, \tag{14.1}$$

where d is so-called memory parameter and $L(x)$ is slowly varying function. A positive measurable function L in $(0, \infty)$ is called slowly varying, if for all $c > 0$

$$\frac{L(cx)}{L(x)} \to 1 \quad \text{for} \quad x \to \infty.$$

Basic examples of slowly varying functions are constants and logarithms.

In the frequency domain, X_t exhibits long range dependence if its spectral density

$$f(\lambda) = (2\pi)^{-1} \sum_{k=-\infty}^{\infty} \exp(-\mathbf{i}\lambda k)\gamma(k) \tag{14.2}$$

with $\gamma(k) = \mathsf{Cov}(z_t, z_{t+k})$ has a pole at the origin. A long memory process is characterized by

$$\lim_{\lambda_j \to 0+} f(\lambda_j) = C \lambda_j^{-2d} \tag{14.3}$$

near the zero frequency where $\lambda_j = \frac{2\pi j}{T}$, with $j = 1, \dots, [T/2]$ and C is a strictly positive constant.

The memory parameter d governs the long-run dynamics of the process and the behaviour of $f(\lambda)$ around the origin. Therefore, if the empirical interest lies in the long-run dynamics of the process, it is useful to specify the spectral density only locally in the vicinity of the origin.

14.2   Fractional Integration and Long-Memory

One feature of macroeconomic time series is that the level changes with time in a rather smooth way. To explain and model these smooth movements, it was common practice to assume that the series were stationary $I(0)$ around a deterministic trend (a polynomial and/or a trigonometric function of time fitted by regression techniques). A second approach argued that these fluctuations in the level were better explained by means of so-called unit roots, $I(1)$ models. In other words, that the change in level was "stochastic" rather than "deterministic". An approach which we would like to present here is in terms of fractionally integrated processes. The term "fractional" refers to a generalized operation of a non-integer order, whereas the "usual" order of the operation has to be an integer.

Suppose $X_t, t \in \mathbb{Z}$, is a stationary process, then the difference process $Y_t = X_t - X_{t-1}$ is also stationary. If Y_t can be expressed in the form $Y_t = (1-L)X_t$, where 1 is the identity operator and L is the shift operator, then the process $X_t = (1-L)^{-1}Y_t$ can be called an *integrated* process. It turns out that one can integrate a process a fractional number of times, while preserving stationarity. This leads to the class of *fractionally integrated* (FI) processes. Let $\varepsilon_t, t \in \mathbb{Z}$, be i.i.d. variables with zero mean and finite variance. Consider a simple generalization of a random walk,

$$(1-L)^d X_t = \varepsilon_t. \tag{14.4}$$

The process ε_t could itself be a stationary, $I(0)$ and invertible ARMA sequence, when its auto-covariances decay exponentially. For $d = 0$, $X_t = \varepsilon_t$, and thus, X_t is "weakly auto-correlated", also termed "weakly dependent". On the other hand, when $d > 0$, X_t is "strongly dependent", so-named because of the strong association between observations widely separated in time.

Constructing a fractionally integrated process depends on a definition of a fractional power of the difference operator $(1 - L)$. The starting point is Taylor expansion for all $d \in \mathbb{R}$:

$$(1-L)^d = 1 - dL + \frac{1}{2!}d(d-1)L^2 + \ldots$$

$$+ \frac{(-1)^j}{j!}d(d-1)\ldots(d-j+1)L^j + \ldots$$

$$= \sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)}L^j$$

where $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the gamma function. From the representation (14.4) X_t is said to be ARFIMA(0,d,0), the following infinite autoregressive and moving average, respectively, representations hold:

$$\sum_{j=0}^{\infty} \frac{\Gamma(j-d)}{\Gamma(-d)\Gamma(j+1)}X_{t-j} = \varepsilon_t$$

$$X_t = \sum_{j=0}^{\infty} \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} \varepsilon_{t-j} \tag{14.5}$$

ε_t are independent variables, so the autocorrelation function ρ_k of X_t is expressed as

$$\rho_k = \frac{\Gamma(1-d)\Gamma(k+d)}{\Gamma(d)\Gamma(k+1-d)}.$$

Using Stirling's formula $\left(\frac{\Gamma(j+a)}{\Gamma(j+b)} \approx j^{a-b} \right)$,

$$\rho_k \sim Ck^{2d-1} \quad as \ k \to \infty$$

where C is a strictly positive constant equal to $\frac{\Gamma(1-d)}{\Gamma(d)}$ and $d \in (0, 0.5)$. For $0 < d < 0.5$, the process is covariance stationary with autocorrelations that decay at the hyperbolic rate of k^{2d-1}, so it possesses the long range property (14.1).

The distinction between $I(d)$ processes is important since the persistence as described in (14.1) has economic interpretations. A particularly interesting region is $0.5 < d < 1$, where the process has infinite variance, displays a high degree of persistence as exemplified by Eq. (14.1) and yet mean reverts in the sense that the impulse response function slowly decays. The process is mean reverting for $d < 1$. For $d \in (-0.5, 0.5)$ the process is covariance stationary. For such a process the effect of a shock ε_t on X_{t+k} decays as k increases but does so much more slowly than when $d = 0$. For $d = 0$, the process $I(0)$ has no long-memory and is weakly correlated. For $d \in (-0.5, 0)$ the process has short memory and has been addressed as "anti-persistent" by Mandelbrot (1977). Table 14.1 presents a summary of time series characteristics associated with long range dependence.

The intuition of fractional integration clearly emerges if one considers the time and frequency domain of the process. Fractionally integrated processes have a time domain representation that naturally extends conventional ARMA models. Essentially, the ARFIMA (as well-known as FARIMA) class is an extension of the classic ARMA models that can exhibit the long range dependence property. The relation between the memory parameter d and persistence of a shock is easily understood in terms of the coefficients in the expansion in Eq. (14.5).

Table 14.1 Time series long memory characteristics

	Series	Memory	Mean reversion	Variance	ACF
$d \in (-0.5, 0)$	Frac. integ.	Anti persist.	Yes	Fin.	Hyperb.
$d = 0$	Stationary	Short	Yes	Fin.	Expon.
$d \in (0, 0.5)$	Frac. integ.	Long	Yes	Fin.	Hyperb.
$d \in [0.5, 1)$	Frac. integ.	Long	Yes	Infin.	Hyperb.
$d = 1$	Integ.	Infinite	No	Infin.	Linear

14.3 Long Memory and Self-similar Processes

The notion of long memory can also be formulated in terms of self-similar processes. The term self-similar was first used by Mandelbrot and Van Ness (1968) to designate those processes that are scalable over time without losing their scalable properties.

Definition 14.1 A stochastic process X_t, $t \in \mathbb{R}$, is self-similar if there exists $H > 0$ such that for any scaling factor $c > 0$

$$X_{ct} \overset{\mathcal{L}}{=} c^H X_t.$$

H is the self-similar exponent also known as the Hurst exponent. In other words, self-similar processes are stochastic models with the property that a scaling in time is equivalent in terms of the distribution to an appropriate scaling in space. Furthermore, if for any k points $t_1, t_2, \ldots, t_k$, the distribution of $(X_{t_1+c} - X_{t_1+c-1}, \ldots, X_{t_k+c} - X_{t_k+c-1})$ does not depend on c, then X_t is self-similar with stationary increments. A non-degenerate (Process X is degenerate when $\{X_t\}_{t \in \mathbb{R}} \equiv 0$) H-self-similar process cannot be stationary. A connection between self-similarity and long memory is that the class of stationary models given as increments of self-similar processes with stationary increments, exhibit long memory if the scaling exponent is "large" enough.

   The FBM is an example of a self-similar process of which the increment process exhibits long memory (Fig. 14.1).

Definition 14.2 A stochastic Gaussian process B_t^H, $t \in \mathbb{R}$, with continuous sample paths, which satisfies:

1. B_t^H is self-similar with $0 < H < 1$,
2. B_t^H has stationary increments

is called a FBM. It is called standard if $\sigma^2 = \mathsf{Var}(B_1^H)$.

Theorem 14.1 *FBM B_t^H satisfies the following conditions:*

1. $B_0^H = 0$,
2. $\mathsf{E}\, B_t^H = 0$,
3. $\mathsf{E}(B_t^H)^2 = \sigma^2 |t|^{2H}$,
4. $\mathsf{E}(B_t^H - B_s^H)^2 = \sigma^2 |t - s|^{2H}$,
5. $\mathsf{Cov}\left(B_t^H, B_s^H\right) = \frac{\sigma^2}{2}\left(|t|^{2H} + |s|^{2H} - |t - s|^{2H}\right)$.

Proof (1) For any $c > 0$ $B_0^H = B_{c0}^H \overset{\mathcal{L}}{=} c^H B_0^H$ what implies the fact that $B_0^H = 0$ almost surely.

(2) For all $t \in \mathbb{R}$ $\mathsf{E}\, B_{2t}^H = 2^H \mathsf{E}\, B_t^H$ and, simultaneously, stationarity of the increments provides $\mathsf{E}\, B_{2t}^H = \mathsf{E}(B_{2t}^H - B_t^H) + \mathsf{E}\, B_t^H = 2\,\mathsf{E}\, B_t^H$. This implies $\mathsf{E}\, B_t^H = 0$.
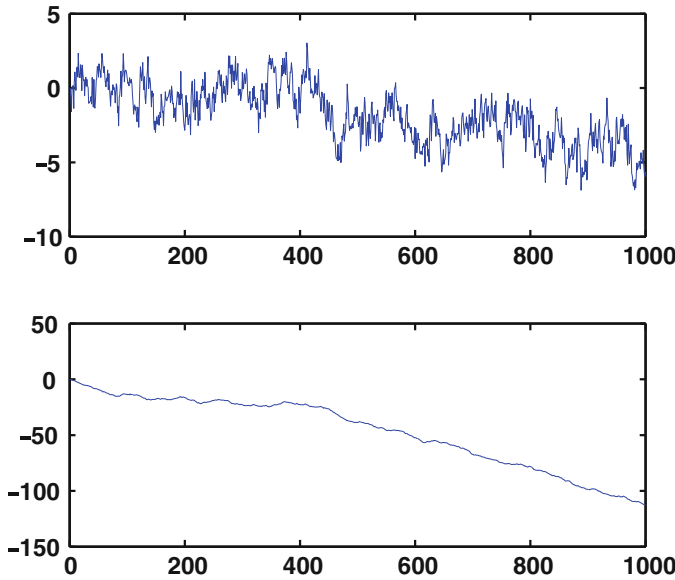
Fig. 14.1 A fractional Brownian motion with Hurst exponent $H = 0.2$ (*top*) and $H = 0.8$ (*bottom*)   🔲 SFEfbmplots

(3) By the self-similarity, $\mathsf{E}(B_t^H)^2 = \mathsf{E}(B_{1t}^H)^2 = |t|^{2H}\,\mathsf{E}(B_1^H)^2 = |t|^{2H}\sigma^2$.

(4) By the stationarity of the increments and the property (3), $\mathsf{E}(B_t^H - B_s^H)^2 = \mathsf{E}(B_{|t-s|}^H)^2 = \sigma^2|t-s|^{2H}$.

(5) By the property (2) and the fact that $xy = \frac{1}{2}\left\{x^2 + y^2 - (x-y)^2\right\}$, the covariance function of B_t^H can be expressed as:

$$
\begin{aligned}
\mathsf{Cov}\left(B_t^H, B_s^H\right) &= \frac{1}{2}\left\{\mathsf{Var}(B_t^H) + \mathsf{Var}(B_t^H) - \mathsf{Var}(B_t^H - B_s^H)\right\} \\
&= \frac{1}{2}\left(\sigma^2|t|^{2H} + \sigma^2|s|^{2H} - \sigma^2|t-s|^{2H}\right) \\
&= \frac{\sigma^2}{2}\left(|t|^{2H} + |s|^{2H} - |t-s|^{2H}\right).
\end{aligned}
$$
                                                                                                       □

When $H = 0.5$, the Brownian motion is obtained (independent increments). The class of FBM is divided into "anti-persistent" (covariance of increments is negative and increments of the FBM tend to be opposite signed) for $0 < H < 0.5$, and persistent, for $0.5 < H < 1$, processes.

Now let us consider the increment process of the FBM B_t^H, $t \in \mathbb{R}$, $0 < H < 1$, i.e.,

$$
Y_k = B_{k+1}^H - B_k^H, \; k \in \mathbb{Z}.
$$
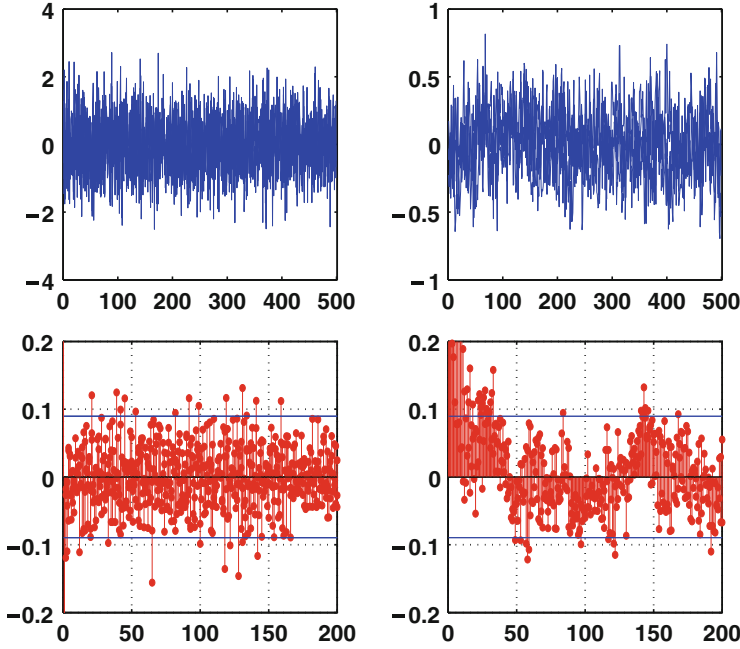
Fig. 14.2 *Upper panel*: simulated paths of the fractional Gaussian noise with Hurst exponent H = 0.2 (*left*) and H = 0.8 (*right*); *lower panel*: sample autocorrelation functions of the fractional Gaussian noise   🔍 SFEfgnacf

$\{Y_k\}_{k\in\mathbb{Z}}$ is called fractional Gaussian Noise (FGN) and possesses following properties:

1. $\{Y_k\}_{k\in\mathbb{Z}}$ is stationary
2. $\mathsf{E}\,Y_k = 0$
3. $\mathsf{E}\,Y_k^2 = \mathsf{E}(B_1^H)^2$

The autocorrelation function of Y_k has the form

$$\rho_k = \frac{1}{2}\left\{(|k|+1)^{2H} - 2|k|^{2H} + (|k|-1)^{2H}\right\}.$$

It is obvious that for $k \leq 0$ $\rho_k = \rho_{(-k)}$ and correlations are given by

$$\rho_k = k^{2H}\frac{1}{2}\left\{\left(1+\frac{1}{k}\right)^{2H} - 2 + \left(1-\frac{1}{k}\right)^{2H}\right\}.$$

For $H \neq 0$, using the first non-zero term in Taylor expansion of function $f(x) = (1+x)^{2H} - 2 + (1-x)^{2H}$, which is equal to $2H(2H-1)x^2$, as $k \to \infty$, the decay of ρ_k is:

$$\rho_k \approx H(2H-1)k^{2H-2}.$$

When $H \in (0, 0.5)$, the sum of the correlation of Y_t is finite what follows from the fact that $\sum_{k=-\infty}^{\infty} |k|^{2H-2} < \infty$, which implies a short-range dependence process. For $H = 0.5$, the Y_t are uncorrelated. For $H \in (0.5, 1)$, the process Y_t satisfies (14.1) and therefore exhibits long range dependence. In connection to fractional integration, a long memory time series is said to be fractionally integrated via the equality $d = H - 0.5$. For $H \in (0.5, 1)$, i.e. $d \in (0, 0.5)$, the series has long memory. If $H > 1$, i.e. $d > 0.5$, the series is non-stationary. If $H \in (0, 0.5)$, i.e. $d \in (-0.5, 0)$, the series is antipersistent (Fig. 14.2).

14.4 Detection of the Long Memory

In an intuitive geometric sense stationary long memory processes may be seen as processes that form a "hyperplane" separating non-stationary processes from the stationary processes. Research results show that stationary long memory processes often exhibit local spurious trends, making it hard to distinguish it from stochastic non-stationarity, Beran (1994). An important problem in econometrics is to distinguish between non-stationary unit-root models and stationary fractional models. Identifying spurious from genuine long memory is a persistent problem that is not easy to solve. It is understood that apparent long memory may be the artefact of the aggregation of short memory processes, structural breaks, level shifts or regime switch. For many financial time series the standard unit root tests cannot reject the null hypothesis of unit root, and most of these tests have low power against the FI alternative. We consider tests of long-memory that do not depend on a specific parametric form of the spectrum in the neighbourhood of the zero frequency. We present here tests of long memory including the rescaled range (R/S), the rescaled variance test (V/S) and the semiparametric (LobRob) tests. Another approach, presented here, is based on estimates of long memory parameter. Some methods of long memory estimation have been used to explain findings of spurious long memory, Diebold and Inoue (2001). It is therefore interesting to identify spurious long memory. We introduce the tests for spurious long memory proposed in Shimotsu (2006).

14.4.1 Rescaled Range and Rescaled Variance Test

One of the first tests for detecting possible long memory in a time series is the R/S statistic proposed by Hurst (1951) in his studies of river discharges. It is also known as the rescaled range or the range over standard deviation. The R/S statistic is

defined as the range of partial sums of deviations of a time series X_t from its mean $\overline{X}_T$, rescaled by its standard deviation:

$$R/S = \frac{1}{s_T}\left[\max_{1\le t\le T}\left\{\sum_{j=1}^{t}(X_j-\overline{X}_T)\right\} - \min_{1\le t\le T}\left\{\sum_{j=1}^{t}(X_j-\overline{X}_T)\right\}\right],$$

where $\sum_{j=1}^{t}(X_j-\overline{X}_T)$ are the partial sums of the observations and $s_T = \sqrt{\frac{1}{T-1}\sum_{j=1}^{T}(X_j-\overline{X}_T)^2}$ is the standard deviation. One looks at sub-intervals, where the partial sums (sums of deviations from a long-run mean) yield the highest and the lowest values, respectively. A (rescaled) difference of these two values assesses a possible change of variation in the series over the considered time span T. Intuitively, for the $I(0)$ time series with no long-range dependence, the value of the R/S statistic should be smaller than the R/S statistics of the $I(d)$ series, where the apparent variability depends on (possibly many) past observations. The R/S is directly linked to the Hurst exponent H: the slope $\hat{H}$ of the regression of $\log\{(R/S)_n\}$ on $\log(n)$, where $n = 1, \ldots, T$ is a sub-sample size and $(R/S)_n$ is the test statistic derived for the n sub-sample, may be taken as an estimator of the Hurst constant, see Beran (1994). The null hypothesis assumes no long range dependence against the fractional alternative $I(d)$.

Motivated by the R/S statistic, Giraitis et al. (2003) have proposed a rescaled variance test by centring the KPSS test statistic based on the partial sum of the deviation from the mean. The rescaled variance test is expressed by

$$V/S(q) = \frac{1}{T^2\hat{\sigma}_T^2(q)}\left[\sum_{k=1}^{T}\left\{\sum_{j=1}^{k}(X_j-\overline{X}_T)\right\}^2 - \frac{1}{T}\left\{\sum_{k=1}^{T}\sum_{j=1}^{k}(X_j-\overline{X}_T)\right\}^2\right],$$

where

$$\hat{\sigma}_T^2(q) = \hat{\gamma}_0 + 2\sum_{j=1}^{q}\left(1-\frac{j}{1+q}\right)\hat{\gamma}_j, \quad (q < T)$$

is the Newey and West (1987) heteroscedastic and autocorrelation consistent (HAC) estimator of the variance. $\hat{\gamma}_0$ is the variance of the process and the sequence $\{\hat{\gamma}_j\}_{j=1}^{q}$ denotes the autocovariances of the process up to the truncation lag q. The null hypothesis assumes $I(0)$ against the fractional alternative $I(d)$. The limiting distribution of this statistic is linked to the standard Kolmogorov statistic on testing cdf's in a sup norm deviation from a hypothesized cdf, so critical values can be obtained using a table of the standard Kolmogorov statistic. Choosing an optimal value of q is a vital problem. It might be useful to consider a range of q's, with respect to the length of the time series.

14.4.2   Semiparametric Test

The semiparametric estimator (LobRob) is based on the approximation of the spectrum of a long memory process. Lobato and Robinson (1998) proposed a test statistic:

$$t_{LobRob} = \sqrt{m}\frac{\hat{C}_1}{\hat{C}_0},$$

with $\hat{C}_k = m^{-1}\sum_{j=1}^{m}\zeta_j^k I(\lambda_j)$ and $\zeta_j = \log(j) - m^{-1}\sum_{j=1}^{m}\log(j)$, where

$$I(\lambda) = (2\pi T)^{-1}\left|\sum_{t=1}^{T} z_t \exp(\mathbf{i}\lambda t)\right|^2 \tag{14.6}$$

is the estimated periodogram. $\lambda_j = \frac{2\pi j}{T}$, $j = 1,\ldots,m \ll [T/2]$ is a degenerate band of Fourier frequencies with bandwidth parameter m. Under the null hypothesis $I(0)$, evaluated against the fractional alternative $I(d)$, the test statistic is asymptotically normally distributed. If the value of the test is in the lower tail of the standard normal distribution, the null hypothesis is rejected against the alternative that the series displays long memory. If the value of the test is in the upper tail of the standard normal distribution, the null hypothesis $I(0)$ is rejected against the alternative that the series is antipersistent.

14.4.3   Tests for Spurious Long Memory

Estimates of the memory parameter d can be used to test for spurious long memory. Shimotsu (2006) proposed identifying spurious long memory by computing tests on data using an adjusted KPSS and Phillips–Perron (PP) tests for the null of $I(0)$ and the null of $I(1)$, respectively. The intuition is that for an $I(1)$ process that appears to have a long memory $I(d)$ (e.g. a mean plus noise or a stochastic break model), taking the d-th difference will result in a new process that is $I(1-d)$. Applying the KPSS test will correctly reject the null that this new, fractional difference process is $I(0)$ but the PP test will fail to reject the null that the new process is $I(1)$. In principle, the contradictory results may be taken as evidence that the statistical behavior of the series in question may not be adequately described by the $I(0)$ or the $I(1)$ process but by $I(d)$.

On the other hand, consider a data generating process characterized by true long memory d. Taking the d-th difference will result in an $I(0)$ process. The KPSS test will correctly fail to reject null $I(0)$, while the PP test will correctly reject null $I(1)$. To evaluate the hypothesis of true long memory versus spurious long memory the adjusted KPSS test and PP tests are applied to the time series.

Let X_t represent a time series. Define the d-th difference of X_t as μ_t:

$$\hat{\mu}_t = (1 - L)^{\hat{d}} \{X_t - \hat{\mu}_0\} = \sum_{j=0}^{t-1} \frac{\Gamma(j - \hat{d})}{\Gamma(-\hat{d})\Gamma(j + 1)} L^j \{X_{t-j} - \hat{\mu}_0\}$$

where $\hat{d}$ is the estimate of d and $\hat{\mu}_0$ represents the (consistent) estimate of the mean of X_t. For stationary processes $I(d)$, where $d < 1$, $\hat{\mu}_0 = \overline{X}_T$. Apply the adjusted PP and KPSS tests to $\hat{\mu}_t$. The adjusted KPSS test statistic for $\hat{\mu}_t$ is defined as

$$\hat{\eta}_\mu = \frac{1}{T} \sum_{t=1}^{T} \frac{s_t^2}{s^2(\ell)},$$

$$s_t = \sum_{k=1}^{t} \varepsilon_k, \quad s^2(\ell) = \frac{1}{T} \sum_{t=1}^{T} \varepsilon_t^2 + \frac{2}{T} \sum_{s=1}^{\ell} \left(1 - \frac{s}{\ell + 1}\right) \sum_{t=s+1}^{T} \varepsilon_t \varepsilon_{t-s},$$

where ε_t is given by $\varepsilon_t = \hat{\mu}_t - \frac{1}{T} \sum_{s=1}^{T} \hat{\mu}_s$. The PP test with an intercept term is applied to the partial sum process of X_t. Critical values for the tests are given in Shimotsu (2006).

14.5   Estimation of the Long Memory Parameter

In this part we introduce some basic estimators of the long memory parameter based on both time and frequency domain of the time series. In the frequency domain, most commonly in empirical studies, the memory parameter is estimated using semiparametric estimators, among them are the log periodogram estimator (GPH) of Geweke and Porter-Hudak (1983) and the Gaussian semiparametric estimator (GSP) of Robinson (1995a). These estimators minimize some distance between the periodogram and the spectral density function, or approximations thereof, at low frequencies represented by the first m Fourier frequencies, $\lambda_j = \frac{2\pi j}{T}$, $j = 1, \ldots, m \ll [T/2]$. Estimation is usually between a set frequency band $(0, m]$ to capture the long run component (14.3) whilst the remainder of the frequencies capture the local variations. On the other hand, in the time domain, parametric estimation is based on the likelihood function.

14.5.1   Exact Maximum Likelihood Estimator

The Exact Maximum Likelihood (EML) estimation method by Sowell (1992) is a recursive procedure that allows for quick evaluation of the likelihood function. Consider the ARFIMA(p, d, q) process defined in Eq. (14.13) with $\mu = 0$,

$$\alpha(L)(1 - L)^d z_t = \beta(L)\varepsilon_t, \tag{14.7}$$

where $\alpha(L) = 1 - \alpha_1 L - \alpha_2 L^2 - \cdots - \alpha_p L^p$ and $\beta(L) = 1 + \beta_1 L + \beta_2 L^2 + \cdots + \beta_q L^q$ are the autoregressive and moving average lag-polynomials, respectively. Let $Z = (z_1, \ldots, z_T)^\top$ follow a normal distribution with $Z \sim N(0, \Sigma)$. The EML procedure allows for simultaneous estimation of both the long memory parameter and ARMA parameters. The maximum likelihood objective function is expressed as

$$\ell_E(\alpha, \beta, d; Z) = -\frac{T}{2} \log |\Sigma| - \frac{1}{2} Z^\top \Sigma^{-1} Z.$$

The EML estimator of d is given by

$$\hat{d}_{EML} = \arg \max_d \left\{ -\frac{T}{2} \log |\Sigma| - \frac{1}{2} Z^\top \Sigma^{-1} Z \right\} \tag{14.8}$$

14.5.2   Regression on the Periodogram

This method is based on the periodogram of the time series defined in (14.6). A series with long-range dependence has a spectral density proportional to $|\lambda|^{-2d}$ close to the origin, see (14.3). Geweke and Porter-Hudak (1983) (GPH) consider a regression of the logarithm of the periodogram against the logarithm of the frequency λ. The GPH estimator is based on the log linearization of (14.6):

$$\log \{I(\lambda_j)\} = C - d \log \{4 \sin^2(\lambda_j/2)\} + \varepsilon \tag{14.9}$$

The memory parameter d is estimated:

$$\hat{d}_{GPH} = -\frac{\sum_{j=1}^m (X_j - \bar{X}) \log \{I(\lambda_j)\}}{2 \sum_{j=1}^m (X_j - \bar{X})}.$$

We consider only harmonic frequencies $\lambda_j = \frac{2\pi j}{T}$, with $j \in (l, m]$, where l is a trimming parameter discarding the lowest frequencies and m is a bandwidth parameter. A necessary condition for consistency is that $\frac{m}{T} \to 0$ as $T \to \infty$. GPH proposed $m = T^\alpha$, $0 < \alpha < 1$. They argue that using $\alpha \in (0.5, 0.6)$ gives confidence intervals with accurate coverage probability. An optimal choice (in the sense of mean squared error minimizing), see Hurvich and Brodsky (1998), has been found to be a multiple of $[T^{0.8}]$ where the multiplicative constant depends on the short-run dynamics of the process. To check the robustness of the estimator it is advisable to apply different constants, e.g. $m = [T^{0.5}], [T^{0.525}], [T^{0.575}], [T^{0.60}], [T^{0.80}]$ where $[x]$ denotes the integer part of x. Velasco (1999) showed that for $d \in [0.5, 0.75]$ asymptotic normality and consistency of the GPH estimator is preserved. In the range $0.5 < d \leq 1$ the GPH for non-stationary fractionally integrated processes is consistent, see Kim and Phillips (2006).

14.5.3 Gaussian Semiparametric Estimator

Robinson (1995a) proposed the Gaussian Semiparametric estimator (GPS) using the local Whittle estimator $L_W(\theta)$. This estimator represents approximately a MLE in the frequency domain, since for large T,

$$I(\lambda_j) \sim \exp\left\{f(\lambda_j)^{-1}\right\}. \tag{14.10}$$

Given (14.10), the likelihood function is:

$$L\{I(\lambda_1), \ldots, I(\lambda_m), \theta\} = \prod_{j=1}^{m} \frac{1}{f_\theta(\lambda_j)} \exp\left\{-\frac{I(\lambda_j)}{f_\theta(\lambda_j)}\right\},$$

where $\theta = (C, d)$ is the parameter vector. The log-likelihood function becomes

$$\ell(\theta) = \sum_{j=1}^{m} \left\{-\log f_\theta(\lambda_j) - \frac{I(\lambda_j)}{f_\theta(\lambda_j)}\right\}. \tag{14.11}$$

In the neighbourhood of $\lambda \approx 0$ we obtain:

$$\ell(C, d) \sim \sum_{j=1}^{m} \left\{\log C - 2d \log(\lambda_j) + \frac{I(\lambda_j)}{C\lambda_j^{-2d}}\right\}, \tag{14.12}$$

$$\frac{\partial \ell(C, d)}{\partial C} = \sum_{j=1}^{m} \left\{\frac{1}{C} - \frac{I(\lambda_j)}{C^2 \lambda_j^{-2d}}\right\} = 0,$$

yielding $\hat{C} = m^{-1} \sum_{j=1}^{m} \left\{I(\lambda_j)/\lambda_j^{-2d}\right\}$. Inserting $\hat{C}$ for C in (14.12),

$$\ell(\hat{C}, d) = m \log \hat{C} - 2d \sum_{j=1}^{m} \log(\lambda_j) + m^2,$$

yields the minimization w.r.t. d:

$$\arg\min_d \ell(\hat{C}, d) = \log \hat{C} - \frac{2d}{m} \sum_{j=1}^{m} \log(\lambda_j) + m,$$

which is equivalent to

$$\hat{d}_{GSP} = \arg\min_d \left[\log\left\{\frac{1}{m} \sum_{j=1}^{m} \frac{I(\lambda_j)}{\lambda_j^{-2d}}\right\} - \frac{2d}{m} \sum_{j=1}^{m} \log(\lambda_j)\right].$$

For $m^* = [T/2]$, an approximation to the Gaussian log-likelihood, Beran (1994), is given by

$$\ell(\theta) = -(2\pi)^{-1} \sum_{j=1}^{m^*} \log f_\theta(\lambda_j) + \frac{I(\lambda_j)}{f_\theta(\lambda_j)},$$

for a given parametric spectral density $f_\theta(\lambda)$. Robinson (1995a) showed that the local Whittle estimator is consistent for $d \in (-0.5, 0.5)$. However, its consistency depends on the bandwidth m, which must satisfy $\frac{1}{m} + \frac{m}{T} \to 0$ as $T \to \infty$. This estimator is attractive due to its nice asymptotic properties, the mild assumptions underlying it and the likelihood interpretation. Robinson (1995a) showed that

$$\sqrt{m}(\hat{d}_{GSP} - d) \xrightarrow{\mathcal{L}} N(0, 1/4).$$

14.6   Long Memory Models

Long memory phenomena can be modelled in the framework of fractional integrated models of the ARMA or ARCH-type. We consider here the Autoregressive Fractional Integrated Moving Average (ARFIMA), the Fractional Integrated GARCH (FIGARCH), the Fractional Integrated Asymmetric Power ARCH (FIAPARCH) and the Hyperbolic-GARCH (HYGARCH) models.

14.6.1   ARFIMA Model

The ARFIMA(p, d, q) model is a generalization of the random walk process (14.4) where ε_t is an ARMA(p, q) process. Stationary process x_t is called the ARFIMA(p, d, q) process if it satisfies

$$\alpha(L)(1 - L)^d (x_t - \mu) = \beta(L)\varepsilon_t, \tag{14.13}$$

where μ is the mean of x_t, $\varepsilon_t \sim i.i.d(0, \sigma_\varepsilon^2)$, $\alpha(L) = 1 - \sum_{j=1}^p \alpha_j L^j$ and $\beta(L) = 1 + \sum_{j=1}^q \beta_j L^j$ are lag polynomials, $d \in (-0.5, 0.5)$ is the long memory parameter and $(1 - L)^d$ is the fractional difference filter. We assume that roots of both polynomials lie outside the unit circle and that the polynomials have no common roots (Fig. 14.3).

In the conditional mean, the ARFIMA model fills the gap between short and long run dependence in that, the short-run behaviour of the time series is captured by the ARMA parameters and the long-run dependence is captured by the fractional
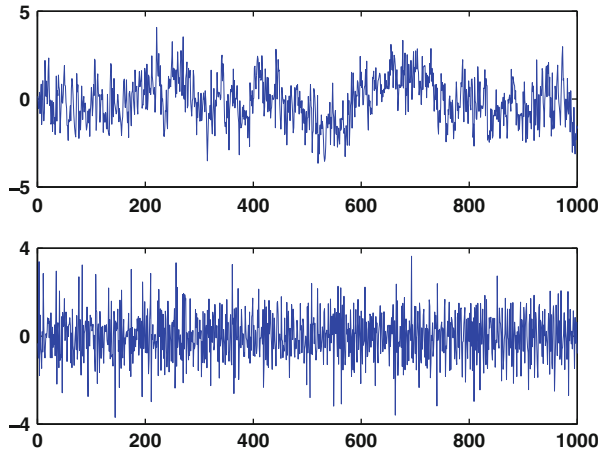
Fig. 14.3 Thousand simulated observations from ARFIMA models with Gaussian white noise N(0, 1). (*top*) ARFIMA(1,d,1) model with $d = 0.4, \alpha = 1, \beta = 1$; (*bottom*) ARFIMA(1,d,1) model with $d = -0.4, \alpha = 1, \beta = 1$ **Q** SFEarfima

differencing parameter. The process displays persistence when $d \in (0, 0.5)$ and anti-persistence when $d \in (-0.5, 0)$. For $d \in (0.5, 1)$ the process is non-stationary but mean-reverting. For $d \geq 1$ it is non-stationary and non-mean-reverting, with shocks effect persisting indefinitely.

The crucial fact in model estimation is the correct specification of the AR and MA orders, otherwise the estimates are likely to be inconsistent. Estimating the parameter d along with other parameters of the model requires estimating the autocovariance function in order to construct the autocorrelation function. Some estimation methods include the EML, (Quasi) Maximum Likelihood and Non-linear least squares.

14.6.2   GARCH Long Memory Models

We consider here a class of models that turn out not only to accommodate some of the stylized facts (time-varying volatility, clustering of volatility) revealed by data, but can capture also slow decaying correlations. Consider a mean process

$$x_t = \mu + \varepsilon_t,$$

with $\mathsf{E}[\varepsilon_t | \mathcal{F}_{t-1}] = 0$, $\mathsf{Var}[\varepsilon_t | \mathcal{F}_{t-1}] = \sigma_t^2$ and $\mathcal{F}_{t-1}$ is the information set up to time $t - 1$.

Let the GARCH(p, q) model for the conditional variance be represented as

$$\varepsilon_t = z_t \sigma_t, \qquad z_t = \frac{\varepsilon_t}{\sigma_t} \sim i.i.d.$$

$$\sigma_t^2 = \omega + \alpha(L)\varepsilon_t^2 + \beta(L)\sigma_t^2, \tag{14.14}$$

with $\alpha(L)$ and $\beta(L)$ as AR and MA lag order polynomials, respectively.

By defining $\upsilon_t = \varepsilon_t^2 - \sigma_t^2$ in Eq. (14.14), the ARMA(m, p) process can be expressed as

$$\{1 - \alpha(L) - \beta(L)\}\varepsilon_t^2 = \omega + \{1 - \beta(L)\}\upsilon_t,$$

where $m = \max(p, q)$. Allowing $\phi(L) = \{1 - \alpha(L) - \beta(L)\}$ to take on a unit root, Engle and Bollerslev (1986) defined the IGARCH(p, q) process:

$$\phi(L)(1 - L)\varepsilon_t^2 = \omega + \{1 - \beta(L)\}\upsilon_t,$$

where $\phi(L) = \sum_{i=1}^{m-1} \phi_i L^i$ is of order $m - 1$.

FIGARCH Model

The FIGARCH model combines features of the fractional integrated process for the mean with the regular GARCH process for the conditional variance.
The general FIGARCH(p, d, q) representation is given as

$$\phi(L)(1 - L)^d \varepsilon_t^2 = \omega + \{1 - \beta(L)\}\upsilon_t, \quad \upsilon_t = \varepsilon_t^2 - \sigma_t^2. \tag{14.15}$$

Rearranging (14.15) gives the conditional variance

$$\sigma_t^2 = \frac{\omega}{1 - \beta(L)} + \left\{ 1 - \frac{\alpha(L)(1 - L)^d}{1 - \beta(L)} \right\} \varepsilon_t^2$$

$$= \frac{\omega}{1 - \beta(L)} + \kappa(L)\varepsilon_t^2$$

where $d \in (0, 1)$, $\omega > 0$ and $\alpha, \beta < 1$. The effect of past innovations on the conditional variance dies out at a hyperbolic rate with increasing lag length. This differs from the GARCH process in which the effect of past squared innovations on the current conditional variance dies out exponentially. For $0 < d < 1$, the FIGARCH model captures the long-run persistence in the conditional volatility. For $d = 0$ or 1, the FIGARCH nests the GARCH and IGARCH models respectively. It should be noted that d in FIGARCH does not have the same interpretation

of persistence as d in ARFIMA, since the fractional differencing operator in the ARFIMA model applies only to the constant term in the mean equation while in FIGARCH it does not apply to the ω term.

Another parametrization for the FIGARCH$(1, d, 1)$ model is given by Chung (2001):

$$\alpha(L)(1 - L)^d(\varepsilon_t^2 - \sigma^2) = \beta(L)(\varepsilon_t^2 - \sigma^2),$$

where

$$\sigma^2 = \frac{\omega}{1 - \alpha_1 - \beta_1}$$

is the unconditional variance of ε_t. The conditional variance is formulated as

$$\sigma_t^2 = \sigma^2 + \left\{ 1 - \frac{\alpha(L)(1 - L)^d}{1 - \beta(L)} \right\} (\varepsilon_t^2 - \sigma^2).$$

14.6.3 FIAPARCH Model

Tse (1998) introduced this model by incorporating a fractional filter $(1 - L)^d$ into the Asymmetric Power ARCH (APARCH) model framework of Ding et al. (1993). The APARCH(p, q) can be expressed as

$$\sigma_t^\delta = \omega + \sum_{i=1}^{q} \alpha_i (|\varepsilon_{t-i}| - \gamma_i \varepsilon_{t-i})^\delta + \sum_{j=1}^{p} \beta_j \sigma_{t-j}^\delta,$$

where $\delta > 0$ and $-1 < \gamma_i < 1$. δ represents a Box–Cox transformation of σ_t and γ_i the "leverage" effect. For the FIAPARCH(p, d, q) model of Tse (1998), the conditional volatility specification is given as

$$\sigma_t^\delta = \omega + \left[1 - \{1 - \beta(L)\}^{-1} \alpha(L)(1 - L)^d \right] (|\varepsilon_t| - \gamma \varepsilon_t)^\delta.$$

The term $(|\varepsilon_t| - \gamma \varepsilon_t)^\delta$ reflects "leverage" effect, $\gamma \neq 0$. When $0 < \gamma < 1$, a positive innovation increases volatility less than a negative innovation and vice versa for $-1 < \gamma < 0$. For $\gamma = 0$, positive and negative innovations of the same magnitude have the same effect on volatility. The FIAPARCH model allows for volatility clustering, leptokurtosis, long memory and features an asymmetry in a relationship between the conditional variance and the lagged squared innovations. It nests the GARCH, IGARCH, GRJ-GARCH, APARCH, FIPARCH, FIGARCH and FIAGARCH models as special cases.

14.6.4 HYGARCH Model

The hyperbolic GARCH model proposed by Davidson (2004) extends the conditional variance of the FIGARCH(p, δ, q) model by introducing weights to the difference operator such that $(1 - L)^d = \{(1 - \alpha) + \alpha(1 - L)^d\}$. The general parametrization is given as

$$\sigma_t^2 = \frac{\omega}{1 - \beta(L)} + \left[1 - \frac{\alpha(L)\{1 + \phi(1 - L)^d\}}{1 - \beta(L)} \right] \varepsilon_t^2$$

where $\phi > 0$ are weights to $(1 - L)^d$ and $d > 0$. The HYGARCH(p, ϕ, d, q) nests the GARCH models (for $\phi = 0$), IGARCH (for $\phi = d = 1$) and the FIGARCH (for $\phi = 1$ or $\log \phi = 0$). When $\phi < 1$ ($\log \phi < 0$) the process is stationary. For added flexibility of the models in order to capture excess kurtosis and measure tail thickness, one can assume a (skewed) Student-t distribution of the innovations.

14.7 An Empirical Illustration

The data consist of daily closing prices for the English FTSE stock index, three German DAX stocks; Bayer, Siemens and Volkswagen, each from 04.01.2000 to 30.10.2006. We use the continuously compounded daily log returns (in %),

$$r_t = \log \frac{P_t}{P_{t-1}}$$

where P_t is the stock price at time t. Summary statistics presented in Table 14.2 show that all returns share similar properties in excess kurtosis. The FTSE index and

Table 14.2 Descriptive statistics: FTSE and DAX stocks (Bayer, Siemens and Volkswagen (VW))

	Stockindex			
	FTSE	Bayer	Siemens	VW
Std. Dev.	1.15	2.31	2.45	2.11
Min	−5.58	−19.42	−10.95	−9.65
Max	5.90	32.99	10.59	10.69
S	−0.16	1.10	0.07	−0.01
K	2.96	27.71	1.79	2.38
$Q^2(24)$	2,329.98	136.64	1,724.63	794.20

$Q^2(24)$ is the Box–Pierce statistic for remaining serial correlation in the squared standardized residuals using 24 lags, Skewness (S), excess Kurtosis (K). Sample: 04.01.2000–30.10.2006   SFEfdescStat
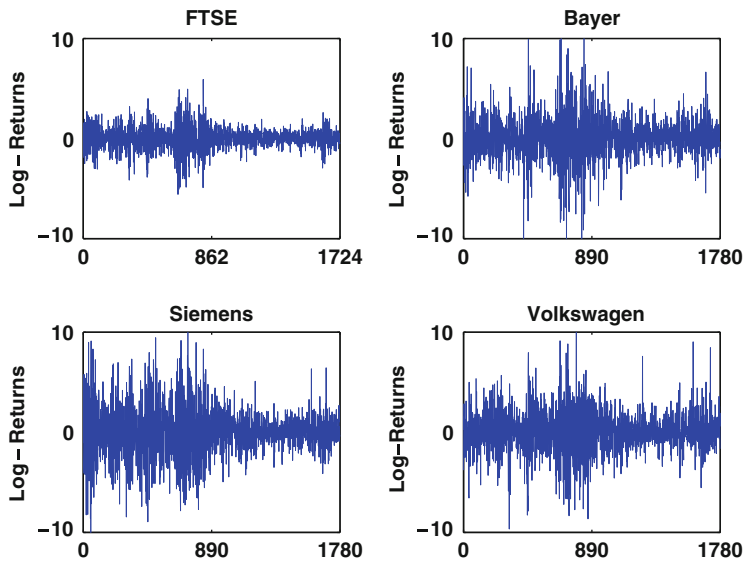
Fig. 14.4 Daily log returns of FTSE and DAX stocks, Bayer, Siemens and Volkswagen

Q SFEcomplogreturns

Table 14.3 Rescaled variance test on log returns for absolute return, $|r_t|$, square returns of the FTSE index and DAX (stocks: Bayer, Siemens (Siem.), Volkswagen (VW))

V/S	$|r_t|$				r_t^2			
q	FTSE	Bayer	Siem.	VW	FTSE	Bayer	Siem.	VW
5	2.51	−0.89	−0.86	−38.49	1.86	1.69	2.56	1.68
10	1.24	−0.58	−0.74	−0.80	0.92	1.01	1.32	0.88
32	0.43	−0.62	−1.30	−0.53	0.34	0.34	0.43	0.35
40	0.36	−0.94	−0.78	−0.40	0.29	0.29	0.35	0.30
110	0.17	−0.34	−0.18	−0.32	0.15	0.15	0.16	0.14

Sample size $T = 1{,}781$ and 5 % significance level   Q SFEresvarTest

the Volkswagen stock are negatively skewed. Box–Pierce statistic for the remaining serial correlation in the squared standardised residuals (using 24 lags) indicates high serial correlation in the variance (Fig. 14.4).

The V/S test results for long range dependence in the absolute and square returns are presented in Table 14.3. At a 5 % level, we fail to reject long memory for r_t^2 since the evaluated statistic is above the critical value, 0.1869. The long memory hypothesis is rejected for absolute returns. The *GPH* estimates (bandwidths, $m = T^{0.5}, m = T^{0.6}, m = T^{0.8}$) and the *GSP* estimates $\left(\text{bandwidths}, m = \left[\frac{T}{4}\right], m = \left[\frac{T}{16}\right], m = \left[\frac{T}{64}\right]\right)$ are reported in Table 14.4. Results from both estimators show that $d \in (0.09, 0.64)$ for r_t^2. To balance the tradeoff between bias and variance, estimates

Table 14.4 The GPH and GSP estimates for absolute return, $|r_t|$, square returns of the FTSE index and DAX stocks (Bayer, Siemens (Siem.), Volkswagen (VW))

m	$|r_t|$				r_t^2			
	FTSE	Bayer	Siem.	VW	FTSE	Bayer	Siem.	VW
GPH								
$T^{0.5}$	0.54	0.17	0.01	−0.18	0.43	0.24	0.64	0.47
$T^{0.6}$	0.53	−0.09	0.10	−0.04	0.59	0.27	0.50	0.40
$T^{0.8}$	0.27	−0.06	−0.03	−0.04	0.28	0.07	0.21	0.19
GSP								
$[T/64]$	0.52	0.09	−0.09	−0.09	0.41	0.34	0.60	0.40
$[T/16]$	0.53	0.02	0.05	−0.01	0.54	0.19	0.42	0.37
$[T/4]$	0.27	−0.02	−0.03	−0.04	0.28	0.10	0.26	0.22

Sample size $T = 1,781$ 🔲 SFEgphTestResults, 🔲 SFEgseTestResults
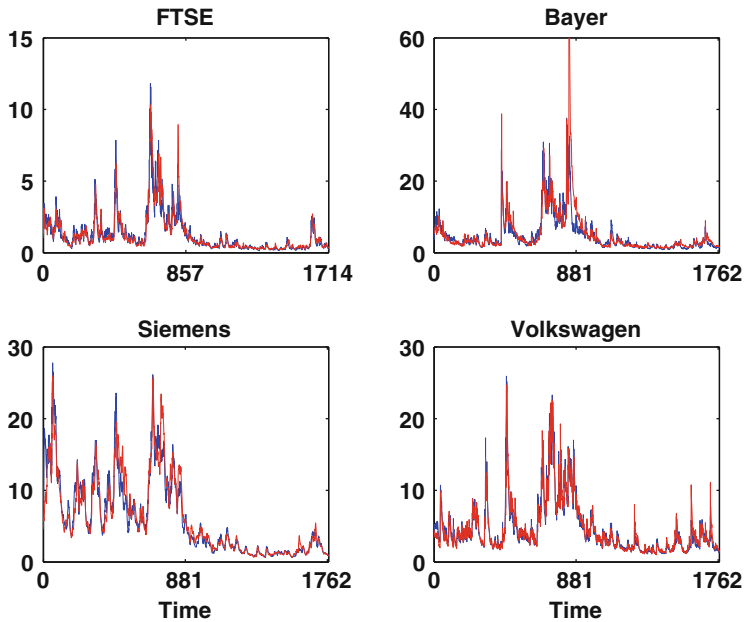
Fig. 14.5 Conditional volatility forecast for each log-return series based on FIAPARCH (*red*) and HYGARCH (*red*) models   🔲 SFECondVola

for bandwidths $m = T^\alpha$ with $\alpha = 0.525, 0.575$ for *GPH* and $m = \left[\frac{T}{8}\right], \left[\frac{T}{32}\right]$ for *GSP* are also investigated. Results in Table 14.4 show that the estimates of d for both estimators fall within $(-0.05 < d < 0.75)$, the range of consistency for fractional integrated process with long range dependence, see Kim and Phillips (2006).

With evidence of long memory in the volatility process, we fit FIAPARCH and HYGARCH models that can accommodate volatility clustering, leptokurtosis, asymmetry and long memory well. Figure 14.5 presents plots of the time path of

conditional standard deviation for each log return series based on the FIAPARCH and HYGARCH models. As observed, the FIAPARCH model is slightly better in estimating the magnitude of volatility than the HYGARCH. Our results suggest that such models hold the promise of improved long-run volatility forecast and more accurate pricing of long-term contracts. The implication can therefore be extended to the context of risk management.

14.8   Recommended Literature

The phenomenon of long memory was first introduced in the non-economic literature by Hurst (1951) and Hurst (1957) in studies on the persistence of the Nile river flow and reservoir capacity. Recent applications have used this phenomenon to study and analyse the fluctuations in financial markets. For an extensive survey on the long range dependence literature, see Beran (1994), Baillie (1996), Robinson and Henry (1999), Giraitis et al. (2003), Phillips and Shimotsu (2004) and the references therein. For details on self-similar processes, see Beran (1994), Samorodnitsky and Taqqu (1994) and Embrechts (2002). Geweke and Porter-Hudak (1983) used a log linearization of the periodogram to extract the long memory parameter d.

Chapter 15
Non-parametric and Flexible Time Series Estimators

With the analysis of (financial) time series, one of the most important goals is to produce forecasts. Using past data one can argue about the future mean, the future volatility, and so on, however a flexible method of producing such estimates will be introduced in this chapter.

Let (Y_t), $t = 0, 1, 2, \ldots$, be a time series. We consider a non-linear autoregressive heteroscedastic model of the form

$$Y_t = f(Y_{t-1}) + s(Y_{t-1})\xi_t, \quad t = 1, 2, \ldots \tag{15.1}$$

Here the innovations ξ_t are i.i.d. random variables with $\mathsf{E}[\xi_t] = 0$ and $\mathsf{E}[\xi_t^2] = 1$, $f : \mathbb{R} \longrightarrow \mathbb{R}$ and $s : \mathbb{R} \longrightarrow (0, \infty)$ are unknown functions, and Y_0 is independent of (ξ_t). Under these assumptions and according to Theorem 3.1 it holds that:

$$\mathsf{E}[Y_t \,|\, Y_{t-1} = x] = f(x) + \mathsf{E}[s(Y_{t-1})\xi_t \,|\, Y_{t-1} = x] = f(x) + s(x)\mathsf{E}[\xi_t] = f(x),$$

where in the second to last equation the independence of ξ_t and Y_{t-1} is used. A similar calculation gives $s^2(x) = \mathsf{Var}[Y_t \,|\, Y_{t-1} = x]$. The unknown functions f and s describe the conditional mean and the conditional volatility of the process, which we want to estimate.

With the specific choice $f(x) = \alpha x$ and $s = \sigma > 0$ the process Y_t is an AR(1) process. Every strong ARCH(1) process (Y_i) satisfies the model (15.1). In this case $f = 0$, so it holds that $s(x) = \sqrt{\omega + \alpha x^2}$ with the parameters $\omega > 0$ and $\alpha \geq 0$. With respect to the structure of the conditional mean and the conditional variance, the model above is another broad generalization of the (strong) ARCH models. The advantage of the *non-parametric* approach is that the model contains no structural assumptions about the functions f and s.

Closely related to our model is the Qualitative Threshold ARCH model (QTARCH) studied in Gouriéroux and Monfort (1992), which for the case of one lag (QTARCH(1)) is a special case of (15.1), where the unknown functions

$f : \mathbb{R} \longrightarrow \mathbb{R}$ and $s : \mathbb{R} \longrightarrow (0, \infty)$ take the form of step functions; see (13.32). On the other hand (15.1) can also be described under certain regularity assumptions on f and s as a limit model of the QTARCH(1) models when $J \to \infty$, in that f and s are approximated with elementary functions.

The work of Gouriéroux and Monfort is the first to consider the conditional mean and the conditional variance *together* in a non-parametric model. The applications of this idea introduced here are taken from Härdle and Tsybakov (1997), and are also considered independently in Franke et al. (2002). In the following we will construct a class of estimators based on the local polynomial regression for the conditional volatility $v(x) = s^2(x)$ and the conditional mean $f(x)$ of the time series (Y_i) under the model assumptions (15.1).

In addition to the model assumptions (15.1) certain regularity assumptions, although no structural assumptions, on f and s will be made. As the main result of this chapter we will show, this combined estimation of the conditional expectation and the conditional volatility is asymptotically normally distributed.

15.1   Non-parametric Regression

In non-parametric regression one is interested in the (functional) relationship between an explanatory variable X and a dependent variable Y, i.e. one is interested in obtaining an estimation for the unknown function $m(x) = \mathsf{E}[Y \mid X = x]$. In doing this, in contrast to parametric statistics, no special assumptions on the form of the function m is made. Only certain regularity and smoothing assumptions are made about m.

One way to estimate m is to use the method of local polynomial regression (LP method). The idea is based on the fact that the function m can be locally approximated with a Taylor polynomial, i.e. in a neighbourhood around a given point x_0 it holds that

$$m(x) \approx \sum_{k=0}^{p} \frac{m^{(k)}(x_0)}{k!} (x - x_0)^k. \tag{15.2}$$

In order to find an estimate for m at point x_0, one therefore tries to find a polynomial based on observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ that is a good approximation of m around x_0. As a measure of the quality of the approximation one usually chooses a LS criterion, i.e. one wants to minimize the expression

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right\}^2 \tag{15.3}$$

with respect to $\beta = (\beta_0, \ldots, \beta_p)^\top$. Since the representation (15.2) holds only locally, one still has to take into consideration that some of the observations X_i may not lie close enough to x_0 and thus (15.2) no longer applies to them. One must then sufficiently localize the observations, i.e. only consider those observations that lie close enough to x_0.

One of the classical methods for localization is based on weighting the data with the help of a kernel. A kernel is a function $K : \mathbb{R} \longrightarrow [0, \infty)$ with $\int K(u)\, du = 1$. The most useful kernels are symmetric and disappear outside a suitable interval around the zero point.

If K is a kernel and $h > 0$, then the kernel K_h

$$K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$$

is re-scaled with the *bandwidth* h, which again integrates to 1. If, for example, the initial kernel K disappears outside the interval $[-1, 1]$, then K_h is zero outside the interval $[-h, h]$. By weighting the i-th term in (15.3) with $K_h(x - X_i)$, one has a minimization problem which, due to the applied localization, can be formulated independently from the point x_0. The coefficient vector $\hat{\beta} = \hat{\beta}(x) = (\hat{\beta}_0(x), \ldots, \hat{\beta}_p(x))^\top$ that determines the polynomial of the point x is thus given by

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (x - X_i)^j \right\}^2 K_h(x - X_i). \tag{15.4}$$

It is obvious that $\hat{\beta}$ depends heavily on the choice of kernel and the bandwidth. Different methods for determining K and h are introduced in Härdle et al. (2004). With the representation

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 - x & (X_1 - x)^2 & \ldots & (X_1 - x)^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & (X_n - x)^2 & \ldots & (X_n - x)^p \end{pmatrix}, \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

$$\mathbf{W} = \begin{pmatrix} K_h(x - X_1) & & 0 \\ & \ddots & \\ 0 & & K_h(x - X_n) \end{pmatrix}$$

the solution $\hat{\beta}$ to the weighted least squares problem (15.4) can be explicitly written as

$$\hat{\beta}(x) = \left(\mathbf{X}^\top \mathbf{W} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{Y} \tag{15.5}$$

The estimation $\hat{m}(x)$ for $m(x)$ can be obtained only by calculating the approximating polynomial at x:

$$\hat{m}(x) = \hat{\beta}_0(x). \tag{15.6}$$

The remaining components of $\hat{\beta}(x)$, due to Eqs. (15.2) and (15.3) deliver estimators for the derivatives of m: $\hat{m}^{(j)}(x) = j!\,\hat{\beta}_j(x)$, $j = 1, \ldots, p$, which will not be discussed in further detail here. In the special case where $p = 0$, $\hat{m}(x)$ is a kernel estimator of Nadaraya–Watson type, see Härdle (1990).

The similarly derived method of local polynomial approximation, or LP method for short, will now be applied to a time series (Y_i). As mentioned before, one is most interested in creating forecasts.

For the simplest case a one-step ahead forecast means that the functional relationship between Y_{i-1} and a function $\lambda(Y_i)$ of Y_i will be analysed, i.e. we want to obtain an estimate for the unknown function

$$m(x) = \mathsf{E}\big[\lambda(Y_i) \,|\, Y_{i-1} = x\big].$$

In order to apply the LP method mentioned above, consider a given sample $Y_0, \ldots, Y_n$ as observations of the form $(Y_0, Y_1), \ldots, (Y_{n-1}, Y_n)$. The process (Y_i) must fulfil certain conditions, so that these observations are identically distributed and in particular so that the function m is independent of the time index i. Such is the case when (Y_i) is stationary. By substituting $X_i = Y_{i-1}$ into (15.4) and replacing Y_i with $\lambda(Y_i)$, we obtain in this situation

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \left\{ \lambda(Y_i) - \sum_{j=0}^{p} \beta_j (x - Y_{i-1})^j \right\}^2 K_h(x - Y_{i-1}), \tag{15.7}$$

and the estimate for $m(x)$ is again given by $\hat{\beta}_0(x)$.

15.2   Construction of the Estimator

The LP method introduced in the previous section will now be applied under the assumption of a non-parametric autoregressive model of the form (15.1) to estimate the volatility function $s(x)$ of the process (Y_i) based on the observations $Y_0, \ldots, Y_n$.

The conditional volatility $s_i(x)$ and the conditional variance $v_i(x)$, respectively, at time i are defined by

$$v_i(x) = s_i^2(x) = \mathsf{E}[Y_i^2 \,|\, Y_{i-1} = x] - \mathsf{E}^2[Y_i \,|\, Y_{i-1} = x]. \tag{15.8}$$

Included in the assumptions of the model (15.1) is the independence from the time index i. An estimate for $v(x) = s^2(x)$ using the LP method is based on the fact that the two dimensional marginal distribution (Y_{i-1}, Y_i) is independent of i. In the following we will see that (Y_i) approaches a stationary process, with which the following application is justified.

Referring back to the representation (15.8) of the conditional variance $v(x)$ we search for an estimator $\hat{v}_n$ for v with the form

$$\hat{v}_n(x) = \hat{g}_n(x) - \hat{f}_n^2(x), \tag{15.9}$$

i.e. we are looking for an estimator $\hat{g}_n(x)$ for $g(x) = f^2(x) + s^2(x)$ and an estimator $\hat{f}_n(x)$ for $f(x)$.

In order to define these two estimators with the LP method, after applying the steps discussed in the previous section we have to solve both of the following minimization problems:

$$\begin{aligned}
\bar{c}_n(x) &= \arg\min_{c\in\mathbb{R}^{p+1}} \sum_{i=1}^n (Y_i^2 - c^\top U_{in})^2 K\left(\frac{Y_{i-1}-x}{h_n}\right), \\
c_n(x) &= \arg\min_{c\in\mathbb{R}^{p+1}} \sum_{i=1}^n (Y_i - c^\top U_{in})^2 K\left(\frac{Y_{i-1}-x}{h_n}\right).
\end{aligned} \tag{15.10}$$

Here $K : \mathbb{R} \longrightarrow \mathbb{R}$ is a kernel and $\{h_n\}$ a series of positive numbers (bandwidth) with $\lim_{n\to\infty} h_n = 0$. The vectors U_{in} from (15.10) are defined by

$$U_{in} = F(u_{in}), \quad u_{in} = \frac{Y_{i-1} - x}{h_n}. \tag{15.11}$$

with $\mathbb{R}^{p+1}$ valued function $F(u) = \{F_0(u), \dots, F_p(u)\}^\top$ given by

$$F_k(u) = \frac{u^k}{k!}.$$

According to the LP method we define $\hat{g}_n$ and $\hat{f}_n$ with

$$\hat{g}_n(x) = \bar{c}_n(x)^\top F(0) \quad \text{and} \quad \hat{f}_n(x) = c_n(x)^\top F(0),$$

which the above-mentioned application ensures that

$$\hat{v}_n(x) = \bar{c}_n(x)^\top F(0) - \{c_n(x)^\top F(0)\}^2. \tag{15.12}$$

This estimate is a direct modification of the estimator from the local polynomial, non-parametric regression in Tsybakov (1986).

15.3   Empirical Illustration

To illustrate the estimator we consider an example. Figure 15.1 above displays the EUR/USD exchange rate from 1 January 2002 to 31 December 2009 in 20 min intervals (volatility time scale). There are $n = 25,434$ observations. We have calculated the returns of this series (see Fig. 15.1 below) and applied the estimator (15.12) to the time series of the returns. Under the model for the geometric Brownian motion for the price, which is based on the Black–Scholes method (see Sect. 6.1), the returns must follow an ordinary Brownian motion. Their volatilities $v(x)$ are thus constant and independent of x. The estimated conditional variance functions (see Fig. 15.2) show a U shaped structure, which is called a "smiling face" structure or *smile*. The estimated volatility functions $\hat{s}(x) = \sqrt{\hat{v}(x)}$ appear to be qualitatively analogous. This means that the expected risk of the returns is significantly higher when extreme values were observed in the period previously.

As an alternative to Eq. (15.9) it is also possible to first determine the sample residuals

$$\hat{Z}_i = Y_i - \hat{f}_n(Y_{i-1}), \quad i = 1, \ldots, n,.$$

They approximate the true residuals $Z_i = Y_i - f(Y_{i-1})$, which under the assumptions of model (15.1) satisfy

$$E[Z_i|Y_{i-1} = x] = 0, \quad E[Z_i^2|Y_{i-1} = x] = v(x).$$

The volatility can be estimated as in the previous section directly from the non-parametric regression of $\hat{Z}_i^2$ on Y_{i-1}. Fan and Yao (1998) have shown that this
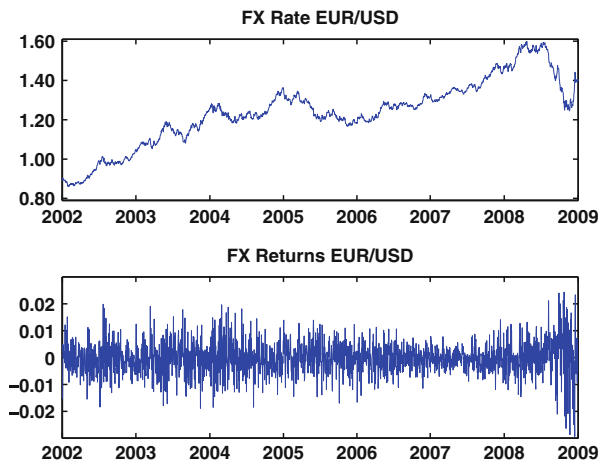
Fig. 15.1 DEM/USD exchange rate and its returns   Q SFEdmusrate
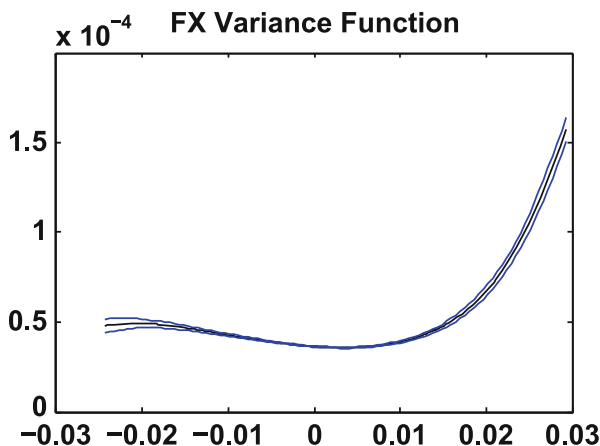
Fig. 15.2 The estimated conditional variance function $\hat{v}(x)$ of the EUR/USD returns

SFEdmusvol

process has advantages in heteroscedastic regression models. In estimating $f(x)$ and $v(x) = s^2(x)$, various bandwidths may be used that do not encounter the danger of the variance estimator taking on negative values. This makes sense when it is expected that the local fluctuations of f and s are of very different proportions.

15.4   Flexible Volatility Estimators

Since their introduction by Engle and Bollerslev, models with autoregressive, conditional heteroscedasticity (*autoregressive conditional heteroscedasticity models* or ARCH) have been successfully applied to financial market data. Thus, it is natural to discuss option pricing models where the underlying instrument follows an ARCH process. From an empirical point of view, the form of the *news impact curve*, which is defined as a function of the current volatility dependent on yesterday's returns, is the dominant factor in determining the price. It is important, for example, to know whether the news impact curve is symmetric or asymmetric. In order to avoid inaccurate pricing due to asymmetries it is necessary to use flexible volatility models. In this way EGARCH models (see Sect. 13.2) can be used when stock prices and volatility are correlated. This model however has a weakness in that the problem of the stationarity conditions and the asymptotic of the Quasi-Maximum-Likelihood-Estimator (QMLE) are not yet completely solved. Another Ansatz, as in the Threshold GARCH-Models, is to introduce thresholds in the news impact curve to create flexible asymmetry.

In this chapter we concentrate on the specification of the volatility. We present the TGARCH process and perform Monte Carlo simulations for three typical parameter groups. In particular, we compare the simulated GARCH option prices with option prices based on the simulations from TGARCH and Black–Scholes models. In the

empirical section of the chapter we show that the market price of call options indeed reflects the asymmetries that were discovered in the news impact curve of the DAX time series.

15.5   Pricing Options with ARCH-Models

Consider an economy in discrete time in which interest and proceeds are paid out at the end of every constant, equally long time interval. Let $S_t, t = 0, 1, 2, \ldots$ be the price of the stock at time t and $Y_t = (S_t - S_{t-1})/S_{t-1}$ the corresponding one period return without dividends. Assume that a price for risk exists in the form of a risk premium which is added to the risk free interest rate r to obtain the expected return of the next period. It seems reasonable to model the risk premium dependent on the conditional variance. As a basis we assume an ARCH-M-Model (see Sect. 13.2.3) with a risk premium, which is a linear function of the conditional standard deviation:

$$Y_t = r + \lambda \sigma_t + \varepsilon_t \tag{15.13}$$

$$\mathcal{L}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \mathrm{N}(0, \sigma_t^2) \tag{15.14}$$

$$\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2. \tag{15.15}$$

In (15.15) ω, α and β are constant parameters that satisfy the stationarity and non-negativity conditions. The constant parameter λ can be understood as the price of one unit of risk. $\mathcal{F}_t$ indicates, as usual, the set of information available up to and including time t. In order to simplify the notation, our discussion will be limited to the GARCH(1,1) case.

The above model is estimated under the empirical measure P. In order to deal with a valuation under no arbitrage, similar to Black–Scholes in continuous time (see Sect. 6.1), assumptions on the valuation of risk must be made. Many studies have researched option pricing with stochastic volatility under the assumption that the volatility has a systematic risk of zero, that is, the risk premium for volatility is zero. Duan (1995) identified an equivalent martingale measure Q for P under the assumption that the conditional distribution of the returns is normal, and in addition it holds that

$$\mathrm{Var}^{\mathrm{P}}(Y_t \mid \mathcal{F}_{t-1}) = \mathrm{Var}^{\mathrm{Q}}(Y_t \mid \mathcal{F}_{t-1}) \tag{15.16}$$

P a.s.. He shows that under this assumption a representative agent with, for example, constant relative risk aversion and a normally distributed relative change of aggregate consumption maximizes his expected utility. The assumption (15.16) contains a constant risk premium for the volatility that directly enters its mean.

In order to obtain a martingale under the new measure a new error term, η_t, needs to be introduced that captures the effect of the time varying risk premium.

When we define $\eta_t = \varepsilon_t + \lambda\sigma_t$, (15.16) it leads to the following model under the new measure Q:

$$Y_t = r + \eta_t \tag{15.17}$$

$$\mathcal{L}_Q(\eta_t \mid \mathcal{F}_{t-1}) = \mathrm{N}(0, \sigma_t^2) \tag{15.18}$$

$$\sigma_t^2 = \omega + \alpha(\eta_{t-1} - \lambda\sigma_{t-1})^2 + \beta\sigma_{t-1}^2. \tag{15.19}$$

In the case of a GARCH(1,1) model according to Theorem 13.10 the variance of the stationary distribution under the empirical measure P is $\mathsf{Var}^P(\varepsilon_t) = \omega/(1-\alpha-\beta)$. For the Duan measure Q the variance of the stationary distribution increases to $\mathsf{Var}^Q(\eta_t) = \omega/\{1 - \alpha(1+\lambda^2) - \beta\}$, because the volatility process under the new measure is determined by the innovations from an asymmetric and a non-symmetric Chi squared distribution. Later on we will see that changes in the unconditional variance depend, in a critical way, on the specification of the news impact curve.

The restriction to a quadratic or symmetric news impact curve is not always optimal, as many empirical studies of stock returns have indicated. Within the context of the above-mentioned model these assumptions can lead to a non-linear news impact function $g(\cdot)$. The following model is a semi-parametric analogue to the GARCH model. Under the empirical measure P we obtain

$$Y_t = r + \lambda\sigma_t + \varepsilon_t$$

$$\mathcal{L}_P(\varepsilon_t \mid \mathcal{F}_{t-1}) = \mathrm{N}(0, \sigma_t^2)$$

$$\sigma_t^2 = g(\varepsilon_{t-1}) + \beta\sigma_{t-1}^2.$$

Under the Duan martingale measure Q the model changes to

$$Y_t = r + \eta_t$$

$$\mathcal{L}_Q(\eta_t \mid \mathcal{F}_{t-1}) = \mathrm{N}(0, \sigma_t^2)$$

$$\sigma_t^2 = g(\eta_{t-1} - \lambda\sigma_{t-1}) + \beta\sigma_{t-1}^2.$$

One notices that as soon as an estimator of $g(\cdot)$ under P is known it can immediately be substituted under the measure Q.

In this general specification the estimation without additional information on $g(\cdot)$ is a difficult matter, since iterative estimation procedures would be necessary in order to estimate the parameters λ, β and the non-parametric function g at the same time. Therefore we will consider a specific, flexible parametric model: the Threshold GARCH Model, see Sect. 13.2. With this model the news impact function can be written as:

$$g(x) = \omega + \alpha_1 x^2 \mathbf{1}(x < 0) + \alpha_2 x^2 \mathbf{1}(x \geq 0)$$
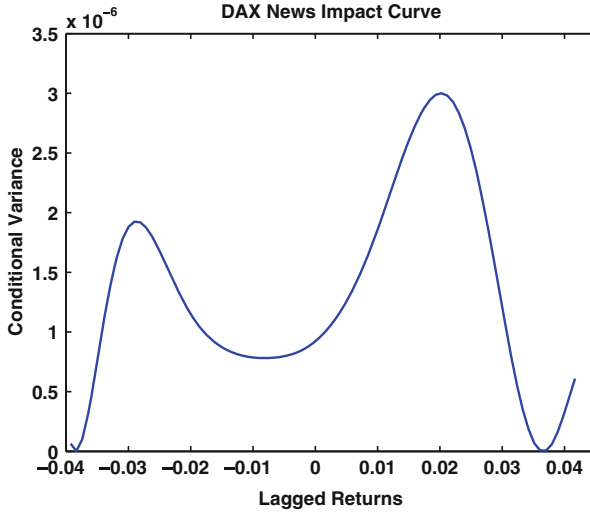
Fig. 15.3 Local linear estimation of the news impact curve for the DAX. The model is $Y_t = \sigma(Y_{t-1})\xi_t$. The estimator of the function $\sigma^2(y)$ with a bandwidth of 0.03 is shown. The tails have been eliminated from the figure   🔲 SFENewsImpactCurve

To motivate this model, consider fitting a very simple non-parametric model $Y_t = \sigma(Y_{t-1})\xi_t$ to the returns of a German stock index, the DAX, where ξ_t is independent and identically distributed with mean 0 and variance 1. The estimator of the news impact curve $\sigma^2(\cdot)$ is given in Fig. 15.3. To get an idea of the underlying distribution of the returns, a non-parametric estimator of the return distribution has been added to Fig. 15.4 over a smoothed normal distribution. Obviously $g(\cdot)$ is not symmetric around zero. The TGARCH model captures this phenomenon when $\alpha_1 > \alpha_2$. Other parametric models can describe these properties as well but the TGARCH model in the case of stock returns has proven to be extremely flexible and technically manageable as claimed, for example, in Rabemananjara and Zakoian (1993).

Remember that the innovations are normally distributed. Thus it follows for the TGARCH model that the unconditional variance, similar to Theorem 13.10, under the measure P is $\mathsf{Var}^P(\varepsilon_t) = \omega/(1 - \bar{\alpha} - \beta)$, where $\bar{\alpha} = (\alpha_1 + \alpha_2)/2$. The following theorem gives the unconditional variance for $\eta_t = \varepsilon_t + \lambda\sigma_t$ under Q.

Theorem 15.1 *The unconditional variance of the TGARCH(1,1) model under the equivalent martingale measure* Q *from Duan is*

$$\mathsf{Var}^Q(\eta_t) = \frac{\omega}{1 - \psi(\lambda)(\alpha_1 - \alpha_2) - \alpha_2(1 + \lambda^2) - \beta} \tag{15.20}$$

where

$$\psi(u) = u\varphi(u) + (1 + u^2)\Phi(u)$$
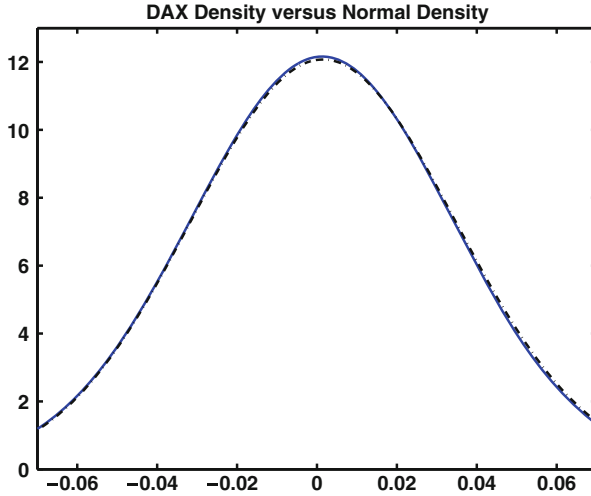
Fig. 15.4 Kernel estimation of the density of DAX returns (*solid line*) against a kernel estimation of a normal distribution (*dotted line*) with the same mean and variance. A bandwidth of 0.03 is used and a quadratic kernel function $K(u) = 15/16(1 - u^2)^2 \mathbf{1}(|u| < 1)$. The tails have been eliminated from the figure   SFEDaxReturnDistribution

and $\varphi(u)$, $\Phi(u)$ *are the density and the distribution function of the standard normal distribution.*

Proof Let $Z_t = \eta_t / \sigma_t - \lambda$. Under measure Q it holds that $\mathcal{L}(Z_t \mid \mathcal{F}_{t-1}) = N(-\lambda, 1)$. The conditional variance σ_t^2 can be written as

$$\sigma_t^2 = \omega + \alpha_1 \sigma_{t-1}^2 Z_{t-1}^2 \mathbf{1}(Z_{t-1} < 0) + \alpha_2 \sigma_{t-1}^2 Z_{t-1}^2 \mathbf{1}(Z_{t-1} \geq 0) + \beta \sigma_{t-1}^2.$$

By calculating the expected value it can be shown that for the integral over the negative values it follows that:

$$\mathsf{E}^Q[Z_t^2 \mathbf{1}(Z_t < 0) \mid \mathcal{F}_{t-1}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} z^2 e^{-\frac{1}{2}(z+\lambda)^2} dz$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\lambda} (u - \lambda)^2 e^{-\frac{1}{2}u^2} du$$

$$= \frac{\lambda}{\sqrt{2\pi}} e^{-\frac{1}{2}\lambda^2} + (1 + \lambda^2)\Phi(\lambda)$$

$$\stackrel{\text{def}}{=} \psi(\lambda). \tag{15.21}$$

Because of

$$\mathsf{E}^{\mathsf{Q}}[Z_t^2 \mid \mathcal{F}_{t-1}] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{1}{2}(z+\lambda)^2} dz = 1 + \lambda^2$$

it follows for the positive values that

$$\mathsf{E}^{\mathsf{Q}}[Z_t^2 \mathbf{1}(Z_t \geq 0) \mid \mathcal{F}_{t-1}] = 1 + \lambda^2 - \psi(\lambda). \tag{15.22}$$

Thus we obtain

$$\mathsf{E}^{\mathsf{Q}}[\sigma_t^2] = \omega + \alpha_1 \psi(\lambda) \mathsf{E}^{\mathsf{Q}}[\sigma_{t-1}^2] + \alpha_2[1 + \lambda^2 - \psi(\lambda)] \mathsf{E}^{\mathsf{Q}}[\sigma_{t-1}^2] + \beta \mathsf{E}^{\mathsf{Q}}[\sigma_{t-1}^2]. \tag{15.23}$$

Since the unconditional variance is independent of t, the theorem follows.

The function ψ is positive and $\psi(\lambda) > 1/2$ for the realistic case $\lambda > 0$. We can make the following statement about the changes in the unconditional variance: for $\alpha_1 = \alpha_2$ in (15.20), one obtains the GARCH(1,1) results. For $\alpha_1 > \alpha_2$ (the case of the leverage effect) the increase in the unconditional variance is even stronger than the symmetric GARCH case. For $\alpha_1 < \alpha_2$, the unconditional variance is smaller as in the leverage case, and we can distinguish between two cases: when the inequality

$$\alpha_1 < \alpha_2 \frac{2\psi(\lambda) - 1 - 2\lambda^2}{2\psi(\lambda) - 1} \tag{15.24}$$

is fulfilled then the unconditional variance under Q is actually smaller than under P. If (15.24) is not fulfilled, then we obtain, as above, $\mathsf{Var}^{\mathsf{P}}(\varepsilon_t) \leq \mathsf{Var}^{\mathsf{Q}}(\eta_t)$. Indeed the quotient on the right-hand side of (15.24) takes on negative values for realistic values of a unit of the risk premium, (for example for small positive values), so that in most empirical studies (15.24) cannot be fulfilled.

Naturally the stationary variance has an effect on an option's price; the larger (smaller) the variance is, the higher (lower) the option price is. This holds true in particular for options with a longer time to maturity where the long-run average of the volatility is the most important determinant of the option's price. Therefore, an option can be undervalued when a GARCH model is used and at the same time a leverage effect is present.

A second feature of the Duan approach is that under Q and with positive risk premia, the current innovation is negatively correlated with the next period's conditional variance of the GARCH risk premium, whereas under P the correlation is zero. More precisely, we obtain $\mathsf{Cov}^{\mathsf{Q}}(\eta_t/\sigma_t, \sigma_{t+1}^2) = -2\lambda\alpha\mathsf{Var}^{\mathsf{Q}}(\eta_t)$ with the GARCH parameter α. It is obvious that small forecasts of the volatility under Q (which influences the option's price) depend not only on the past squared innovations, but also on their sign. In particular a negative (positive) past innovation for $\lambda > 0$ leads to the fact that the volatility increases(falls) and with it, the

option price. The following theorem claims that the covariance is dependent on the asymmetry of the news impact function when a TGARCH instead of a GARCH model is used.

Theorem 15.2 *For the TGARCH(1,1) model the covariance between the innovation in t and the conditional variance in t + 1 under the equivalent martingale measure* Q *from Duan is given by*

$$Cov^Q(\frac{\eta_t}{\sigma_t}, \sigma_{t+1}^2) = -2\mathsf{Var}^Q(\eta_t)[\lambda\alpha_2 + \{\varphi(\lambda) + \lambda\Phi(\lambda)\}(\alpha_1 - \alpha_2)], \qquad (15.25)$$

where $\mathsf{Var}^Q(\eta_t)$ *follows from the previous theorem.*

Proof First the conditional covariance is determined:

$$\mathsf{Cov}_{t-1}^Q(\frac{\eta_t}{\sigma_t}, \sigma_{t+1}^2) = \mathsf{E}_{t-1}^Q\left[\frac{\eta_t}{\sigma_t}\sigma_{t+1}^2\right] = \omega\mathsf{E}_{t-1}^Q\left[\frac{\eta_t}{\sigma_t}\right]$$

$$+ \alpha_1\mathsf{E}_{t-1}^Q\left[\frac{\eta_t}{\sigma_t}(\eta_t - \lambda\sigma_t)^2\mathbf{1}(\eta_t - \lambda\sigma_t < 0)\right]$$

$$+ \alpha_2\mathsf{E}_{t-1}^Q\left[\frac{\eta_t}{\sigma_t}(\eta_t - \lambda\sigma_t)^2\mathbf{1}(\eta_t - \lambda\sigma_t \geq 0)\right]$$

$$+ \beta\sigma_t\mathsf{E}_{t-1}^Q[\eta_t], \qquad (15.26)$$

where $\mathsf{E}_t(\cdot)$ and $\mathsf{Cov}_t(\cdot)$ are abbreviations of $\mathsf{E}(\cdot \mid \mathcal{F}_t)$ and $\mathsf{Cov}(\cdot \mid \mathcal{F}_t)$, respectively. Due to (15.18) the first and the fourth expectation values on the right side of (15.26) are zero. The second conditional expected value is

$$\mathsf{E}_{t-1}^Q\left[\frac{\eta_t}{\sigma_t}(\eta_t - \lambda\sigma_t)^2\mathbf{1}(\eta_t - \lambda\sigma_t < 0)\right]$$

$$= -2\sigma_t^2\left[\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\lambda^2) + \lambda\Phi(\lambda)\right]. \qquad (15.27)$$

Since $\mathsf{E}_{t-1}^Q\left[\frac{\eta_t}{\sigma_t}(\eta_t - \lambda\sigma_t)^2\right] = -2\lambda\sigma_t^2$, we can write for the third conditional expected value in (15.26):

$$\mathsf{E}_{t-1}^Q\left[\frac{\eta_t}{\sigma_t}(\eta_t - \lambda\sigma_t)^2\mathbf{1}(\eta_t - \lambda\sigma_t \geq 0)\right]$$

$$= -2\sigma_t^2\left[\lambda - \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}\lambda^2) - \lambda\Phi(\lambda)\right]. \qquad (15.28)$$

Inserting (15.27) and (15.28) into (15.26), it follows that

$$\mathsf{Cov}_{t-1}^Q(\frac{\eta_t}{\sigma_t}, \sigma_{t+1}^2) = -2\sigma_t^2[\lambda\alpha_2 + \{\varphi(\lambda) + \lambda\Phi(\lambda)\}(\alpha_1 - \alpha_2)]. \qquad (15.29)$$

One notices that $\mathsf{Cov}^Q(\eta_t/\sigma_t, \sigma_{t+1}^2) = \mathsf{E}^Q[\mathsf{Cov}_{t-1}^Q(\eta_t/\sigma_t, \sigma_{t+1}^2)]$, thus the claim follows immediately.

In the following we assume that a positive risk premium λ exists per unit. Three cases can be identified: for $\alpha_1 = \alpha_2$ (in the symmetric case) we obtain $\mathsf{Cov}^Q(\eta_t/\sigma_t, \sigma_{t+1}^2) = -2\lambda\alpha_2\mathsf{Var}^Q(\eta_t)$, i.e. the GARCH(1,1) result. For $\alpha_1 < \alpha_2$ (the case of the reverse leverage effect) the covariance increases,

$$\lambda\alpha_2 + \left\{ \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\lambda^2\right) + \lambda\Phi(\lambda) \right\} (\alpha_1 - \alpha_2) < 0, \qquad (15.30)$$

the correlation is positive. In the last case, $\alpha_1 > \alpha_2$ (the leverage case), the covariance is negative and increases with the total.

This also shows that the return of the volatility to a stationary variance under Q is different from the symmetric GARCH case. The negative covariance in the leverage case is actually larger. This could indicate that options are over (under) valued when for positive (negative) past innovation a TGARCH process with $\alpha_1 > \alpha_2$ is used for the price process and then mistakenly a GARCH model ($\alpha_1 = \alpha_2$) is used for the volatility forecast.

15.6 Application to the Valuation of DAX Calls

The valuation method with GARCH is applied to the German stock index and options data. For the stock index we use the daily closing values of the DAX from 1 January 1988 to 31 March 1992. The closing values are usually set at 13:30 (Frankfurt time). For the options data on this index we have taken the recorded values of the transaction prices from the German derivative exchange (DTB) from January to March 1992. In order to synchronize the observation time periods of the index and options we interpolate between the last option price before 13:30 h and the first price after, as long as the difference is no more than 2 h.

No evidence for autocorrelated DAX returns was found but the squared and absolute returns are highly autocorrelated. We estimate a GARCH(1,1)-M model

$$Y_t = \lambda\sigma_t + \varepsilon_t \qquad (15.31)$$

$$\mathcal{L}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \mathrm{N}(0, \sigma_t^2) \qquad (15.32)$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2 \qquad (15.33)$$

for the DAX with the Quasi-Maximum-Likelihood-Method; see Sect. 13.1.6. A possible constant in (15.31) is not significant and is thus ignored from the very beginning. Table 15.1 shows the results of the estimation. All parameters are significantly different from zero. The degree of persistence $\alpha + \beta = 0.9194$ is significantly smaller than 1 and thus the unconditional variance is finite, see

Table 15.1 The GARCH and TGARCH estimation results for
DAX returns, 1 January 1988 to 30 December 1991 (QMLE
standard error in parentheses)

	GARCH	TGARCH
ω	1.66E−05 (1.04E−06)	1.91E−05 (1.359E−06)
α	0.144 (0.006)	
α_1		0.201 (0.008)
α_2		0.045 (0.011)
β	0.776 (0.012)	0.774 (0.016)
λ	0.069 (0.018)	0.039 (0.018)
$-2logL$	−7698	−7719

Theorem 13.10. The parameter of the risk premium λ is positive as is expected
from economic theory.

The Quasi-Maximum-Likelihood-Estimator of the TGARCH model

$$\sigma_t^2 = \omega + \alpha_1\varepsilon_{t-1}^2\mathbf{1}(\varepsilon_{t-1} < 0) + \alpha_2\varepsilon_{t-1}^2\mathbf{1}(\varepsilon_{t-1} \geq 0) + \beta\sigma_{t-1}^2 \tag{15.34}$$

is also given in Table 15.1. Taking the value of the log-likelihood into consideration,
the ability of the TGARCH model is better than that of GARCH model. A likelihood
quotient test rejects the GARCH model at every rational confidence level. α_1 and
α_2 are significantly different; thus the asymmetry of the news impact function is
significant. Since $\alpha_1 > \alpha_2$ we observe the usual leverage effect for financial time
series.

After the model was fitted to the data from 1988 to 1991, the next step in
calculating the option price for the observed time period from January to March
1992 is to use the simulation method described above and then compare this to the
market prices. Here we will concentrate on call options. Since the DAX option,
which is traded on the DTB, is a European option, the results for put options can
be calculated as usual from the put–call-parity. We consider nine call options with
maturity dates 17 January, 20 March and 19 June 1992. In order to distinguish the
case of in, out and at the money, we have chosen the strike prices 1,550, 1,600 and
1,650 for the January option and 1,600, 1,650 and 1,700 for the March and June
options. We simulate the price of the January option from 3 January to the 16th (10
days), for the March option from 3 January to the 19th (57 days) and for the June
option from 3 January to the 31 March (64 days). The June option with a strike
price of 1,700 began on 16 January so that there are no observations for the first 10
trading days. Due to low trading volume not all market prices are available, thus we
reduced the number of observations, k in Table 15.2, even further.

A remaining question is how to choose the starting value of the volatility process.
We set the starting value equal to the running estimator of the volatility (GARCH
or TGARCH), in which the volatility process is extrapolated and the parameters are
held constant. Alternatively one can use the implied volatility, see Sect. 6.4.5.

To calculate the Black–Scholes price at time t the implied volatility at time $t - 1$ is used. To obtain a measure of the quality of the estimate, we define the relative residuals as

$$u_t \overset{\text{def}}{=} \frac{C_t - C_{Market,t}}{C_{Market,t}}$$

where C_t is either the Black–Scholes or the GARCH or the TGARCH price and $C_{Market,t}$ is the price observed on the market. Residuals should be considered as relative values, since a trader would always prefer the cheaper option, which is undervalued by the same amount than a more expensive option, simply because he can multiply his position in the cheaper option. A similar argument holds for the sale of an overvalued option. For reasons of symmetry we use a squared loss criterion, i.e.

$$U = \sum_{t=1}^{k} u_t^2.$$

The results for the three models are given in Table 15.2.

Overall the GARCH as well as the TGARCH option valuation model performs substantially better than the Black–Scholes model. For options in and at the money the improvement of the TGARCH forecast compared to the GARCH model is small. When the option, however, is out of the money there is a large reduction in the loss criterion. In the simulation study out of the money options react most sensitively to stochastic volatility and the leverage effect. In situations with real data this is most obvious for the January-1,650 option, where Black–Scholes performs poorly and TGARCH performs better than GARCH. For the March and June options

Table 15.2 The loss criterium U for the DAX calls with maturity at T and a strike price K using BS, GARCH and TGARCH option prices

T	K	k	BS	GARCH	TGARCH
	1,550	10	0.017	0.014	0.014
January	1,600	10	0.099	0.029	0.028
	1,650	10	4.231	1.626	1.314
	1,600	47	1.112	0.961	0.954
March	1,650	53	1.347	1.283	1.173
	1,700	56	1.827	1.701	1.649
	1,600	53	1.385	1.381	1.373
June	1,650	56	2.023	1.678	1.562
	1,700	51	2.460	2.053	1.913
Sum		346	14.500	10.725	9.980

The number of observations is given by k

the difference is not so obvious. This can be explained by the fact that the index increased to a level of 1,736 points on 20 March 1,717 points on 30 March, so that the option with a strike price of 1,700 became in the money. This is also the explanation for the fact that U is the highest for the January-1,650 option. There were only 10 trading days, but the option was out of the money for several days. For example, the DAX closed on 8 January at 1,578 points. Since in every case TGARCH performs better than GARCH, we conclude that the market follows the asymmetry of the volatility. Therefore, specifying the volatility model correctly plays an important role in determining option prices.

15.7 Recommended Literature

The model (15.1) is thoroughly studied together with financial time series, in particular under the assumptions of the ARCH structure, in Engle (1982). Until recently, academic research focused mainly on the (linear) conditional mean, or it was assumed that the conditional variance was constant or, as in the ARCH models, that it had a special form. At the beginning of the 1980s this deficit in the literature was corrected by Engle (1982), and Robinson (1983, 1984) and in the statistic literature by Collomb (1984) and Vieu (1995). There have also been non-parametric and semi-parametric approximations suggested in Gregory (1989) and in Engle and Gonzalez-Rivera (1991). Since then the interest in the non-parametric situation discussed here, in which the form of the functions f and s is not identified ahead of time, has clearly grown in economics and statistics literature, see Fan and Yao (2003).

The QTARCH models in Gouriéroux and Monfort (1992) create a generalization of the threshold models for the conditional mean in Tong (1983). The methods from Gouriéroux and Monfort (1992) and McKeague and Zhang (1994) are based on histogram estimations of the volatility. The works from Chen and Tsay (1993a,b) concentrate on additive modelling of the mean function f. Additive or multiplicative structures of volatility are considered in Härdle et al. (1997), Yang et al. (1999) and Hafner (1998). The general non-parametric ARCH model is handled in Härdle et al. (1996). Franke (1999) discusses the connection between the non-parametric AR-ARCH model and the discrete version of geometric Brownian motion which is used as a foundation for the Black–Scholes applications. Franke, Härdle and Kreiss (2003) study, in connection with a special stochastic volatility model, a non-parametric de-convolution estimator for the volatility function as the first step towards the non-parametric handling of general GARCH models.

The idea of the local polynomial estimation originates in Stone (1977), Cleveland (1979) and Katkovnik (1979, 1985), who have all used it on non-parametric regression models. Statistical properties of LP estimators by nonparametric regression models (convergence, convergence rate and pointwise asymptotic normality) are derived in Tsybakov (1986). References to more recent studies in this area can be found in Fan and Gijbels (1996).

Apart from the statistical studies of the model (15.1), the utilized theoretical probability properties of the constructed process (Y_i) are also of importance. This is studied in the works of Doukhan and Ghindès (1981), Chan and Tong (1985), Mokkadem (1987), Diebolt and Guégan (1990) and Ango Nze (1992). In these articles the ergodicity, geometric ergodicity and mixture properties of the process (Y_i) are derived.

The presentation of this chapter closely follows the work of Härdle and Hafner (2000). The standard ARCH model originated in Engle (1982), the development of EGARCH in Nelson (1991) and TGARCH in Zakoian (1994) (for the standard deviation) and Glosten et al. (1993) (for the variance). Non-parametric and semi-parametric variants of the ARCH model were suggested and studied by Carroll et al. (2002) and Hafner (1998). The classic option pricing model with stochastic volatility originated in Hull and White (1987). Hull and White implicitly assume that the market price of the risk of the volatility is zero, whereas in Melino and Turnbull (1990) it is different from zero, constant and exogenous. Empirical evidence for the valuation of risk of the volatility is given in Wiggins (1987). Renault and Touzi (1996) generalize the model from Hull and White (1987), in that they allow a market price of the risk for the volatility, which itself can vary over time. The concept of minimizing the quadratic loss of a hedge portfolio is given in Föllmer and Sondermann (1991) and Föllmer and Schweizer (1991). The practical procedure to implement "15 min old" implied volatility into the Black/Scholes formula, was successfully used in Bossaerts and Hillion (1993).

Part III
Selected Financial Applications

Chapter 16
Value-at-Risk and Backtesting

The Value-at-Risk (VaR) is probably the most known measure for quantifying and controlling the risk of a portfolio. The establishment of VaR is of central importance to a credit institute, since it is the basis for a regulatory notification technique and for required equity investments. The description of risk is carried out with the help of an "internal model", whose job is to reflect the market risk of portfolios and similar risky investments over time. This often occurs through the choice of suitable portfolios with a specific risk factor, i.e. through principal component analysis (Chap. 20). With risks from option trading, a linear transformation is often applied using the "Greeks" (Chap. 6).

The objective parameter in the model is the probability forecasts of portfolio changes over a given time horizon. Whether the model and its technical application correctly identify the essential aspects of the risk, remains to be checked. The back testing procedure serves to evaluate the quality of the forecast of a risk model in that it compares the actual results to those generated with the VaR model. For this the daily VaR estimates are compared to the results from hypothetical trading that are held from the end-of-day position to the end of the next day, the so-called clean back testing. The concept of clean back testing is differentiated from that of "mark-to-market" profit and loss ("dirty $P\&L$") analyses in which intra-day changes are also observed. In judging the quality of the forecast of a risk model it is advisable to concentrate on the clean back testing.

The interest of an institute in a correct VaR calculation can be traced back to a rule of equity investing, which we will briefly describe here. Since 1997 (modification of the Basle market risk paper) institutes have been allowed to replicate specific risks using internal models. Included here under specific risks are those associated with daily price developments ("residual risks") and others realized from rare occurrences ("event risks" such as rating changes). Models that only consider residual risks are called "surcharge models", and those that consider event risks are called "non-surcharge models".To calculate capital investments institutes need to use the following formula, Graumert and Stahl (2001):

Table 16.1 Traffic light as a factor of the exceeding amount

Exceedances	Increase of M	Zone
0 bis 4	0	**Green**
5	0.4	**Yellow**
6	0.5	**Yellow**
7	0.65	**Yellow**
8	0.75	**Yellow**
9	0.85	**Yellow**
More than 9	1	**Red**

$$EMU_t = \max \left(VaR_{t-1} + d \cdot SR_{t-1}; M \cdot \frac{1}{60} \sum_{i=1}^{60} VaR_{t-i} + d \cdot \frac{1}{60} \sum_{i=1}^{60} SR_{t-i} \right)$$
(16.1)

- EMU_t = Capital investment for the price risks determined by the risk model at day t
- VaR_{t-i} = VaR estimation at day $t - i$ for the general and the specific price risk
- d = Indicator variable with $d = 1$ for surcharge models and $d = 0$ for non-surcharge models and for models that only model general risk
- M = Multiplier with $M = 3 + ZBT + ZQM$
- ZBT = Backtesting surcharge factor according to § 37 *II* GI ($0 \leq ZBT \leq 1$)
- ZQM = Surcharge factor for qualitative deficiencies ($0 \leq ZQM \leq 1$)
- SR_{t-i} = Surcharge for not modelling event risk in surcharge models on day $t - i$

The multiplier M in (16.1) contains the backtesting surcharge factor which is calculated from the so-called traffic light. According to the "traffic light" the value M increases with the number of times the VaR values exceed the actual loss. Table 16.1 explains the "traffic light" zones.

16.1 Forecast and VaR Models

Value-at-Risk (VaR) models are used in many financial applications. Their goal is to quantify the profit or loss of a portfolio which could occur in the near future. The uncertainty of the development of a portfolio is expressed in a "forecast distribution" P_{t+1} for period $t + 1$.

$$P_{t+1} = \mathcal{L}(L_{t+1}|\mathcal{F}_t)$$

is the conditional distribution of the random variable L_{t+1}, which represents the possible profits and losses of a portfolio in the following periods up to date $t + 1$, and $\mathcal{F}_t$ stands for the information in the available historical data up to date t. An estimator for this distribution is given by the forecast model. Consequently the possible conditional distributions of L_{t+1} come from a parameter class $\mathcal{P}_{t+1} = \{P_{t+1}^{\theta(t)} \,|\, \theta(t) \in \Theta\}$. The finite-dimensional parameter $\theta(t)$ is typically estimated from $n = 250$ historical return observations at time t, that is approximately the trading days in a year. Letting $\hat{\theta}(t)$ stand for this estimator then $\mathcal{L}(L_{t+1}|\mathcal{F}_t)$ can be approximated with $P_{t+1}^{\hat{\theta}(t)}$.

An important example of $\mathcal{P}_{t+1}$ is the Delta-Normal Model, RiskMetrics (1996). In this model we assume that the portfolio is made up of d linear (or linearized) instruments with market values $X_{k,t}, k = 1, \ldots, d$, and that the combined conditional distribution of the log returns of the underlying

$$Y_{t+1} \in \mathbb{R}^d , Y_{k,t+1} = \log X_{k,t+1} - \log X_{k,t}, k = 1, \ldots, d,$$

given the information up to time t is a multivariate normal distribution, i.e.

$$\mathcal{L}(Y_{t+1}|\mathcal{F}_t) = \mathrm{N}_d(0, \Sigma_t) \tag{16.2}$$

where Σ_t is the (conditional) covariance matrix of the random vector Y_{t+1}. We consider first a single position ($d = 1$), which is made up of λ_t shares of a single security with an actual market price $X_t = x$. With $w_t = \lambda_t x$ we represent the *exposure* of this position at time t, that is its value given $X_t = x$. The conditional distribution of the changes to the security's value $L_{t+1} = \lambda_t(X_{t+1} - X_t)$ is approximately:

$$\begin{aligned}
\mathcal{L}(L_{t+1}|\mathcal{F}_t) &= \mathcal{L}(\lambda_t(X_{t+1} - x) \,|\, \mathcal{F}_t) \\
&= \mathcal{L}(w_t \frac{X_{t+1} - x}{x} \,|\, \mathcal{F}_t) \\
&\approx \mathcal{L}(w_t Y_{t+1} \,|\, \mathcal{F}_t) = \mathrm{N}(0, w_t^2 \sigma_t^2) \tag{16.3}
\end{aligned}$$

with $\sigma_t^2 = \mathsf{Var}(Y_{t+1} \,|\, \mathcal{F}_t)$. Here we have used the Taylor approximation

$$\log X_{t+1} - \log x = \frac{X_{t+1} - x}{x} + \mathcal{O}(X_{t+1} - x). \tag{16.4}$$

The generalization to a portfolio that is made up of $\lambda_t^1, \cdots, \lambda_t^d$ shares of d (linear) instruments is quite obvious. Let w_t be the d-dimensional exposure vector at time t

$$w_t = (w_t^1, \cdots, w_t^d)^\top = (\lambda_t^1 x^1, \cdots, \lambda_t^d x^d)^\top. \tag{16.5}$$

$$L_{t+1} = \sum_{k=1}^{d} \lambda_t^k (X_{k,t+1} - X_{k,t})$$

is the change in the value of the portfolio. For a single position the conditional distribution of L_{t+1} given the information $\mathcal{F}_t$ is approximately equal to the conditional distribution of

$$w_t^\top Y_{t+1} = \sum_{k=1}^{d} w_t^k Y_{k,t+1}.$$

In the framework of Delta-Normal models this distribution belongs to the family

$$\mathcal{P}_{t+1} = \{\mathrm{N}(0, \sigma_t^2) : \sigma_t^2 \in [0, \infty)\}, \tag{16.6}$$

with $\sigma_t^2 = w_t^\top \Sigma_t w_t$. The goal of the VaR analysis is to approximate the parameter $\theta(t) = \sigma_t$ and thus to approximate the forecast distribution of P_{t+1}.

Now consider the problem of estimating the forecast distribution from the view point of the following model's assumptions. The change in the value of the portfolio is assumed to be of the form

$$L_{t+1} = \sigma_t \, Z_{t+1} \tag{16.7}$$

$$\sigma_t^2 = w_t^\top \Sigma_t w_t, \tag{16.8}$$

where Z_t is i.i.d. N(0,1) distributed random variable, w_t is the exposure vector at time t and Σ_t is the (conditional) covariance matrix of the vector Y_{t+1} of the log returns. We combine the last n realizations of $Y_t = y_t, \ldots, Y_{t-n+1} = y_{t-n+1}$ from the log return vector with a $(n \times d)$ matrix $\mathcal{Y}_t = (y_i^\top)_{i=t-n+1,\ldots,t}$. From these observations we calculate two estimators from Σ_t; first the naive RMA, i.e. *rectangular moving average*:

$$\hat{\Sigma}_t = \frac{1}{n} \mathcal{Y}_t^\top \mathcal{Y}_t. \tag{16.9}$$

Since the expected value of the vector of returns Y_t is zero according to the Delta-Normal model, this is exactly the empirical covariance matrix. The second so-called EMA estimator, i.e. *exponentially moving average*, is based on an idea from Taylor (1986) and uses an exponential weighting scheme. Define for γ, $0 < \gamma < 1$

$$\tilde{y}_{t-k} = \gamma^k y_{t-k}, k = 0, \ldots, n-1, \quad \tilde{\mathcal{Y}}_t = (\tilde{y}_i^\top)_{i=t-n+1,\ldots,t}$$

a log return vector is exponentially weighted over time and a $(n \times d)$ matrix is constructed from this, then Σ_t is estimated with

$$\hat{\Sigma}_t = (1 - \gamma)^{-1} \tilde{\mathcal{Y}}_t^\top \tilde{\mathcal{Y}}_t. \tag{16.10}$$

This normalization makes sense, since the sum $\sum_{i=1}^{n} \gamma^{i-1} = \frac{1-\gamma^n}{1-\gamma}$ for $\gamma \to 1$ converges to n, thus the RMA estimator is the boundary case of the EMA estimator. Both estimators can be substituted in (16.7) and (16.8), and we obtain with

$$\hat{P}_{t+1} = N(0, \hat{\sigma}_t^2), \quad \hat{\sigma}_t^2 = w_t^\top \hat{\Sigma}_t w_t$$

an approximation of the forecast distribution, i.e. the conditional distribution of L_{t+1}. It should be noted that the Bundesanstalt für Finanzdienstleistungsaufsicht (http://www.bafin.de) currently proposes the RMA technique.

The *Value-at-Risk* VaR is determined for a given level α by

$$VaR_t = F_{t+1}^{-1}(\alpha) \stackrel{\text{def}}{=} \inf\{x; F_{t+1}(x) \geq \alpha\} \tag{16.11}$$

and estimated with

$$\widehat{VaR}_t = \hat{F}_{t+1}^{-1}(\alpha) \stackrel{\text{def}}{=} \inf\{x; \hat{F}_{t+1}(x) \geq \alpha\}. \tag{16.12}$$

Here F_{t+1}, $\hat{F}_{t+1}$ represent the distribution function of P_{t+1}, $\hat{P}_{t+1}$. The quality of the forecast is of particular interest in judging the VaR technique. It can be empirically checked using the realized values $(\hat{P}_t, L_t), t = 1, \ldots, N$. In the event that the model assumptions, for example, (16.7) and (16.8), are correct for the form of the forecast's distribution, then the sample $U_t = F_t(L_t), t = 1, \ldots, N$, should have independent uniformly distributed random values over the interval $[0, 1]$ and $\hat{U}_t = \hat{F}_t(L_t), t = 1, \ldots, N$, approximately independent identically uniformly distributed random values. Then the ability of the forecasts distribution to fit the data is satisfied.

16.2 Backtesting with Expected Shortfall

Below we consider the expected shortfall from the profit and loss variable L_{t+1} as an alternative to the VaR and develop a backtesting method for this risk measurement. The *expected shortfall*, also called the *Tail-VaR*, is in the Delta-Normal Model, i.e. under the assumptions from (16.7) and (16.8), defined by

$$E(L_{t+1} \mid L_{t+1} > VaR_t) = E(L_{t+1} \mid L_{t+1} > z_\alpha \sigma_t)$$
$$= \sigma_t \, E(L_{t+1}/\sigma_t \mid L_{t+1}/\sigma_t > z_\alpha). \tag{16.13}$$

Here $z_\alpha = \Phi^{-1}(\alpha)$ represents the α quantile of the standard normal distribution, where Φ is the standard normal distribution function.

Under this model (16.7) and (16.8) $Z_{t+1} = L_{t+1}/\sigma_t$ has a standard normal distribution. For a defined threshold value u we obtain

$$\vartheta = E(Z_{t+1} \mid Z_{t+1} > u) = \frac{\varphi(u)}{1 - \Phi(u)} \tag{16.14}$$

$$\varsigma^2 = \mathsf{Var}(Z_{t+1} \mid Z_{t+1} > u) = 1 + u \cdot \vartheta - \vartheta^2, \qquad (16.15)$$

where φ is the standard normal density. For given observations from a forecast distribution and its realizations $(\hat{F}_{t+1}(\cdot/\hat{\sigma}_t), L_{t+1}/\hat{\sigma}_t)$ we consider (16.14) as the parameter of interest. Replacing the expected value with a sample mean and the unobservable Z_{t+1} with

$$\hat{Z}_{t+1} = \frac{L_{t+1}}{\hat{\sigma}_t}, \qquad (16.16)$$

where σ_t in (16.8) is estimated with (16.9) or (16.10), we obtain an estimator for ϑ

$$\hat{\vartheta} = \frac{1}{N(u)} \sum_{t=0}^{n} \hat{Z}_{t+1} \, \mathbf{1}(\hat{Z}_{t+1} > u). \qquad (16.17)$$

$N(u)$ is the random number of times that the threshold value u is exceeded:

$$N(u) = \sum_{t=1}^{n} \mathbf{1}(\hat{Z}_{t+1} > u).$$

Inferencing on the expected shortfall, i.e. on the difference $\hat{\vartheta} - \vartheta$, we obtain the following asymptotical result:

$$\sqrt{N(u)} \left(\frac{\hat{\vartheta} - \vartheta}{\hat{\varsigma}} \right) \xrightarrow{\mathcal{L}} \mathrm{N}(0, 1) \qquad (16.18)$$

(16.18) can be used to check the adequacy of the Delta-Normal model.

16.3   Backtesting in Action

The data used in this section is a bond portfolio of a German bank from 1994 to 1995. The portfolio is not adjusted so that the exposure vector $w_t = w$ is time dependent. We assume that (16.7) and (16.8) hold. The VaR forecast is based on both prediction rules introduced in Sect. 16.1 that are used to estimate the parameters σ_t of the forecast distribution in RMA and EMA given $\gamma = 0.94$. In light of the bond crisis in 1994 it is interesting how both techniques respond to this stress factor.

The significance level under consideration is $\alpha = 1\,\%$ for large losses and $\alpha = 99\,\%$ for large profits. To investigate we include plots of time series from the realized P/L (i.e. profit-loss) data L_t as compared to the respective VaR estimator $\widehat{VaR}_t$ calculated with (16.12). If the model and the estimation of the parameter σ_t based on forecast distribution are adequate, then approximately $1\,\%$ of the data
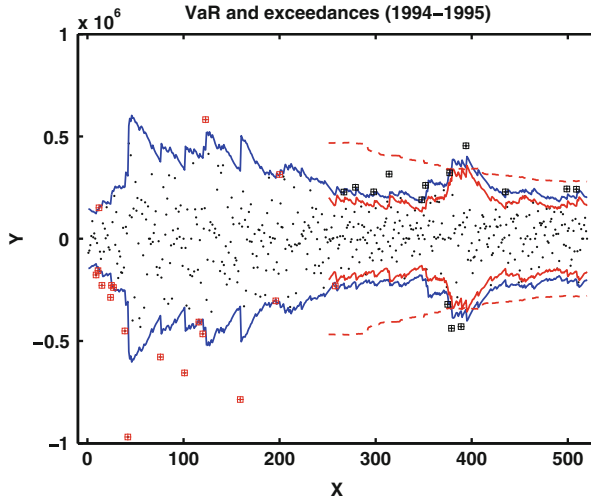
Fig. 16.1 The *dots* show the observed changes L_t in the value of the portfolio. The *dashed line* represents the forecasted VaRs based on RMA (99 and 1 %). The *solid red line* represents the same for EMA. The *solid blue line* shows the estimation results of the Delta-Normal model

Q SFEVaRbank

should lie below the 1 % and above the 99 % VaR Estimators. In addition, in Fig. 16.1 the crossings for the case where VaR is estimated with EMA are marked. We recognize that in 1994 (1995) there were a total of 10 (9) crossings determined for the EMA method. This strongly contrasts the 17 (3) observed values for the RMA method. It is clear that the RMA technique leads to, above all during the bond crisis in 1994, too many crossings for the 1 % VaR estimator, which means that the probability of larger losses is underestimated. This tendency to underestimate the risk is produced from the observation width of 250 days, when the market is moving towards a more volatile phase. The opposite is true when moving in the other direction; RMA overestimates risk. The EMA adapts more quickly to market phases since data in the past has less influence on the estimator due to the exponentially deteriorating weights.

The poor forecast quality of the RMA, in particular for the left side of the distribution, can also be seen in that for a particular day the VaR was exceeded by 400 %. If the model (16.7) and (16.8) is correct, then the variable (16.19) must have a standard deviation of about 0.41. The empirical standard deviation calculated from the data is about 0.62. According to the volatility scale of the RMA the risk is underestimated on average by $\frac{0.62-0.41}{0.41} \approx 50$ %. The EMA plot in Fig. 16.1 shows a better calibration. The empirical standard deviation of (16.19) is in this case around 0.5, which corresponds to an underestimation of risk by approximately 25 %.

All other diagnostic measurements are entered into the QQ plot of the variable
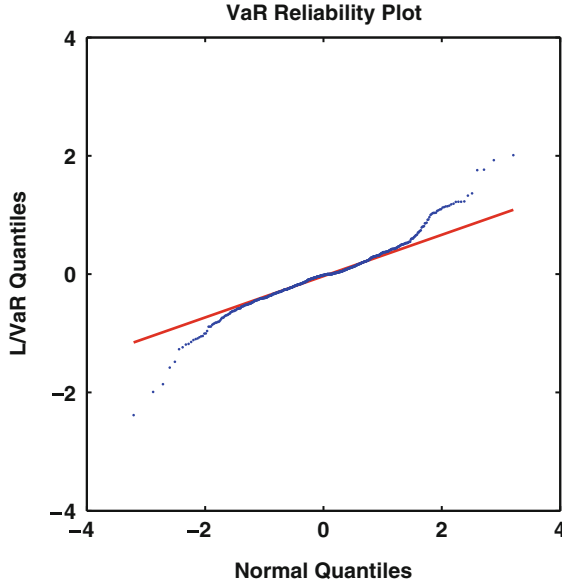
Fig. 16.2 QQ plot of $L_{t+1}/\widehat{VaR}_t$ for RMA in 1994   Q SFEVaRqqplot

$$\frac{L_{t+1}}{\widehat{VaR}_t} = \frac{L_{t+1}}{2.33\hat{\sigma}_t}, \qquad (16.19)$$

see Figs. 16.2 and 16.3. If the VaR forecast $\widehat{VaR}_t$ was perfect, the QQ plot would produce a straight line and fill out the area in $[-1, 1]$.

A comparison of the graphs in Figs. 16.2 and 16.3 show that the EMA method is calibrated better than the RMA method. The RMA method clearly shows outliers at both ends. The interval boundaries of $[-1, 1]$ are in both cases clearly exceeded. This indicates a possible inadequacy of an assumed normal distribution. QQ plots for the year 1995 are not shown, which also clearly show the dominance of EMA over RMA.

Another important assumption of our model is the independence of the re-scaled random variable Z_t. Figure 16.4 shows the outliers of another bank

$$\{t, \mathbf{1}(L_{t+1} > \widehat{VaR}_t)\}, \ t = 1, \ldots, 750, \qquad (16.20)$$

as a function of t. The contradictory temporal non-uniform distribution of the outliers from the independence of Z_t is clearer to see by the RMA method than by the EMA method.

The exploratory analysis clearly shows the differences between RMA and EMA. As a supplement we now compare both estimation techniques with an appropriate

VaR Reliability Plot
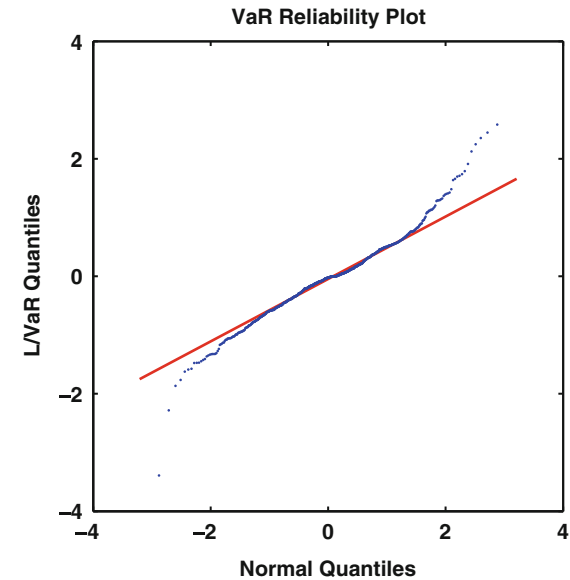
Fig. 16.3 QQ plot of $L_{t+1}/\widehat{VaR}_t$ for EMA in 1994   Q SFEVaRqqplot
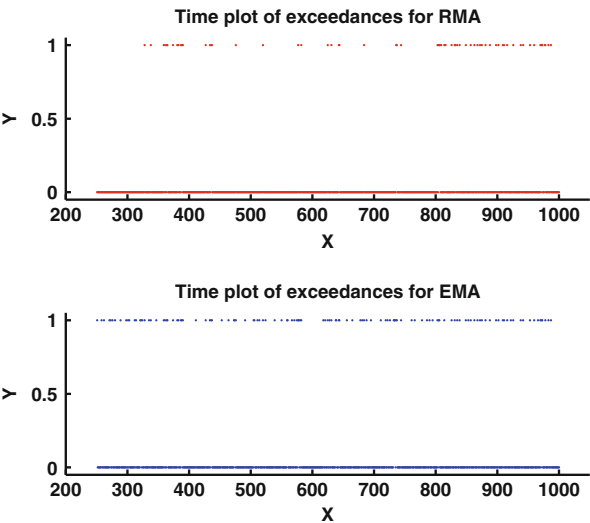
Fig. 16.4 Time diagram of the exceedances at the 80 % significance level from VaR for RMA (*left*) and EMA. The superiority of EMA is obvious   Q SFEVaRtimeplot2

test within the framework of the model (16.7) and (16.8). We again consider the sample residuals $\hat{Z}_{t+1}$ from (16.16) and set the threshold value in (16.14) to $u = 0.8416$, i.e. to the 80 % quantile of the distribution of $Z_{t+1} = L_{t+1}/\sigma_t$. From this

Table 16.2 $H_0 : \vartheta \stackrel{(\leq)}{=} 1.4$

Method	$\vartheta = 1.4$	$\varsigma = 0.46$	$\sqrt{N(u)}\frac{\hat{\vartheta}-\vartheta}{\hat{\varsigma}}$	Significance	$N(u)$
EMA	$\hat{\vartheta} = 1.72$	$\hat{\varsigma} = 1.01$	2.44	0.75 %	61
RMA	$\hat{\vartheta} = 1.94$	$\hat{\varsigma} = 1.3$	3.42	0.03 %	68

Table 16.3 $H_0 : \vartheta \stackrel{(\leq)}{=} 1.47$

Method	$\vartheta = 1.47$	$\varsigma = 0.546$	$\sqrt{N(u)}\frac{\hat{\vartheta}-\vartheta}{\hat{\varsigma}}$	Significance	$N(u)$
EMA	$\hat{\vartheta} = 1.72$	$\hat{\varsigma} = 1.01$	2.01	2.3 %	61
RMA	$\hat{\vartheta} = 1.94$	$\hat{\varsigma} = 1.3$	3.04	0.14 %	68

Table 16.4 $H_0 : \vartheta \stackrel{(\leq)}{=} 1.4$ largest outlier excluded

Method	$\vartheta = 1.4$	$\varsigma = 0.46$	$\sqrt{N(u)}\frac{\hat{\vartheta}-\vartheta}{\hat{\varsigma}}$	Significance	$N(u)$
EMA	$\hat{\vartheta} = 1.645$	$\hat{\varsigma} = 0.82$	2.31	1 %	60
RMA	$\hat{\vartheta} = 1.83$	$\hat{\varsigma} = 0.93$	3.78	0.00 %	67

we obtain $\vartheta = 1.4$ according to (16.14). Due to the asymptotic distribution (16.18) we can check the significance of the hypothesis

$$H_0 : \vartheta \stackrel{(\leq)}{=} 1.4. \tag{16.21}$$

A better approximation than the standard normal distribution for the sample is the Student $t(20)$ distribution, if we generalize the degrees of freedom.

$$\mathcal{L}(\hat{Z}_{t+1}) = \mathcal{L}\left(\frac{L_{t+1}}{\hat{\sigma}_t}\right) \approx t(20). \tag{16.22}$$

The value of ϑ obtained differs from the value given above by 5 %, the corresponding variances ς^2 by 18 %. Therefore, we also consider the hypothesis

$$H_0 : \vartheta \stackrel{(\leq)}{=} 1.47. \tag{16.23}$$

The following Tables 16.2, 16.3, 16.4 and 16.5 summarizes our results.

From Tables 16.2 and 16.3 it is obvious that the observed outliers for EMA are calibrated better than for the RMA method. For a random sample of 260 values we expect 52 outliers (standard deviation 6.45). For EMA we observe 61 ($61-52 \approx 1.5 \cdot$ standard deviation) outliers and for RMA 68 ($68 - 52 \approx 2.5 \cdot$ standard deviation). Naturally the outliers influence the test considerably. We can therefore repeat the analysis excluding the outliers and obtain (16.4) and (16.5).

To conclude we can say that the EMA method gives better calibrated results than the RMA method. Both methods are extremely sensitive to outliers and should

Table 16.5 $H_0 : \vartheta \overset{(\leq)}{=} 1.47$ largest outlier excluded

Method	$\vartheta = 1.47$	$\varsigma = 0.546$	$\sqrt{N(u)}\frac{\hat\vartheta - \vartheta}{\varsigma}$	Significance	$N(u)$
EMA	$\hat\vartheta = 1.645$	$\hat\varsigma = 0.82$	1.65	5 %	60
RMA	$\hat\vartheta = 1.83$	$\hat\varsigma = 0.93$	3.1	0.15 %	67

both be considered. Even the EMA method suffers from the assumptions (16.7) and (16.8), which are based on the Delta-Normal Model, can only be approximately fulfilled. The residuals Z_t are neither normally distributed nor independent, although the EMA method is not strongly effected by the independence assumption due to its exponentially decreasing memory.

16.4   Recommended Literature

The classic start to Value-at-Risk (VaR) estimation lies in the consideration of linear or linearized portfolios, see RiskMetrics (1996). The linear structure transforms the multi-dimensional normally distributed random variables into one dimensional Gaussian values whose quantile can be estimated. An introduction to the asymptotic distributions of extreme values can be found in Leadbetter, Lindgren and Rootzen (1983) and Embrechts et al. (1997). McAllister and Mingo (1996) describe the advantages from (16.13) in a RAROC (risk-adjusted return on capital) setup. Artzner, Delbaen, Eber and Heath (1997) claim that the expected shortfall is a coherent measurement of risk. Jaschke and Küchler (1999) show that (16.13) is a reasonable approximation for a worst case scenario. Leadbetter et al. (1983) show how (16.13) can be used in the context of the theory of extreme values. A good overview of the VaR problems is given in Jorion (2000). The majority of the German laws can be found under http://www.bafin.de. Taleb (2001) is a critic of specific VaR definitions and gives several examples in which Value-at-Risk definitions can be "blinding" given certain trading strategies ("Peso Problem Traders"). A complete literature review can be found in Franke, Härdle and Stahl (2000). The VaR calculations from Overbeck (2000) based on the ability to pay process are discussed and country risk is evaluated in Lehrbass (2000).

16.5   Exercises

Exercise 16.1 *Discuss the standard methodologies for calculating VaR and explain how they work. Are there advantages and disadvantages of the presented methods?*

Exercise 16.2 *Name and discuss a few important problems in the implementation of VaR.*

Exercise 16.3 *Let* Z *be a* $N(0, 1)$ *rv, prove that* $\vartheta = E[Z|Z > u] = \varphi(u)/\{1 - \Phi(u)\}.$

Exercise 16.4 *Recall the definitions of* Z *and* ϑ *for given exceedance level* u *from Exercise 16.3. Prove that* $\varsigma^2 = Var[Z|Z > u] = 1 + u\vartheta - \vartheta^2.$

Exercise 16.5 *Consider a portfolio with two stocks. The portfolio has 2,000 EUR invested in* S_1 *and 4,000 EUR in* S_2*. Given the following variance–covariance matrix of the returns* R_1 *and* R_2*:*

$$\Sigma = \begin{pmatrix} 0.1^2 & 0 \\ 0 & 0.06^2 \end{pmatrix}$$

calculate the VaR (at 95 %) for each stock and VaR of the portfolio using the Delta-Normal Model.

Exercise 16.6 *Component VaR, CVaR, is the partition of the portfolio VaR that indicates how much the portfolio VaR would change approximately if the given component was deleted. It can be calculated with* $CVaR_i = w_i \Delta VaR_i$ *where* ΔVaR_i *is the incremental VaR of the position* i*, i.e. how much the VaR of the portfolio increases if we increase the position* i *by 1. This value can be calculated directly by revaluating the portfolio or be approximated using the marginal VaR. Calculate both incremental and marginal VaR in the case of Exercise 16.5 (change by 1 for each position) and compare them. Calculate the approximated* $CVaR_i$*. What do you discover?*

Exercise 16.7 *Suppose a portfolio manager manages a portfolio which consists of a single asset. The natural logarithm of the portfolio value is normally distributed with an annual mean of 10 % and annual standard deviation of 30 %. The value of the portfolio today is 100 million EUR. Taking VaR as a quantile, answer the following:*

(a) What is the probability of a loss of more than 20 million EUR by year end (i.e. what is the probability that the end-of-year value is less than 80 million EUR)?
(b) With 1 % probability, what is the maximum loss at the end of the year? This is the VaR at 1 %.
(c) Calculate the daily, weekly and monthly VaRs at 1 %.

Exercise 16.8 *Calculate the daily VaR in a delta normal framework for the following portfolio with the given correlation coefficients. Do the same calculation for the cases of complete diversification and perfect correlation.*

Assets	Estimated daily VaR	$\rho_{S,FX}$	$\rho_{B,FX}$	$\rho_{S,B}$
Stocks(S)	400,000.00 EUR	-0.10	0.25	0.80
Bonds(B)	300,000.00 EUR			
Foreign Exchange(FX)	200,000.00 EUR			

Exercise 16.9 *A derivatives portfolio has a current market value of 250 million EUR. Marking this derivatives position, to obtain the market value that would have been obtained on the previous 201 trading days, yields the following worst cases for the daily fall in its value (in million EUR):*

−152	−132	−109	−88	−85	−76	−61	−55	−45	−39
−37	−32	−30	−26	−22	−21	−18	−15	−14	−12

Using the above data, what is the daily VaR on this portfolio at the 1 % threshold? At the 5 % threshold? Comment on the relative accuracy of these two calculations.

Exercise 16.10 *A risk measure ρ is subadditive when the risk of the total position is less than, or equal to, the sum of the risk of individual portfolios. Intuitively, subadditivity requires that risk measures should consider risk reduction by portfolio diversification effects. Subadditivity can be defined as follows: Let X and Y be random variables denoting the losses of two individual positions. A risk measure ρ is subadditive if the following equation is satisfied.*

$$\rho(X + Y) \leq \rho(X) + \rho(Y)$$

Using the above definition, show that the VaR based on the Delta-Normal Model is subadditive.

Exercise 16.11 *The digital option (also called the binary option) is the right to earn a fixed amount of payment conditional on whether the underlying asset price goes above (digital call option) or below (digital put option) the strike price. Consider the following two digital options on a stock, with the same exercise date T. The first option denoted by A (initial premium u) pays 1,000 if the value of the stock at time T is more than a given U, and nothing otherwise. The second option denoted by B (initial premium l) pays 1,000 if the value of the stock at time T is less than L (with L < U), and nothing otherwise. Suppose L and U are chosen such that $P(S_T < L) = P(S_T > U) = 0.008$, where S_T is the stock price at time T. Consider two traders, trader A and trader B, writing one unit of option A and option B, respectively.*

(a) Calculate the VaR at the 99 % confidence level for each trader.
(b) Calculate the VaR at the 99 % confidence level for the combined position on options A and B.
(c) Is the VaR for this exercise subadditive?

Exercise 16.12 *In the traffic light approach the backtesting surcharge factor increases with the number of exceptions evaluated on 250 historical returns. Consider a correct 1 % VaR model and assume the independence of the returns. It would mean that the appearance exceptions on 250 days follows a binomial*

distribution with parameter $n = 250$ and $p = 0.01$. Calculate the probability that there are more than four exceptions.

Exercise 16.13 *Similarly to Exercise 16.12 consider an incorrect 1 % VaR model which has the true probability of exception $p = 0.025$. Calculate the probability that there are less than five exceptions.*

Exercise 16.14 *(a) Consider a portfolio which consists of 20 stocks of type A. The price of the stock today is 10 EUR. What is the 95 % 1 year Value-at-Risk (VaR) of this portfolio if the 1-year return arithmetic of the stock R_A is normally distributed $N(0, 0.04)$?*

(b) Consider again a portfolio of ten stocks of type A and 20 stocks of type B. The prices of the stocks today are 10 EUR and 5 EUR respectively. The joint distribution of the yearly arithmetic returns follows a two-dimensional normal distribution $N(\mu, \Sigma)$, where $\mu = (0, 0)^T$ and the covariance matrix of the returns R_A and R_B is given by:

$$\Sigma = \begin{pmatrix} 0.04 & 0.02 \\ 0.02 & 0.08 \end{pmatrix}$$

What is the yearly 95 % VaR of the portfolio in this case?

Exercise 16.15 *Let $X \sim N(\mu, \sigma^2)$ be a random variable, c is a constant, please calculate $E[\max(X - c, 0)]$.*

Exercise 16.16 *Calculate the expected shortfall for $X \sim N(\mu, \sigma^2)$.*

Chapter 17
Copulae and Value at Risk

The capital requirement from financial institutions is based on the amount of risk carried in their portfolios. The risk associated with a portfolio may originate from:

1. Fluctuations in the value of financial assets composing the portfolio (*market* risk),
2. Fluctuations in the credibility of debtors (*credit* risk),
3. Uncertainty connected with technical, personal and natural factors that may influence the portfolio value (*operational* risk).

In order to investigate the risk of a portfolio, the assets subjected to risk (*risk factors*) should be identified and the changes in the portfolio value caused by the risk factors evaluated. Especially relevant for risk management purposes are *negative changes*— the portfolio *losses*.

The *Value-at-Risk* (VaR) is a measure that quantifies the riskiness of a portfolio. This measure and its accuracy are of crucial importance in determining the capital requirement from financial institutions and is one of the reasons why increasing attention has been paid to VaR computing methods.

The losses and the probabilities associated with them (the *distribution of losses*) are necessary to describe the degree of portfolio riskiness. The riskier the portfolio, the higher the probability of losses being larger than a certain amount is. In other words, the riskier the portfolio, the larger are the minimal losses for a certain probability (also called *level*). This is the precise definition of VaR: VaR is a quantile of the distribution of portfolio losses representing the minimal losses for a certain level.

Looking carefully at the distribution of losses, one verifies that large losses are influenced by simultaneous losses in risk factors. Hence, the *distribution of losses depends on joint distribution of risk factors*.

Understanding the joint distribution of risk factors is fundamental in investigating and computing the Value-at-Risk. The conventional procedure to model joint distributions of financial returns is to approximate them with *multivariate normal distributions*.

That implies, however, that the dependence structures of the returns are reduced to a fixed type. Even if the autocorrelation structure is neglected, predetermining a multivariate normal distribution means that the following assumptions hold:

1. symmetric distribution of returns,
2. the tails of the distribution are not too heavy,
3. linear dependence.

Empirical evidence for these assumptions are barely verified and an alternative model is needed, with more flexible dependence structure and arbitrary marginal distributions. These are exactly the characteristics of *copulae*.

Copulae are very useful for modelling and estimating multivariate distributions. The flexibility of copulae basically follows from *Sklar's Theorem*, which says that each joint distribution can be "decomposed" into its marginal distributions and a copula C "responsible" for the dependence structure:

$$F(x_1 \ldots, x_d) = C\{F_1(x_1), \ldots, F_d(x_d)\}.$$

Two important factors for practical applications rely on this theorem:

1. The construction of multivariate distributions may be done in two independent steps: the specification of marginal distributions—not necessarily identical—and the specification of a dependence structure. Copulae "couple together" the marginal distributions into a multivariate distribution with the desired dependence structure.
2. Joint distributions can be separately estimated from a sample of observations: the marginal distributions are estimated first, the dependence structure later.

The copula approach gives us more freedom than the normality assumptions, marginal distributions with asymmetric heavy tails (typical for financial returns) can be combined with different dependence structures, resulting in multivariate distributions (far different from the multivariate normal) that better describe the empirical characteristics of financial returns distribution.

Moreover, copulae allow for dynamical modelling and adaption to portfolios, different copulae with distinct properties can be associated to different portfolios according to their specific dependence structures. Furthermore, copulae may change as time evolves, reflecting the evolution of the dependence between financial assets. To summarize, the Value-at-Risk estimation with copulae is more efficient and flexible than any method based on normality assumption.

17.1 Copulae

The description of copulae for measuring and modelling dependence with its main properties is the subject of this section. There are a lot of different equivalent definitions that could define the copula, but the most general is the following one.

Definition 17.1 (Copula) A d-dimensional copula is a function C from $[0, 1]^d \to [0, 1]$ with the following properties:

1. $C(u_1, \ldots, u_d)$ is increasing in each component $u_i \in [0, 1]$, $i = 1, \ldots, d$.
2. $C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$ for all $u_i \in [0, 1]$, $i = 1, \ldots, d$.
3. For all $(u_1, \ldots, u_d)$, $(u'_1, \ldots, u'_d) \in [0, 1]^d$ with $u_i < u'_i$ we have

$$\sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1 + \cdots + i_d} C(v_{j1}, \ldots, v_{jd}) \geq 0,$$

where $v_{j1} = u_j$ and $v_{j2} = u'_j$, for all $j = 1, \ldots, d$.

The first and third properties state that copulae are grounded functions and that all d-dimensional boxes with vertices in $[0, 1]^d$ have *non-negative C-volume*. Together they guarantee that copulae are distribution functions on the d-dimensional unit cube, while the second property reveals that copulae have *uniform marginal distributions*. From this we conclude the following equivalent definition of the copulae.

Definition 17.2 (Copula) A d-dimensional copula is a d-dimensional distribution with all uniform marginal distributions.

Note that by considering random variables $X_1, \ldots, X_d$ with univariate distribution functions $F_{X_1}, \ldots, F_{X_d}$ and the random variables $U_i = F_{X_i}(X_i)$, $i = 1, \ldots, d$ uniformly distributed in $[0, 1]$, a copula may be interpreted as *the joint distribution of the marginal distributions*.

Copulae gained popularity through Sklar's (1959) work where the term was first coined. However, many results had already been proved by Hoeffding (1940) and Hoeffding (1941), who could have been the founder of a copula theory, if he had considered the stochastically more intuitive dependency over the unit cube $[0, 1]^2$ rather than over $[-1/2, 1/2]^2$ as he had done. Copulae allow marginal distributions to be separated from the dependency structure. Sklar's theorem connects copulae with distribution functions such that from the one side every distribution function can be "decomposed" into its marginal distribution and (at least) one copula and from the other side a (unique) copula is obtained from "decoupling" every (continuous) multivariate distribution function from its marginal distributions.

Theorem 17.1 (Sklar's Theorem) *Let F be a multivariate distribution function with margins $F_1, \ldots, F_d$, then exists the copula C such that*

$$F(x_1, \ldots, x_d) = C\{F_1(x_1), \ldots, F_d(x_d)\}, \quad x_1, \ldots, x_d \in \overline{\mathbb{R}}.$$

If F_i are continuous for $i = 1, \ldots, d$ then C is unique. Otherwise C is uniquely determined on $F_1(\overline{\mathbb{R}}) \times \cdots \times F_d(\overline{\mathbb{R}})$.

Conversely, if C is a copula and $F_1, \ldots, F_d$ are univariate distribution functions, then the function F defined above is a multivariate distribution function with margins $F_1, \ldots, F_d$.

The representation in Sklar's Theorem can be used to construct new multivariate distributions by changing either the copula function or marginal distributions. For an arbitrary continuous multivariate distribution we can determine its copula from the transformation

$$C(u_1, \ldots, u_d) = F\{F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)\}, \quad u_1, \ldots, u_d \in [0, 1], \qquad (17.1)$$

where F_i^{-1} are inverse marginal distribution functions.

Copulae are *invariant under monotone increasing transformations*. This property is very useful in practical applications, where it is often of interest to describe dependency or independency rather than correlation or uncorrelation.

Theorem 17.2 *Let $X = (X_1, \ldots, X_d)^\top$ be a random vector with continuous marginals and copula C_X and $T_1, \ldots, T_d$ be strictly increasing functions on $range(X_1), \ldots, range(X_d)$. Let $Y = (Y_1, \ldots, Y_d)^\top$, $Y_i = T_i(X_i)$ be a random vector with copula C_Y. Then $C_X = C_Y$ almost everywhere.*

A d-dimensional random variable determines a copula through its joint and marginal distributions. Moreover, monotone increasing transformations on the random variable do not affect the copula. These are the main ideas used to obtain the Gaussian copula: the random variable $X = (X_1, \ldots, X_d)^\top$ with multivariate normal distribution and copula C_X is transformed into the standardized variable $Z = (Z_1, \ldots, Z_d)^\top$, $Z_j \sim N(0, 1)$. The copula of the random variable Z is C_X.

Since the copula function is a multivariate distribution with uniform margins, it follows that the copula density can be determined in the usual way

$$c(u_1, \ldots, u_d) = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d}, \quad u_1, \ldots, u_d \in [0, 1],$$

Being armed with Theorem 17.1 and (17.1) we can write density function $f(\cdot)$ of the bivariate distribution F in terms of copula as follows

$$f(x_1, \ldots, x_d) = c\{F_1(x_1), \ldots, F_d(x_d)\} \prod_{i=1}^{d} f_i(x_i), \quad x_1, \ldots, x_d \in \overline{\mathbb{R}}.$$
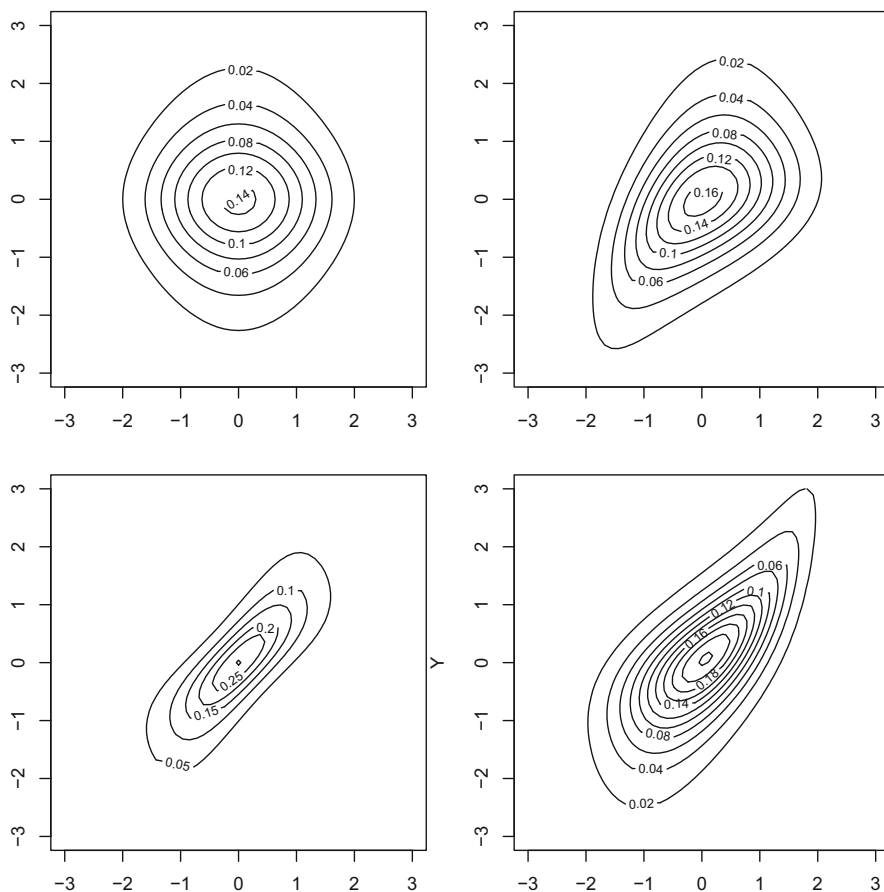
Fig. 17.1 pdf contour plots, $F(x_1, x_2) = C\{\Phi(x_1), t_3(x_2)\}$ with (*clockwise*) Gaussian ($\rho = 0$), Clayton ($\theta = 0.9$), Frank ($\theta = 8$) and Gumbel ($\theta = 2$) copulae   ◘ SFEplotCop

The contour plots from $f(x_1, x_2)$ are shown in Fig. 17.1 for the respective copula choices.

17.2   Copula Classes

Since there are plenty of functions satisfying the assumption of Theorem 17.1 they should be classified by construction and properties. Here we consider several main classes, like *simplest*, *elliptical*, *Archimedean copulae* and *hierarchical Archimedean copulae*.

17.2.1   Simplest Copulae

Special cases, like independence and perfect positive or negative dependence can be represented by copulae. If d random variables $X_1, \ldots, X_d$ are stochastically independent from Theorem 17.1, then the structure of such a relationship is given by the product copula

$$\Pi(u_1, \ldots, u_d) = \prod_{j=1}^{d} u_j. \tag{17.2}$$

Copulae are bounded, this means that for all $u = (u_1, \ldots, u_d)^\top \in [0, 1]^d$:

$$W(u_1, \ldots, u_d) \le C(u_1, \ldots, u_d) \le M(u_1, \ldots, u_d)$$

where

$$M(u_1, \ldots, u_d) = \min(u_1, \ldots, u_d)$$

is called the *Fréchet–Hoeffding upper bound* and

$$W(u_1, \ldots, u_d) = \max\left(\sum_{i=1}^{d} u_i - d + 1, 0\right)$$

is the *Fréchet–Hoeffding lower bound*. While W is not a copula for $d > 2$, M is a copula for all d. Both structures represent the perfect negative and perfect positive dependence. From this observation we may conclude that an arbitrary copula C reflects dependence which lies between the perfect negative and positive one. Figure 17.2 illustrates the Fréchet–Hoeffding bounds and the product copulae.

17.2.2   Elliptical Copulae

Due to the popularity of Gaussian and t-distributions elliptical copulae also play an important role. The construction of this type of copula is based directly on Theorem 17.1.

Definition 17.3 (Elliptical Distribution) If Y is a d-dimensional random vector and, for some vector $\mu \in \mathbb{R}^d$, some non-negative definite symmetric $d \times d$ matrix Σ and some function $\phi : [0; \infty) \to \mathbb{R}^d$, the characteristic function $\varphi_{Y-\mu}$ is of the form $\varphi_{Y-\mu}(t) = \phi(t^\top \Sigma t)$, we say that Y has an elliptical distribution with parameters μ, Σ and ϕ. The function ϕ is referred to as the characteristic function of Y.

The elliptical copulae are derived from the elliptical distributions using Theorem 17.1.
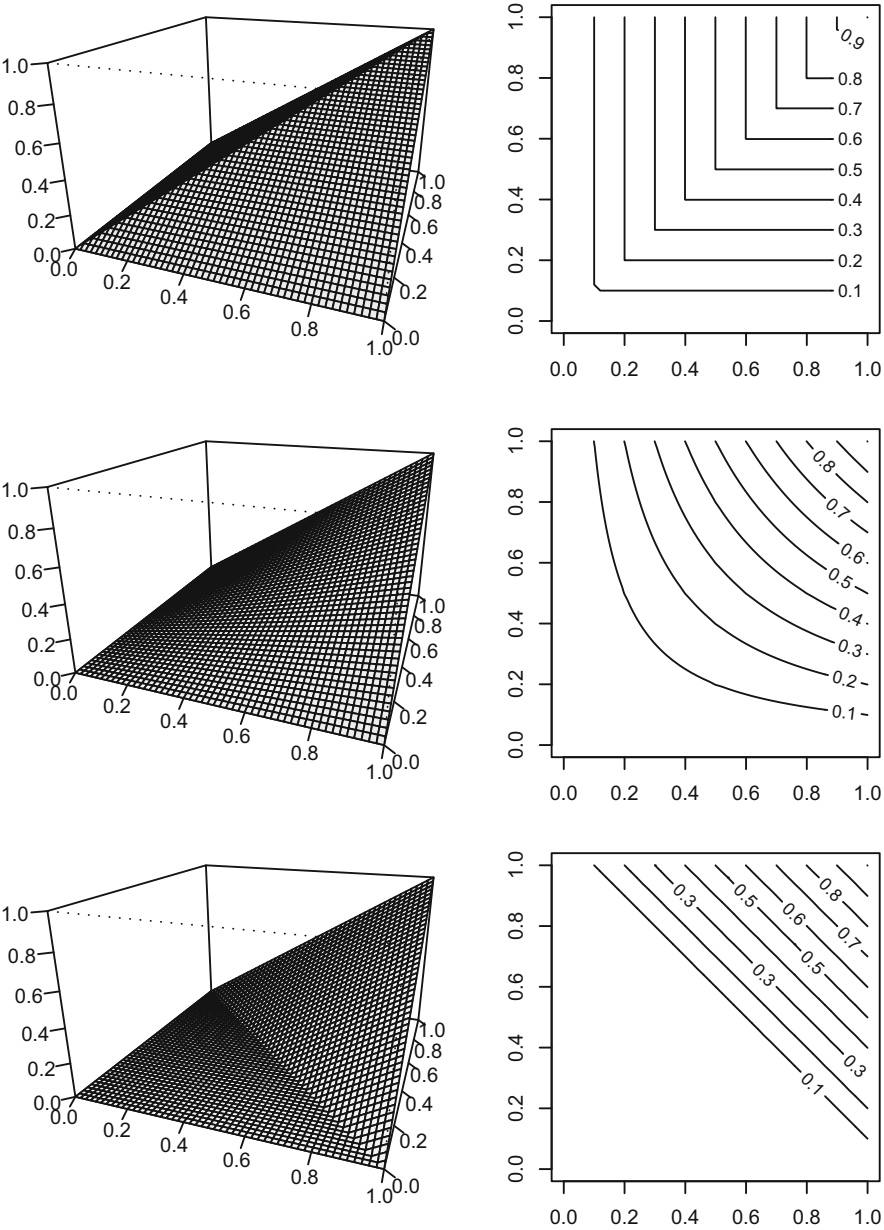
Fig. 17.2 Fréchet–Hoeffding upper bound $W(u_1, u_2)$ (*upper panel*), product copula $\Pi(u_1, u_2)$ (*middle panel*), Fréchet–Hoeffding lower bound $M(u_1, u_2)$ (*lower panel*)   SFEfrechet

Gaussian Copula

The Gaussian copula represents the *dependence structure* of the multivariate normal distribution, that means that *normal* marginal distributions are combined with a Gaussian copula to form multivariate normal distributions. The combination of *non-normal* marginal distributions with a Gaussian copula results in *meta-Gaussian* distributions, i.e. distributions where *only* the dependence structure is Gaussian.

To obtain the Gaussian copula, let $X = (X_1, \ldots, X_d)^\top \sim N_d(\mu, \Sigma)$ with $X_j \sim N(\mu_j, \sigma_j)$ for $j = 1, \ldots, d$. A copula C_X exists:

$$F_X(x_1, \ldots, x_d) = C_X\{F_1(x_1), \ldots, F_d(x_d)\},$$

where F_j is the distribution function of X_j and F_X the distribution function of X.

Let $Y_j = T_j(X_j)$, $T_j(x) = (x - \mu_j)/\sigma_j$. Then $Y_j \sim N(0, 1)$ and $Y = (Y_1, \ldots, Y_d)^\top \sim N_d(0, \Psi)$ where Ψ is the correlation matrix associated with Σ. A copula C_Ψ^{Ga}, called *Gaussian copula* exists as follows:

$$F_Y(y_1, \ldots, y_d) = C_\Psi^{Ga}\{\Phi(y_1), \ldots, \Phi(y_d)\}. \tag{17.3}$$

An explicit expression for the Gaussian copula is obtained by rewriting (17.3) with $u_j = \Phi(y_j)$:

$$
\begin{aligned}
C_\Psi^{Ga}(u_1, \ldots, u_d) &= F_Y\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)\} \\
&= \int_{-\infty}^{\Phi^{-1}(u_1)} \cdots \int_{-\infty}^{\Phi^{-1}(u_d)} (2\pi)^{-\frac{d}{2}} \mid \Psi \mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2} r^\top \Psi^{-1} r\right) dr_1 \ldots dr_d.
\end{aligned}
$$

As $T_j(x)$ is increasing, it follows from Theorem 17.2 that

$$C_X = C_\Psi^{Ga}.$$

Thus, *any* multivariate normal distribution can be constructed from its marginal distributions and the Gaussian copula C_Ψ^{Ga} with the desired correlation matrix Ψ.

Remark 17.1 If $\Psi = I_d$ the Gaussian copula becomes the product copula as

$$
\begin{aligned}
C_{I_d}^{Ga}(u_1, \ldots, u_d) &= \int_{-\infty}^{\Phi^{-1}(u_1)} \cdots \int_{-\infty}^{\Phi^{-1}(u_d)} (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2} \sum_{j=1}^{d} r_j^2\right) dr_1 \ldots dr_d \\
&= \int_{-\infty}^{\Phi^{-1}(u_1)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} r_1^2\right) dr_1 \times \cdots \times \\
&\quad \times \int_{-\infty}^{\Phi^{-1}(u_d)} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} r_d^2\right) dr_d
\end{aligned}
$$

$$= \Phi\{\Phi^{-1}(u_1)\}\ldots\Phi\{\Phi^{-1}(u_d)\}$$
$$= \Pi(u_1,\ldots,u_d).$$

The *density of the Gaussian copula* (Fig. 17.3) is obtained by differentiating (17.3),

$$\mid 2\pi\Psi\mid^{-\frac{1}{2}} \exp\left(-\frac{1}{2}x^\top\Psi^{-1}x\right) = c_\Psi^{Ga}\{\Phi(x_1),\ldots,\Phi(x_d)\}$$

$$\times\prod_{j=1}^{d}(2\pi)^{-\frac{1}{2}}\exp\left(-\frac{1}{2}x_j^2\right)$$
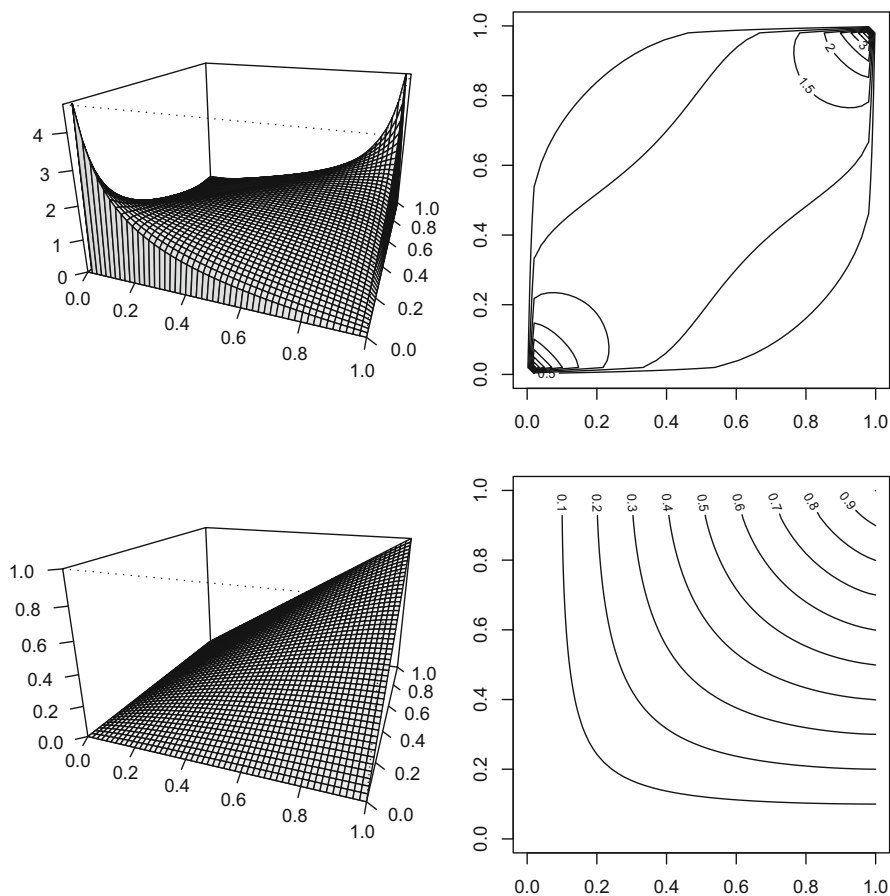
Fig. 17.3   Density and the cumulative distribution of the Gaussian copula, $c_\Psi^{Ga}(u_1,u_2)$, $\psi_{12}=0.5$

SFEgausscop

rearranging terms and defining $\zeta_j = \Phi^{-1}(u_j)$, $j = 1, \ldots, d$, $\zeta = (\zeta_1, \ldots, \zeta_d)^\top$:

$$c_\Psi^{Ga}(u_1, \ldots, u_d) = |\Psi|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\text{``}^\top(\Psi^{-1} - I_d)\text{``}\right\}. \tag{17.4}$$

Student's t-Copula

The t-copula, containing the dependence structure from the multivariate t-distribution, may be obtained in a similar way.

Let $X = (X_1, \ldots, X_d)^\top \sim t_d(v, \mu, \Sigma)$ and $Y = (Y_1, \ldots, Y_d)^\top \sim t_d(v, 0, \Psi)$ where Ψ is the correlation matrix associated with Σ. The unique copula from Y is the *Student's t-copula* $C_{v,\Psi}^t$. Similarly for normal copula it follows from Theorem 17.2 that $C_X = C_{v,\Psi}^t$.

For $u = (u_1, \ldots, u_d)^\top \in [0, 1]^d$, the *Student's t-copula* is given by

$$C_{v,\Psi}^t(u_1, \ldots, u_d) = t_{v,\Psi}\{t_v^{-1}(u_1), \ldots, t_v^{-1}(u_d)\}$$

where t_v^{-1} is the quantile function from the univariate t-distribution and $t_{v,\Psi}$ the distribution function of Y.

The *density of the t-copula* (Fig. 17.4) is given by

$$c_{v,\Psi}^t(u_1, \ldots, u_d) = \frac{t_{v,\Psi}\{t_v^{-1}(u_1), \ldots, t_v^{-1}(u_d)\}}{\prod_{j=1}^d t_{v,\Psi}\{t_v^{-1}(u_j)\}}. \tag{17.5}$$

With $\zeta_j = t_v^{-1}(u_j)$ for $j = 1, \ldots, d$ the density of the t-copula can be expressed as:

$$c_{v,\Psi}^t(u_1, \ldots, u_d) = |\Psi|^{-\frac{1}{2}} \frac{\Gamma(\frac{v+d}{2})\left\{\Gamma(\frac{v}{2})\right\}^{d-1}\left(1 + \frac{1}{v}\zeta^\top\Psi^{-1}\zeta\right)^{-\frac{v+d}{2}}}{\left\{\Gamma(\frac{v+1}{2})\right\}^d \prod_{j=1}^d \left(1 + \frac{1}{v}\zeta_j^2\right)^{-\frac{v+1}{2}}}. \tag{17.6}$$

17.2.3 Archimedean Copulae

As opposed to elliptical copulae, Archimedean copulae are not constructed using Theorem 17.1, but are related to Laplace transforms of univariate distribution functions. Let $\mathbb{L}$ denote the class of Laplace transforms which consists of strictly decreasing differentiable functions Joe (1997), i.e.

$$\mathbb{L} = \{\phi : [0; \infty) \to [0, 1] \mid \phi(0) = 1, \phi(\infty) = 0; (-1)^j \phi^{(j)} \geq 0; j = 1, \ldots, \infty\}.$$
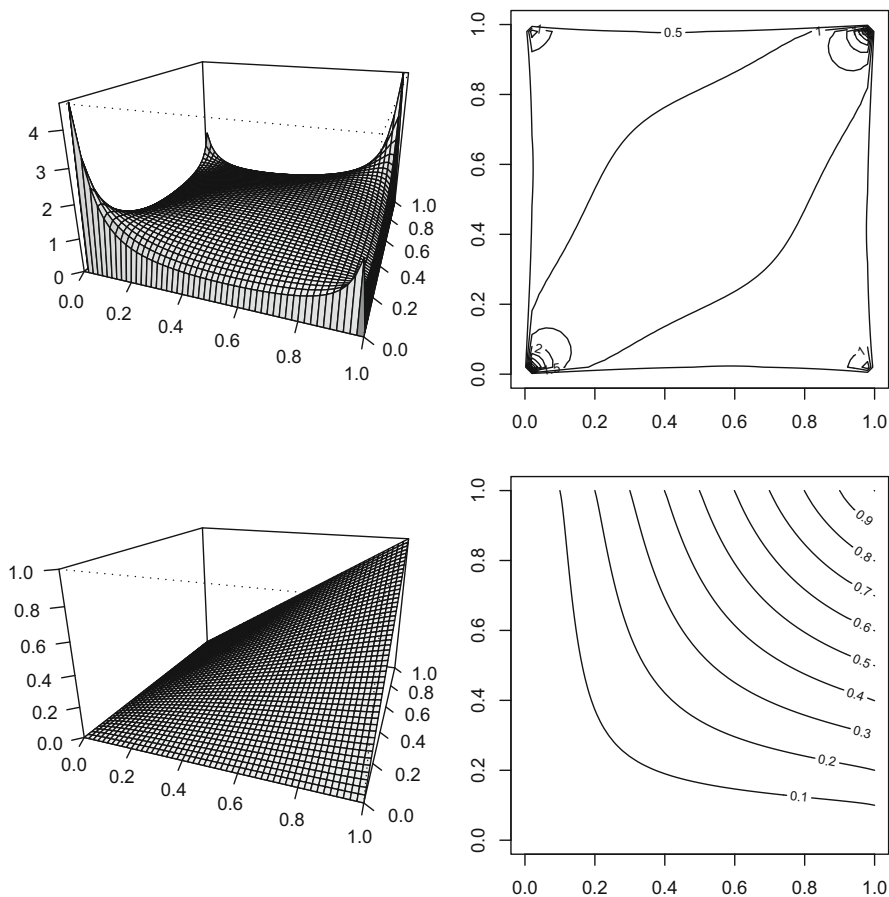
Fig. 17.4   Density of t-copula, $c_{\nu,\Psi}^t(u_1, u_2)$, $\psi_{12} = 0.2$, $\nu = 3$ ▣ SFEtcop

The function $C : [0, 1]^d \to [0, 1]$ defined as

$$C(u_1, \ldots, u_d) = \phi\{\phi^{-1}(u_1) + \cdots + \phi^{-1}(u_d)\}, \quad u_1, \ldots, u_d \in [0, 1]$$

is a d-dimensional Archimedean copula, where $\phi \in \mathbb{L}$ and is called the *generator of the copula*. It is straightforward to show that $C(u_1, \ldots, u_d)$ satisfies the conditions of Definition 17.1.

Some d-dimensional Archimedean copulae are presented below.

Frank (1979) Copula, $0 \leq \theta < \infty$

The first popular Archimedean copula is the so-called Frank copula, which is the only elliptical Archimedean copula. Its generator and copula functions are

$$\phi(x, \theta) = \theta^{-1} \log\{1 - (1 - e^{-\theta})e^{-x}\}, \quad 0 \leq \theta < \infty, \; x \in [0, \infty).$$

$$C_\theta(u_1, \ldots, u_d) = -\frac{1}{\theta} \log \left[1 + \frac{\prod_{j=1}^{d} \{\exp(-\theta u_j) - 1\}}{\{\exp(-\theta) - 1\}^{d-1}} \right].$$

The dependence becomes maximal when θ tends to infinity and independence is achieved when $\theta = 0$.

Gumbel (1960) Copula, $1 \leq \theta < \infty$

The Gumbel copula is frequently used in financial applications. Its generator and copula functions are

$$\phi(x, \theta) = \exp\{-x^{1/\theta}\}, \quad 1 \leq \theta < \infty, \; x \in [0, \infty)$$

$$C_\theta(u_1, \ldots, u_d) = \exp \left[-\left\{ \sum_{j=1}^{d} (-\log u_j)^\theta \right\}^{\theta^{-1}} \right].$$

Consider a bivariate distribution based on the Gumbel copula with univariate extreme value marginal distributions. Genest and Rivest (1989) showed that this distribution is the only bivariate extreme value distribution based on an Archimedean copula. Moreover, all distributions based on Archimedean copulae belong to its domain of attraction under common regularity conditions. In contrary to the elliptical copulae, the Gumbel copula leads to asymmetric contour diagrams. The Gumbel copula shows stronger linkage between positive values, however, it also shows more variability and more mass in the negative tail.

For $\theta > 1$ this copula allows for the generation of dependence in the upper tail. For $\theta \to 1$, the Gumbel copula reduces to the product copula and for $\theta \to \infty$ we obtain the Fréchet–Hoeffding upper bound.

Clayton (1978) Copula, $-1/(d-1) \leq \theta < \infty, \; \theta \neq 0$

The Clayton copula which, in contrast to the Gumbel copula, has more mass on the lower tail, and less on the upper. The generator and copula function are

$$\phi(x, \theta) = (\theta x + 1)^{-\frac{1}{\theta}}, \quad -1 \le \theta < \infty, \, \theta \ne 0, \, x \in [0, \infty),$$

$$C_\theta(u_1, \ldots, u_d) = \left\{ \left(\sum_{j=1}^{d} u_j^{-\theta} \right) - d + 1 \right\}^{-\theta^{-1}}.$$

The Clayton copula is one of few copulae that has a simple explicit form of density for any dimension

$$c_\theta(u_1, \ldots, u_d) = \prod_{j=1}^{d} \{1 + (j-1)\theta\} u_j^{-(\theta+1)} \left(\sum_{j=1}^{d} u_j^{-\theta} - d + 1 \right)^{-(\theta^{-1}+d)}.$$

As the parameter θ tends to infinity, dependence becomes maximal and as θ tends to zero, we have independence. As $\theta \to -1$, the distribution tends to the lower Fréchet bound.

Ali–Mikhail–Haq Copula, $-1 \le \theta < 1$

Another well-known copula is the so-called Ali–Mikhail–Haq copula. Its generator and copula functions are:

$$\phi(x, \theta) = (1 - x \exp \theta)/(1 - x), \quad -1 \le \theta < 1, \, x \in [0, \infty),$$

$$C_\theta(u_1, \ldots, u_d) = \frac{\displaystyle\prod_{j=1}^{d} u_j}{1 - \theta \left\{ \displaystyle\prod_{j=1}^{d} (1 - u_j) \right\}}.$$

If $\theta = 0$, then $C_0 = \Pi$, the product copula.

17.2.4   Hierarchical Archimedean Copulae

A recently developed flexible method is provided by hierarchical Archimedean copulae (HAC). The special, so-called fully nested case of the copula function is:

$$\begin{aligned}
C(u_1, \ldots, u_d) &= \phi_{d-1}\{\phi_{d-1}^{-1} \circ \phi_{d-2}(\ldots [\phi_2^{-1} \circ \phi_1\{\phi_1^{-1}(u_1) + \phi_1^{-1}(u_2)\} \\
&\quad + \phi_2^{-1}(u_3)] + \cdots + \phi_{d-2}^{-1}(u_{d-1})) + \phi_{d-1}^{-1}(u_d)\} \\
&= \phi_{d-1}[\phi_{d-1}^{-1} \circ C(\{\phi_1, \ldots, \phi_{d-2}\})(u_1, \ldots, u_{d-1}) + \phi_{d-1}^{-1}(u_d)]
\end{aligned}$$

for $\phi_{d-i}^{-1} \circ \phi_{d-j} \in \mathbb{L}^*$, $i < j$, where

$$\mathbb{L}^* = \{\omega : [0; \infty) \to [0, \infty) \,|\, \omega(0) = 0,$$
$$\omega(\infty) = \infty; \, (-1)^{j-1}\omega^{(j)} \geq 0; \, j = 1, \dots, \infty\}.$$

In contrast to the Archimedean copula, the HAC defines the whole dependency structure in a recursive way. At the lowest level the dependency between the first two variables is modelled by a copula function with the generator ϕ_1, i.e. $z_1 = C(u_1, u_2) = \phi_1\{\phi_1^{-1}(u_1) + \phi_1^{-1}(u_2)\}$. At the second level another copula function is used to model the dependency between z_1 and u_3, etc. Note that the generators ϕ_i can come from the same family and they differ only through the parameter or, to introduce more flexibility, they come from different generator families. As an alternative to the fully nested model, we can consider copula functions, with arbitrary chosen combinations at each copula level. Okhrin, Okhrin and Schmid (2013) provide several methodologies in determining the structure of the HAC from the data. The case of $d = 3$ which we use further in applications is quite a simple one. If τ_{12}, τ_{13} and τ_{23} are Kendall's τ, pairwise rank correlation coefficients, we join together those X_i and X_j such that $\max_{i,j \in \{1,2,3\}, \, i \neq j} = \tau_{ij}$. Next we introduce $z = \hat{C}\{\hat{F}_i(X_i), \hat{F}_i(X_j)\}$. Estimation techniques will be considered later. Variable X_{i^*}, $i^* \in \{1, 2, 3\}/\{i, j\}$ is joined afterwards with the z.

17.2.5   Generalizations

There are numerous techniques which allow one to construct new types of copulae by using combinations of simple, elliptical or Archimedean copulae. For example, copula families B11 and B12 which is also known as a *Spearman* or *Cuadras-Augé* copula arise as a convex combination of the upper Fréchet–Hoeffding bound and the product copula

$$C_{B11}(u_1, u_2, \theta) = \theta M(u_1, u_3) + (1 - \theta)\Pi(u_1, u_2)$$
$$= \theta \min\{u_1, u_2\} + (1 - \theta)u_1 u_2,$$
$$C_{B12}(u_1, u_2, \theta) = M^\theta(u_1, u_2)\Pi^{1-\theta}(u_1, u_2)$$
$$= (\min\{u_1, u_2\})^\theta (u_1 u_2)^{1-\theta}, \quad u_1, u_2, \theta \in [0, 1].$$

Nelsen (2006) provides a list of methods of constructing multivariate copulae. This family of copulae arises from simple multivariate Archimedean copula from reparametrization $\lambda = e^{-\phi t}$:

$$C(u_1, \dots, u_d) = \lambda^{-1}\{\lambda(u_1) \dots \lambda(u_d)\} = \lambda^{-1}[\Pi\{\lambda(u_1), \dots, \lambda(u_d)\}].$$

The function λ is known as a *multiplicative generator* of C. Replacing product copula Π with an arbitrary copula C_1 of dimension d we get a new copula family, investigated in Morillas (2005).

Another popular and very flexible approach to modelling multivariate distributions is based on *vines* (Joe 1996). The idea is based on the decomposition of a multivariate density into $d(d-1)/2$ bivariate densities. For the D-vine copula model the density is given by

$$f(x_1,\ldots,x_d) =$$
$$= \prod_{m=1}^{d} f(x_m) \prod_{j=1}^{d-1} \prod_{i=1}^{d-j} c_{ji}\{F(x_i|x_{i+1},\ldots,x_{i+j-1}), F(x_{i+j}|x_{i+1},\ldots,x_{i+j-1})\},$$

where the conditional distribution is computed as a derivative with respect to the known arguments.

17.3   Monte Carlo Simulation

The Monte Carlo simulation is often a single reliable solution to many financial problems. Within the simulation study the random variables are generated from some prescribed distributions.

17.3.1   Conditional Inverse Method

The simulation from d pseudo random variables with joint distribution defined by a copula C and d marginal distributions F_j, $j = 1,\ldots,d$, may follow different techniques.

Defining the copula j-dimensional marginal distribution C_j for $j = 2,\ldots,d-1$ sas

$$C_j(u_1,\ldots,u_j) = C(u_1,\ldots,u_j,1,\ldots,1)$$

and the derivative of C_j with respect to the first $j-1$ arguments as

$$c_{j-1}^{j}(u_1,\ldots,u_j) = \frac{\partial^{j-1} C_j(u_1,\ldots,u_j)}{\partial u_1,\ldots,\partial u_{j-1}}$$

the probability $P(U_j \leq u_j, U_1 = u_1,\ldots,U_{j-1} = u_{j-1})$ can be written as

$$\lim_{\Delta u_1,\ldots,\Delta u_{j-1}\to 0} \frac{C_j(u_1+\Delta u_1,\ldots,u_{j-1}+\Delta u_{j-1},u_j) - C_j(u_1,\ldots,u_j)}{\Delta u_1,\ldots,\Delta u_{j-1}}$$
$$= c_{j-1}^{j}(u_1,\ldots,u_j).$$

Thus, the conditional distribution $\Lambda(u_j)$ (given fixed $u_1, \ldots, u_{j-1}$) is a function of the ratio of derivatives:

$$\Lambda(u_j) = P(U_j \le u_j \mid U_1 = u_1, \ldots, U_{j-1} = u_{j-1})$$

$$= \frac{P(U_j \le u_j, U_1 = u_1, \ldots, U_{j-1} = u_{j-1})}{P(U_1 = u_1, \ldots, U_{j-1} = u_{j-1})}$$

$$= \frac{c_{j-1}^j(u_1, \ldots, u_j)}{c_{j-1}^{j-1}(u_1, \ldots, u_{j-1})}.$$

The generation of d pseudo random numbers with given marginal distributions F_j, $j = 1, \ldots, d$ and dependence structure given by the copula C follows the steps:

1. generate pseudo random numbers $v_1, \ldots, v_d$ independent and uniformly distributed in $[0, 1]$.
2. for $j = 1, \ldots, d$ generate the pseudo random numbers as $u_j = \Lambda^{-1}(v_j)$. The pseudo random numbers $u_1, \ldots, u_d$ have uniform marginal distributions in $[0, 1]$ and dependence structure given by the copula C (Fig. 17.5).
3. set $x_j = F_j^{-1}(u_j)$. The pseudo random numbers $x_1, \ldots, x_d$ are distributed with the desired marginal distributions and dependence structure (Fig. 17.6).
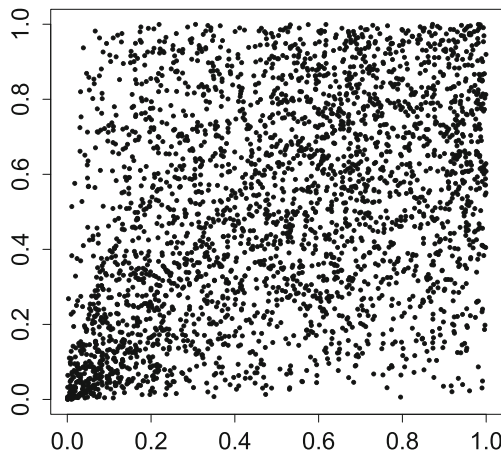
Fig. 17.5 Monte Carlo sample of 10,000 realizations of pseudo random variable with uniform marginals in $[0, 1]$ and dependence structure given by Clayton copula, $\theta = 0.79$

SFEclaytonMC
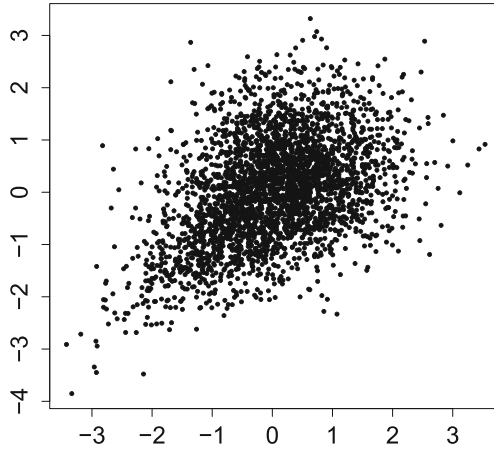
Fig. 17.6 Monte Carlo sample of 5,000 realizations of pseudo random variable with standard normal marginals and dependence structure given by Clayton copula with $\theta = 0.79$

Q SFEclaytonMC

If C is the Gaussian copula, the simulation is as follows:

1. generate pseudo random numbers $v_1, \ldots, v_d$ distributed as N$(0, \Psi)$
2. set $u_j = \Phi(v_j)$, $j = 1, \ldots, d$. The pseudo random numbers $u = (u_1, \ldots, u_d)$ have uniform marginal distributions in $[0, 1]$ and dependence structure given by C_Ψ^{Ga}.
3. set $x_j = F_j^{-1}(u_j)$. The pseudo random numbers $x_1, \ldots, x_d$ are distributed with the desired marginal distributions and dependence structure.

As already mentioned, if the marginal distributions are normal, the pseudo random numbers are multivariate normal distributed. Otherwise their distribution is called a Meta-Gaussian distribution.

If C is the t-copula, the simulation is as follows:

1. generate pseudo random numbers $v_1, \ldots, v_d$ distributed as $t_d(\nu, 0, \Psi)$
2. set $u_j = t_\nu(v_j)$, $j = 1, \ldots, d$ where t_ν is the univariate t distribution with ν degrees of freedom. The pseudo random numbers $u = (u_1, \ldots, u_d)$ have uniform marginal distributions in $[0, 1]$ and dependence structure given by $C_{\nu, \Psi}^t$.
3. set $x_j = F_j^{-1}(u_j)$. The pseudo random numbers $x_1, \ldots, x_d$ are distributed with the desired marginal distributions and dependence structure.

If the marginal distributions are t_ν, the pseudo random numbers are multivariate t distributed. Otherwise their distribution is called a *Meta-t* distribution.

Repeating one of the procedures above T times yields a Monte Carlo sample $\{x_{j,t}\}_{t=1}^T$, for $j = 1, \ldots, d$ of a random variable distributed as desired.
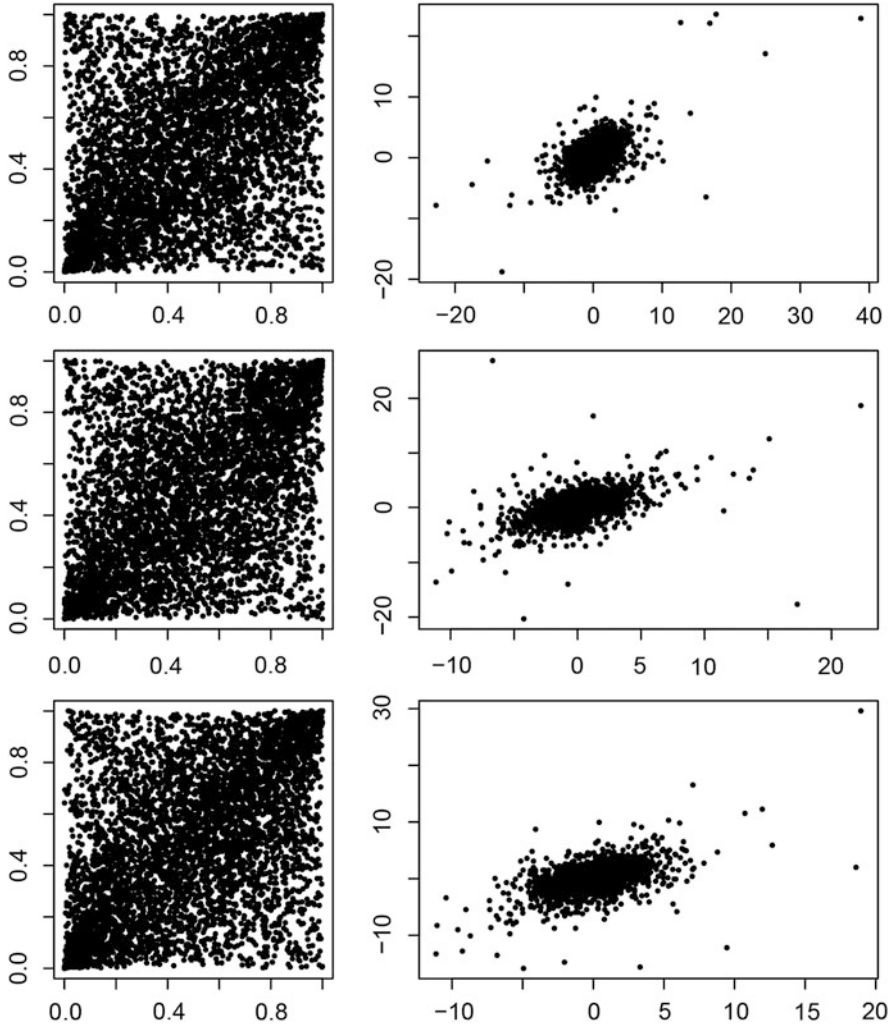
Fig. 17.7 Scatterplots of Monte Carlo sample (5,000 realizations) of pseudo random variable $X = (X_1, X_2, X_3)^\top$ with uniform (*left*) and t_3 marginal distributions (*right*). Dependence structure given by t-copula with $\nu = 3$ and $\psi_{i,j} = 0.5$, $i, j = 1, 2, 3$, $i \neq j$ [Q] SFEtMC

Scatterplots of Monte Carlo sample of pseudo random variable with uniform and t_3 marginal distributions and dependence structure given by t-copula are plotted in Fig. 17.7.

17.3.2   Marshal–Olkin Method

The idea of the Marshal–Olkin method is based on the fact that the Archimedean copulae are derived from Laplace transforms. Let M be a univariate cdf of a positive random variable (so that $M(0) = 0$) and ϕ be the Laplace transform of M, i.e.

$$\phi(s) = \int_0^\infty \exp\{-sw\}\, dM(w), \ s \geq 0.$$

For any univariate distribution function F, a unique distribution G exists:

$$F(x) = \int_0^\infty G^\alpha(x)\, dM(\alpha) = \phi\{-\log G(x)\}.$$

Considering d different univariate distributions $F_1, \ldots, F_d$, we obtain

$$C(u_1, \ldots, u_d) = \int_0^\infty \prod_{i=1}^d G_i^\alpha\, dM(\alpha) = \phi\left[\sum_{i=1}^d \phi^{-1}\{F_i(u_i)\}\right]$$

which is a multivariate distribution function. By replacing the product of univariate distributions G_i for $i = 1, \ldots, d$ with an arbitrary copula function R we get:

$$C(u_1, \ldots, u_d) = \int_0^\infty \cdots \int_0^\infty R(G_1^\alpha, \ldots, G_d^\alpha)\, dM(\alpha).$$

Note that for the classical Archimedean copula R is equal to a product copula.

One proceeds with the following three steps to make a draw from a distribution described by an Archimedean copula:

1. generate an observation u from M;
2. generate an observations $(v_1, \ldots, v_d)$ from R;
3. the generated vector is computed by $x_j = G_j^{-1}(v_j^{1/u})$.

This method works faster than the conditional inverse technique. The drawback is that the distribution M can be determined explicitly only for a few generator functions ϕ like, for example for the Frank, Gumbel and Clayton families. The same problem arises in the case of hierarchical copulae, where $\phi_i \circ \phi_{i+1}^{-1}$ should satisfy the properties of generator functions.

17.4   Copula Estimation

The estimation of a copula based multivariate distribution involves both the estimation of the copula parameters θ and the estimation of the margins F_j, $j = 1, \ldots, d$, however all the parameters from the copula and from the margins could be also

estimated in one step. The properties and goodness of the estimator of θ heavily depend on the estimators of F_j, $j = 1, \ldots, d$. We distinguish between a parametric and a non-parametric specification of the margins. If we are interested only in the dependency structure, the estimator of $\{\delta_1, \ldots, \delta_d, \theta\}$ should be independent of any parametric models for the margins. In practical applications, however, we are interested in a complete distribution model and, therefore, parametric models for margins are preferred.

For non-parametrically estimated margins, one may show the consistency and asymptotic normality of maximum-likelihood (ML) estimators and derive the moments of the asymptotic distribution. The ML estimation can be performed simultaneously for the parameters of the margins and of the copula function. Alternatively, a two-stage procedure can be applied, where we estimate the parameters of margins at the first stage and the copula parameters at the second stage.

Let X be a d-dimensional random variable with parametric univariate marginal distributions $F_j(x_j; \delta_j)$, $j = 1, \ldots, d$. Further let a copula belong to a parametric family $\mathcal{C} = \{C_\theta, \theta \in \Theta\}$. The distribution of X can be expressed as

$$F_X(x_1, \ldots, x_d) = C\{F_1(x_1; \delta_1), \ldots, F_d(x_d; \delta_d); \theta\}$$

and its density as

$$f(x_1, \ldots, x_d; \delta_1, \ldots, \delta_d, \theta) = c\{F_1(x_1; \delta_1), \ldots, F_d(x_d; \delta_d); \theta\} \prod_{j=1}^{d} f_j(x_j; \delta_j)$$

where

$$c(u_1, \ldots, u_d) = \frac{\partial^d C(u_1, \ldots, u_d)}{\partial u_1 \ldots \partial u_d}.$$

For a sample of observations $\{x_t\}_{t=1}^{T}$, $x_t = (x_{1,t}, \ldots, x_{d,t})^\top$ and a vector of parameters $\alpha = (\delta_1, \ldots, \delta_d, \theta)^\top \in \mathbb{R}^{d+1}$ the likelihood function is given by

$$L(\alpha; x_1, \ldots, x_T) = \prod_{t=1}^{T} f(x_{1,t}, \ldots, x_{d,t}; \delta_1, \ldots, \delta_d, \theta)$$

and the log-likelihood function by

$$\ell(\alpha; x_1, \ldots, x_T) = \sum_{t=1}^{T} \log c\{F_1(x_{1,t}; \delta_1), \ldots, F_d(x_{d,t}; \delta_d); \theta\}$$

$$+ \sum_{t=1}^{T} \sum_{j=1}^{d} \log f_j(x_{j,t}; \delta_j).$$

The vector of parameters $\alpha = (\delta_1, \ldots, \delta_d, \theta)^\top$ contains d parameters δ_j from the marginals and the copula parameter θ. All these parameters can be estimated *in one step*. For practical applications, however, a two-step estimation procedure is more efficient.

17.4.1   Full Maximum Likelihood Estimation

In the Maximum Likelihood estimation method [also called *full maximum likelihood* (FML)], the vector of parameters α is estimated in one single step through

$$\tilde{\alpha}_{FML} = \arg\max_{\alpha} \ell(\alpha)$$

The estimates $\tilde{\alpha}_{FML} = (\tilde{\delta}_1, \ldots, \tilde{\delta}_d, \tilde{\theta})^\top$ solve

$$(\partial\ell/\partial\delta_1, \ldots, \partial\ell/\partial\delta_d, \partial\ell/\partial\theta) = 0.$$

Following the standard theory on ML estimation it is efficient and asymptotically normal. However, it is often computationally demanding to solve the system simultaneously.

17.4.2   Inference for Margins

In the IFM (*inference for margins*) method, the parameters δ_j from the marginal distributions are estimated in the first step and used to estimate the dependence parameter θ in the second step:

1. for $j = 1, \ldots, d$ the log-likelihood function for each of the marginal distributions is

$$\ell_j(\delta_j) = \sum_{t=1}^{T} \log f_j(x_{j,t}; \delta_j)$$

   and the estimated parameters

$$\hat{\delta}_j = \arg\max_{\delta} \ell_j(\delta_j)$$

2. the *pseudo log-likelihood* function

$$\ell(\theta, \hat{\delta}_1, \ldots, \hat{\delta}_d) = \sum_{t=1}^{T} \log c\{F_1(x_{1,t}; \hat{\delta}_1), \ldots, F_d(x_{d,t}; \hat{\delta}_d); \theta\}$$

is maximized over θ to get the dependence parameter estimate $\hat{\theta}$.

The estimates $\hat{\alpha}_{IFM} = (\hat{\delta}_1, \ldots, \hat{\delta}_d, \hat{\theta})^{\top}$ solve

$$(\partial \ell_1/\partial \delta_1, \ldots, \partial \ell_d/\partial \delta_d, \partial \ell/\partial \theta) = 0.$$

Detailed discussion on this method could be found in Joe and Xu (1996). Note, that this procedure does not lead to efficient estimators, however, as argued by Joe (1997) the loss in the efficiency is modest. The advantage of the IFM procedure lies in the dramatic reduction of the numerical complexity.

17.4.3   *Canonical Maximum Likelihood*

In the CML (*canonical maximum likelihood*) method, the univariate marginal distributions are estimated through the edf $\hat{F}$. The asymptotic properties of the multistage estimators of θ do not depend explicitly on the type of the non-parametric estimator, but on its convergence properties. For $j = 1, \ldots, d$

$$\hat{F}_j(x) = \frac{1}{T+1} \sum_{t=1}^{T} \mathbf{1}(x_{j,t} \leq x).$$

The *pseudo log-likelihood* function is

$$\ell(\theta) = \sum_{t=1}^{T} \log c\{\hat{F}_1(x_{1,t}), \ldots, \hat{F}_d(x_{d,t}); \theta\}$$

and the copula parameter estimator $\hat{\theta}_{CML}$ is given by

$$\hat{\theta}_{CML} = \arg \max_{\theta} \ell(\theta).$$

Notice that the first step of the IFM and CML methods estimates the marginal distributions. After marginals are estimated, a *pseudo sample* $\{u_t\}$ of observations transformed in the unit d-cube is obtained and used in the *copula* estimation. As in the IFM, the semiparametric estimator $\hat{\theta}$ is asymptotically normal under suitable regularity conditions.

17.4.4 *Gaussian Copula Estimation*

From a sample $\{u_t\}_{t=1}^T$ where $u = (u_1, \ldots, u_d)^\top \in [0,1]^d$, the density of the Gaussian copula is given by

$$c_\Psi^{Ga}(u_1, \ldots, u_d) = |\Psi|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\zeta^\top(\Psi^{-1} - I_d)\zeta\right\}$$

and the pseudo log-likelihood function by

$$\ell(\Psi; u_{1,t}, \ldots, u_{d,t}) = -\frac{T}{2}\log|\Psi| - \frac{1}{2}\sum_{t=1}^{T}\zeta_t^\top(\Psi^{-1} - I_d)\zeta_t$$

where $\zeta_t = (\zeta_{1,t}, \ldots, \zeta_{d,t})^\top$ and $\zeta_{j,t} = \Phi^{-1}(u_{j,t})$.

The maximum-likelihood estimator for Ψ is

$$\hat{\Psi} = \arg\max_{\Psi \in \mathcal{P}} \ell(\Psi)$$

where $\mathcal{P}$ is the set of all lower-triangular matrices with one in the diagonal. The maximization is feasible but very slow for high dimensions. An approximate solution can be obtained using the ML estimator for the covariance matrix Σ as

$$\hat{\Sigma} = \arg\max_{\Sigma} \ell(\Sigma).$$

The estimator is then

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}\zeta_t\zeta_t^\top$$

and defining

$$\Lambda = \text{diag}(\hat{\Sigma}_{ii})$$

we obtain

$$\hat{\Psi} = \Lambda^{-1}\hat{\Sigma}\Lambda^{-1}.$$

17.4.5 *t*-Copula Estimation

One possible estimation method for the Student's *t*-copula is based on the estimation of Kendall's τ with method of moments. For a pseudo sample $\{u_t\}_{t=1}^T$ where $u = (u_1, \ldots, u_d)^\top \in [0, 1]^d$, Kendall's τ coefficient for each pair of observations $i, j = 1, \ldots, d$ is given by

$$\hat{\rho}_\tau(u_i, u_j) = \binom{T}{2}^{-1} \sum_{1 \le t_1 \le t_2 \le T} \text{sign}(u_{i,t_1} - u_{i,t_2})(u_{j,t_1} - u_{j,t_2}).$$

Each element from the correlation matrix Ψ is estimated as

$$\hat{\psi}_{ij} = \sin\left\{\frac{\pi}{2}\hat{\rho}_\tau(u_i, u_j)\right\}.$$

The parameter ν is estimated through maximum-likelihood with the estimated matrix $\hat{\Psi}$ held fixed. In this case the pseudo log-likelihood function is given by

$$\ell(\nu; u_{1,t}, \ldots, u_{d,t}) = \sum_{t=1}^T \log\{c_{\nu,\hat{\Psi}}^t(u_{1,t}, \ldots, u_{d,t})\}$$

where $c_{\nu,\hat{\Psi}}^t(u_{1,t}, \ldots, u_{d,t})$ is defined in Eq. (17.5). The estimator for the number of degrees of freedom is then

$$\hat{\nu} = \underset{\nu \in \mathbb{N}_+}{\arg\max}\, \ell(\nu).$$

17.5   Asset Allocation

We illustrate the extension of the classical asset allocation problem to copula-based models. We consider an investor with a CRRA utility function $U(x) = (1 - \gamma)^{-1}x^{1-\gamma}$ willing to allocate his wealth to d risky assets. We denote the d-dimensional vector of d asset prices by $S_t = (S_{1,t}, \ldots, S_{d,t})^\top$ and their continuously compounded asset returns at time $t + 1$ by $X_{t+1} = (X_{1,t+1}, \ldots, X_{d,t+1})^\top$ where $X_{t+1} = \log S_{t+1} - \log S_t$. The vector of portfolio weights by $w = (w_1, \ldots, w_d)^\top$. Let F_{t+1} be the d-dimensional distribution function of X_{t+1} with the mean μ_{t+1} and covariance matrix Σ_{t+1}. The aim is to forecast F_{t+1} for the time period $t + 1$ using the data up to time t. The estimator is denoted by $\hat{F}_{t+1}$ with the mean $\hat{\mu}_{t+1}$, the covariance matrix $\hat{\Sigma}_{t+1}$ and the density $\hat{f}_{t+1}$. The objective of the investor is to maximize the expected utility at the time point $t + 1$. This leads to the optimization problem

$$\max_{w \in \mathcal{W}} \mathsf{E}_{\hat{F}_{t+1}} U(1 + w^\top X_{t+1}). \tag{17.7}$$

In the case of no short sales constraint we set $\mathcal{W} = \{w \in [0, 1]^d : w^\top 1 = 1\}$ else we set $\mathcal{W} = \{w \in \mathbb{R}^d : w^\top 1 = 1\}$. The conditional expectation in (17.7) implies that we integrate the utility with respect to the forecasted distribution $\hat{F}_{t+1}$. This reduces the problem (17.7) to the problem

$$\max_{w \in \mathcal{W}} \int \cdots \int U(1 + w^\top X_{t+1}) \hat{f}_{t+1}(X_{t+1}) dX_{t+1}.$$

There are several alternative parametric approaches to modelling F_{t+1}. Let $\Sigma_{d,t+1}$ denote the diagonal matrix containing only the main diagonal of Σ_{t+1}. Then $\Sigma_{t+1} = \Sigma_{d,t+1}^{1/2} R_{t+1} \Sigma_{d,t+1}^{1/2}$, where R_{t+1} denotes the correlation matrix. A standard approach is to define the model of the asset returns in the form

$$\Sigma_{d,t}^{-1/2}(X_t - \text{---}_t) \sim N_d(0, R_t), \tag{17.8}$$

where the conditional moments μ_t and Σ_t are modelled by a GARCH type process.

To introduce a copula-based distribution into the asset allocation we deviate from the normality assumption and assume that $F = C(F_1, \ldots, F_d)$. Thus (17.8) is replaced by:

$$\Sigma_{d,t}^{-1/2}(X_t - \text{---}_t) \sim C(F_1, \ldots, F_d) \tag{17.9}$$

with some given functional forms of the copula and the marginal distributions. Similarly as above, the parameters of the conditional moments of the copula and of the marginal distributions are estimated using the ML method.

17.6   Value-at-Risk of the Portfolio Returns

If the return of the stock i at time point t is denoted as X_{it} then the portfolio value V at time t is defined recursively as

$$V_t = V_{t-1} \left(1 + \sum_{i=1}^{d} w_i X_{it}\right),$$

where w_i for $i = 1, \ldots, d$ are the corresponding portfolio weights. Ruled with this notation the portfolio return is then given by

$$R_{tp} = \frac{V_t}{V_{t-1}} - 1 = \sum_{i=1}^{d} X_{it} w_i.$$

In our study we consider the case of equally weighted portfolio, i.e. $w_i = \frac{1}{d}$ for $i = 1, \ldots, d$. The portfolio return is the random variable and its distribution strongly depends on the underlying distribution of the indices.

The distribution function of R_p, dropping the time index, is given by

$$F_{R_p}(\xi) = \mathrm{P}(R_p \le \xi). \tag{17.10}$$

One of the main advantages of copulae is the fact that they allow flexible modelling of the tail behaviour of multivariate distributions. Since the tail behaviour explains the simultaneous outliers of asset returns, it is of special interest in risk management. The *Value-at-Risk* of a portfolio at level α is defined as the lower α-quantile of the distribution of the portfolio return, i.e.

$$\mathrm{VaR}(\alpha) = F_{R_p}^{-1}(\alpha). \tag{17.11}$$

The VaR is a reasonable measure of risk if we assume that the returns are elliptically distributed. Moreover, the assumption of ellipticity implies that minimizing the variance in the Markowitz problem also minimizes the VaR, the expected shortfall and any other coherent measure of risk. However, this statement is false in the non-elliptical case. Moreover, regarding the effect of diversification the variance is the smallest (highest) for perfect negative (positive) correlation of the assets. This also holds for the VaR in the elliptical case, however, not for the non-elliptical distributions. This implies that for copula based distribution the VaR should be used with caution and its computation should be awarded more attention.

Our aim is to determine such ξ that $\mathrm{P}(R_p \le \xi) = \alpha$. Note that

$$R_p = w^\top X = \sum_{i=1}^{d} w_i X_i = \sum_{i=1}^{d} w_i F_i^{-1}(u_i),$$

where F_i denotes the marginal distributions of individual asset returns, $u_i = F_i(X_i) \sim U[0, 1]$ for all $i = 1, \ldots, d$ and $u_1, \ldots, u_d \sim C$. The copula C defines the dependency structure between the asset returns. This implies that

$$F_{R_p}(\xi) = \mathrm{P}(R_p \le \xi) = \int_{\mathcal{U}} c(u_1, \ldots, u_d) du_1 \ldots du_d, \tag{17.12}$$

with

$$\mathcal{U} = \{[0, 1]^{d-1} \times [0, u_d(\xi)]\}, \quad u_d(\xi) = F_d \Big\{ \xi/w_d - \sum_{i=1}^{d-1} w_i F_i^{-1}(u_i)/w_d \Big\}. \tag{17.13}$$

For fixed α, the VaR is determined by solving (17.12) numerically for ξ. Direct multidimensional numerical integration is a tedious task which can be substantially

simplified by using the Monte Carlo integration. For this purpose we have to generate random samples from C using the methods described in Sect. 17.3.

In the empirical study we consider four countries Canada, Germany, the USA and UK from the MCSI index and eleven models of the joint multivariate distribution of indices, which include t-copula, Gaussian copula, simple exchangeable Archimedean copula, binary HAC and aggregated binary HAC, with normally and t-distributed margins. As a benchmark we use the empirical VaR, based purely on the real data.

In the cases where margins are t-distributed, we consider t-distribution with three degrees of freedom, while estimated t-distributions for this data are $t_{3.163}$, $t_{3.420}$, $t_{3.023}$, $t_{2.879}$. Multivariate t-copula in this case has eight degrees of freedom. Let us consider the simulation procedure, where on the first stage we estimate the covariance matrix $\hat{\Sigma} = \{\hat{\Sigma}_{ij}\}_{i,j=1,\dots,d}$, mean vector $\hat{\mu} = \{\hat{\mu}_i\}_{i=1,\dots,d}$ from the real data set and assume, or estimate, the marginal distributions $\hat{F}_i(\cdot)$ (in our case they are normally or t-distributed), for $i = 1,\dots,d$. Next we show how to sample $u_1,\dots,u_d \in \mathcal{U}$ from (17.13). First we simulate the vector u of a dimension $d-1$

$$u_1,\dots,u_{d-1} \sim U(0,1).$$

Based on u we consider $x = \{x_i\}_{i=1,\dots,d-1}$ which for normal margins is equal to

$$x_i = \Phi^{-1}(u_i)\sqrt{\hat{\Sigma}_{ii}} + \hat{\mu}_i, \quad i = 1,\dots,d-1,$$

and for t margins is

$$x_i = t^{-1}(u_i)\sqrt{\frac{v_i-2}{v_i}\hat{\Sigma}_{ii}} + \hat{\mu}_i, \quad i = 1,\dots,d-1,$$

where v_i, $i = 1,\dots,d$ are degrees of freedom for marginal distributions. This transformation returns a normally or t-distributed vector x with the same parameters as the real data set.

Theoretically, in further steps we have to find bounds for the last stock (or index) to gain the portfolio ξ which is the α quantile. Thus, we separate our maximally reachable portfolio return ξ into two parts

$$\xi = \sum_{i=1}^{d-1} \frac{1}{d} X_i + \frac{1}{d} X_d,$$

then the return of the last index given the return of the portfolio is

$$X_d = d\xi - \sum_{i=1}^{d-1} X_i,$$

where the upper bound for our last value in vector u is then

$$u_d^* = \hat{F}_d \left(d\xi - \sum_{i=1}^{d-1} x_i \right).$$

Value u_d^* is uniformly distributed on $[0, 1]$ and we simulate the last element of the vector $u_d \sim U(0, u_d^*)$.

As mentioned above, the goal is to compute (17.12) which for this setting is

$$F_{R_p}(\xi) = \int \cdots \int_{[0,1]^{d-1} \times [0,u_d^*]} c(u_1, \ldots, u_d) du_1 \ldots du_d.$$

Then by solving $F_{R_p}(\xi) = \alpha$ we find $R_\alpha = \mathrm{VaR}(\alpha)$. In our study we solve the equations numerically using the golden section method. The integration is performed using the Monte Carlo technique

$$\widehat{P(R_p \leq \xi)} = \frac{1}{n_s} \sum_{i=1}^{n_s} c(u_{1i}, \ldots, u_{di})$$

where n_s is equal to 10^8, α is set to be 1% and the values $u_{1i}, \ldots, u_{di}$ for $i = 1, \ldots, n_s$ are simulated using the method described above. The precision of R is set at 0.00015. The final results for all methods are given in Table 17.1. In the left-hand column we provide the models with normal margins and in the right-hand column with t margins. From top to bottom we have five different copula functions like Gaussian, t, simple Archimedean copula, binary HAC and binary aggregated HAC. The empirical VaR which is at the bottom of the table is derived from the empirical quantile. Bold fonts in the table emphasize those results which are closest in absolute value to the empirical one in each column, and italic fonts the worst cases in absolute value.

As can be seen from Table 17.1, the results which are the best in absolute value are those returned by the model with t-copula and t margins. The model based on the simple Archimedean copula is the worst one. This is quite natural, since this copula needs exchangeability between variables, which is not observable here (see previous section). HAC with binary as well as aggregated binary structures, unfortunately, give us results that are not much worse compared to t-copula and Gaussian copula. For $\mathrm{VaR}(0.01)$ the t-copula with t margins provided the best result.

Table 17.1 VaR for the four-dimensional data set

	N	t_3
N	−0.0194	−0.0210
t_8	**−0.0199**	**−0.0213**
AC	*−0.0174*	*−0.0154*
HAC_{binary}	−0.0187	−0.0194
$HAC_{binary\ aggr.}$	−0.0188	−0.0194
Empirical	−0.0235	

17.6.1   VaR of the P&L

This sub-section introduces the main assumptions and steps necessary to estimate the VaR from a Profit and Loss of a linear portfolio using copulae. Static and time-varying methods and their VaR performance evaluation through backtesting are described below.

In this section w is the portfolio, which is represented by the number of assets for a specified stock in the portfolio, $w = \{w_1, \ldots, w_d\}$, $w_i \in \mathbb{Z}$. The value V_t of the portfolio w is given non-recursively by

$$V_t = \sum_{j=1}^{d} w_j S_{j,t} \tag{17.14}$$

and the random variable

$$L_{t+1} = (V_{t+1} - V_t)$$

$$= \sum_{j=1}^{d} w_j S_{j,t} \left\{ \exp(X_{j,t+1}) - 1 \right\}.$$

also called *profit and loss (P&L) function*, expresses the absolute change in the portfolio value in one period.

Similarly to the previous case, the distribution function of L, dropping the time index, is given by

$$F_L(x) = P(L \leq x). \tag{17.15}$$

As usual the *Value-at-Risk* at level α from a portfolio w is defined as the α-quantile from F_L:

$$\mathrm{VaR}(\alpha) = F_L^{-1}(\alpha). \tag{17.16}$$

It follows from (17.15) that F_L depends on the d-dimensional distribution of log-returns F_X. In general, the *loss distribution* F_L depends on a random process representing the *risk factors* influencing the P&L from a portfolio. In the present case log-returns are a suitable risk factor choice. Thus, modelling their distribution is essential to obtain the quantiles from F_L.

Contrary to the previous section, here log-returns are assumed to be time-dependent, thus a log-returns process $\{X_t\}$ can be modelled as

$$X_{j,t} = \mu_{j,t} + \sigma_{j,t}\varepsilon_{j,t}$$

where $\varepsilon_t = (\varepsilon_{1,t}, \ldots, \varepsilon_{d,t})^\top$ are standardized *i.i.d.* innovations with $\mathsf{E}[\varepsilon_{j,t}] = 0$ and $\mathsf{E}[\varepsilon_{j,t}^2] = 1$ for $j = 1, \ldots, d$; $\mathcal{F}_t$ is the available information at time t;

$$\mu_{j,t} = E[X_{j,t} \mid \mathcal{F}_{t-1}]$$

is the conditional mean given $\mathcal{F}_{t-1}$ and

$$\sigma_{j,t}^2 = E[(X_{j,t} - \mu_{j,t})^2 \mid \mathcal{F}_{t-1}]$$

is the conditional variance given $\mathcal{F}_{t-1}$. The innovations $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)^\top$ have joint distribution

$$F_\varepsilon(\varepsilon_1, \ldots, \varepsilon_d) = C_\theta\{F_1(\varepsilon_1), \ldots, F_d(\varepsilon_d)\} \tag{17.17}$$

where C_θ is a copula belonging to a parametric family $\mathcal{C} = \{C_\theta, \theta \in \Theta\}$, and F_j, $j = 1, \ldots, d$ are continuous marginal distributions of ε_j. To obtain the Value-at-Risk in this set up, the dependence parameter and distribution function from residuals are estimated from a sample of log-returns and used to generate P&L Monte Carlo samples. Their quantiles at different levels are the estimators for the Value-at-Risk.

For a portfolio w on d assets and a sample $\{x_{j,t}\}_{t=1}^T$, $j = 1, \ldots, d$ of log-returns, the Value-at-Risk at level α is estimated according to the following steps:

1. Estimation of residuals $\hat{\varepsilon}_t$ from the prespecified time-series model;
2. Specification and estimation of marginal distributions $F_j(\hat{\varepsilon}_j)$;
3. Specification of a parametric copula family $\mathcal{C}$ and estimation of dependence parameter θ;
4. Generation of Monte Carlo sample of innovations ε and losses L, for the forecast on the one day;
5. Estimation of $\widehat{VaR}(\alpha)$, the empirical α-quantile from the forecasted L.

The application of the (*static*) procedure described above on sliding windows of a time series $\{x_{j,t}\}_{t=1}^T$ delivers a sequence of parameters for a copula family. Hence the denomination *time-varying copulae*.

Using moving windows of size r in time t

$$\{x_t\}_{t=s-w+1}^{s}$$

for $s = r, \ldots, T$, the procedure described in the section above generates the time series $\{\widehat{VaR}_t\}_{t=r}^{T}$ of Value-at-Risk and $\{\hat{\theta}_t\}_{t=r}^{T}$ dependence parameters estimates.

Afterwards *Backtesting* is used to evaluate the performance of the specified copula family $\mathcal{C}$. The estimated values for the VaR are compared with the true realizations $\{l_t\}$ of the P&L function, an *exceedance* occurring for each l_t smaller than $\widehat{VaR}_t(\alpha)$. The ratio of the number of exceedances to the number of observations gives the *exceedances ratio* $\hat{\alpha}$:

$$\hat{\alpha} = \frac{1}{T-r} \sum_{t=r}^{T} \mathbf{1}\{l_t < \widehat{VaR}_t(\alpha)\}.$$

The estimation methods described before are used on two portfolio, the first composed of two positions, the second of three positions. Different copulae are used in static and dynamic setups and their VaR performance is compared based on backtesting.

In this section, the Value-at-Risk of portfolios for two companies (Tyssenkrupp (TKA) and Volkswagen (VOW) from 01.12.1997 to 03.07.2007) is computed using different copulae.

Assuming the log-returns $\{X_{j,t}\}$ follow a GARCH(1,1) process we have

$$X_{j,t} = \mu_{j,t} + \sigma_{j,t}\varepsilon_{j,t}$$

where

$$\sigma_{j,t}^2 = \omega_j + \alpha_j \sigma_{j,t-1}^2 + \beta_j (X_{j,t-1} - \mu_{j,t-1})^2$$

and $\omega > 0$, $\alpha_j \geq 0$, $\beta_j \geq 0$, $\alpha_j + \beta_j < 1$.

The fit of a GARCH(1,1) model to the sample of log returns $\{x_t\}_{t=1}^{T}$, $X_t = (X_{1,t}, X_{2,t})^\top$, $T = 2,500$, gives the estimates $\hat{\omega}_j$, $\hat{\alpha}_j$ and $\hat{\beta}_j$, as in Table 17.2, and empirical residuals $\{\hat{\varepsilon}_t\}_{t=1}^{T}$, where $\hat{\varepsilon}_t = (\hat{\varepsilon}_{1,t}, \hat{\varepsilon}_{2,t})^\top$. The marginal distributions are

Table 17.2   GARCH(1,1) parameters, two-dimensional portfolio

	$\hat{\omega}_j$	$\hat{\alpha}_j$	$\hat{\beta}_j$
$j = 1$	0.00	0.07	0.93
$j = 2$	0.00	0.09	0.88

Table 17.3   Parameters from
marginal distributions

	$\hat{\mu}_j$	$\hat{\sigma}_j$
$j = 1$	-0.0149	0.9976
$j = 2$	-0.0193	0.9987
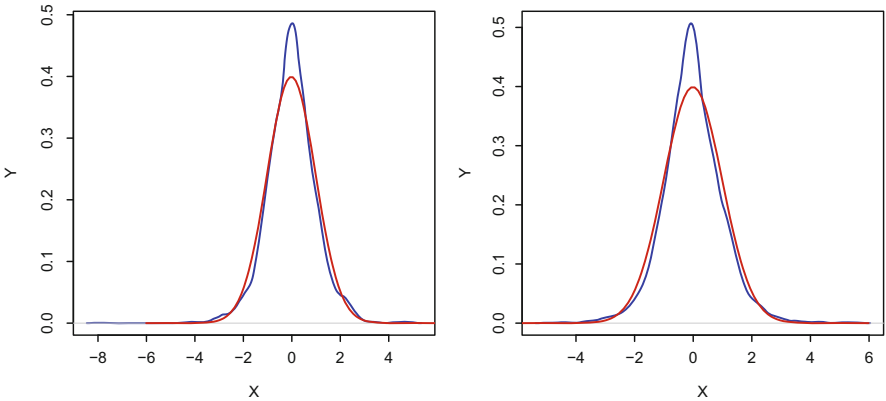
Fig. 17.8   Kernel density estimator of the residuals and of the normal density from TKA (*left*) and
VOW (*right*). Quartic kernel, $\hat{h} = 2.78\hat{\sigma}n^{-0.2}$ SFEresDens

Table 17.4   Dependence parameter
for different static copulae

Copula	$\hat{\theta}$
Gaussian	0.462
Clayton	0.880
Gumbel	1.439

specified as normal, i.e. $\hat{\varepsilon}_j \sim N(\hat{\mu}_j, \hat{\sigma}_j)$ with parameters $\hat{\delta}_j = (\hat{\mu}_j, \hat{\sigma}_j)$ estimated
from the data as in Table 17.3.

   Figure 17.8 displays the Kernel density estimator of the residuals and of the
normal density, estimated with an Quartic kernel. The dependence parameters
are estimated (Table 17.4) for different copula families (Gaussian, Clayton and
Gumbel). Residuals $\hat{\varepsilon}$ and fitted copulae (Gaussian, Clayton and Gumbel) are plotted
in Fig. 17.9.

   In the dynamic approach, the empirical residuals are sampled in moving windows
with a fixed size $r = 250$, $\{\hat{\varepsilon}_t\}_{t=s-r+1}^s$, for $s = r, \ldots, T$. The time series from
estimated dependence parameters for each copula family are in Fig. 17.10.
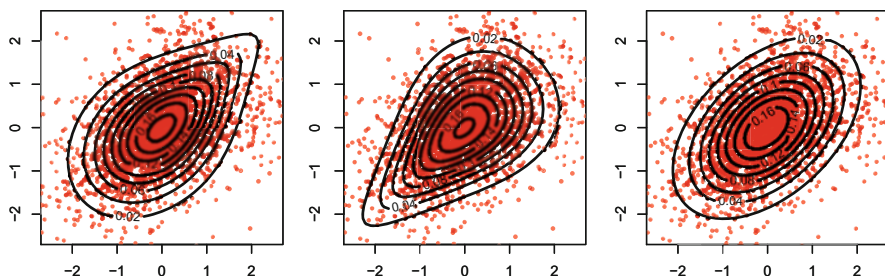
Fig. 17.9 Residuals $\hat{\varepsilon}$ and fitted copulae: Gumbel ($\hat{\theta} = 1.439$), Clayton ($\hat{\theta} = 0.880$), Gaussian ($\hat{\rho} = 0.462$)
                                                                    Q SFEstaticCop

The same portfolio compositions as in the static case are used to generate P&L samples. The series of estimated Value-at-Risk and the P&L function for selected portfolios are plotted in Figs. 17.11, 17.12 and 17.13.

17.6.2 Three-Dimensional Portfolio

In this section, the Value-at-Risk of portfolios composed of three positions (Merck (MRK), Tyssenkrupp (TKA) and Volkswagen (VOW) from 01.12.1997 to 03.07.2007) is computed using a time-varying simple Gumbel copula and time-varying hierarchical Archimedean copula with generators from the Gumbel family.

The estimation of the parameters of the three-dimensional copula was done by the IFM method. Concerning the HAC, we determine the structure under each window and re-estimate the parameters (Fig. 17.14).

The fit of a GARCH(1,1) model to the sample of log returns $\{X_t\}_{t=1}^{T}$, $X_t = (X_{1,t}, X_{2,t}, X_{3,t})^{\top}$, $T = 2{,}500$, gives the estimates $\hat{\omega}_j$, $\hat{\alpha}_j$ and $\hat{\beta}_j$, as in Table 17.5, and empirical residuals $\{\hat{\varepsilon}_t\}_{t=1}^{T}$, where $\hat{\varepsilon}_t = (\hat{\varepsilon}_{1,t}, \hat{\varepsilon}_{2,t}, \hat{\varepsilon}_{3,t})^{\top}$, as in upper right part of Fig. 17.15. The marginal distributions are specified as normal, $\hat{\varepsilon}_j \sim \mathrm{N}(\hat{\mu}_j, \hat{\sigma}_j)$, the estimated parameters $\hat{\delta}_j = (\hat{\mu}_j, \hat{\sigma}_j)$ are in Table 17.6.

The estimated Value-at-Risk at level α together with the P&L function is plotted in Fig. 17.16 for the simple archimedean copula (AC) and in Fig. 17.17 for the HAC. As can be seen from the backtesting results for different VaR levels, HAC outperforms the simple AC in all levels. This implies the necessity of dependence flexibility in modelling of log-returns.

To conclude, a summary of the main findings of this chapter. Three different copulae—Gumbel, Clayton and Gaussian—were used to estimate the Value-at-Risk from the two-dimensional portfolio (MRK and TKA). From the time series of
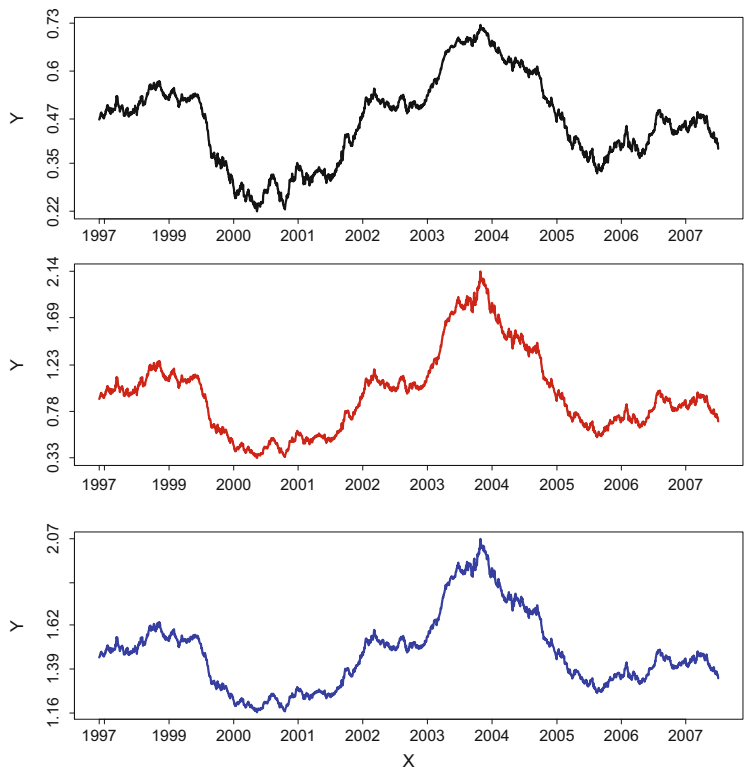
Fig. 17.10   Dependence parameter $\hat{\theta}$, estimated using the IFM method, Gaussian (*upper panel*), Gumbel (*middle panel*) and Clayton (*lower panel*) copulae, moving window ($w = 250$)

Q SFEdynCop
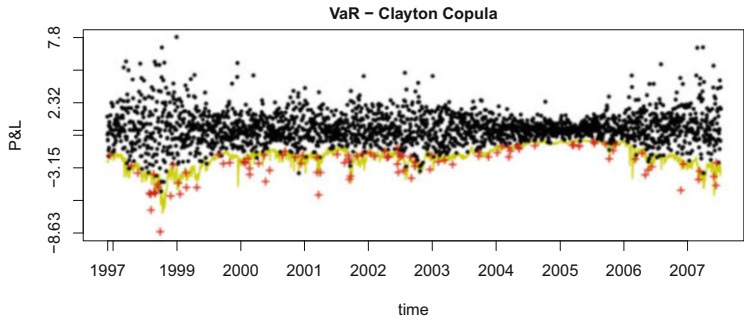
Fig. 17.11 $\widehat{VaR}(\alpha)$ (*solid line*), P&L (*dots*) and exceedances (*crosses*), $\alpha = 0.05$, $\hat{\alpha} = 0.0424$. P&L samples generated with Clayton copula

Q SFEVaRcopulaSIM2ptv
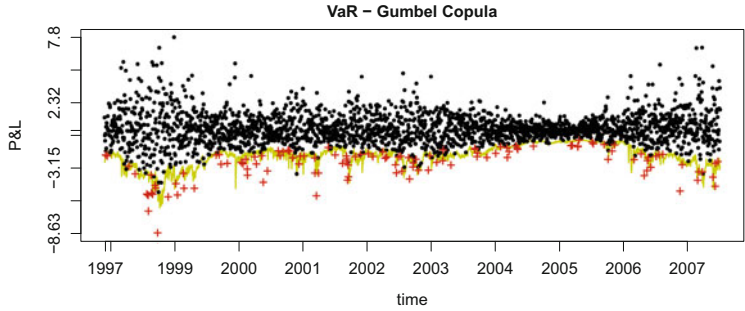
Fig. 17.12 $\widehat{VaR}(\alpha)$ (*solid line*), P&L (*dots*) and exceedances (*crosses*), $\alpha = 0.05$, $\hat{\alpha} = 0.0508$. P&L samples generated with Gumbel copula   Q SFEVaRcopulaSIM2ptv
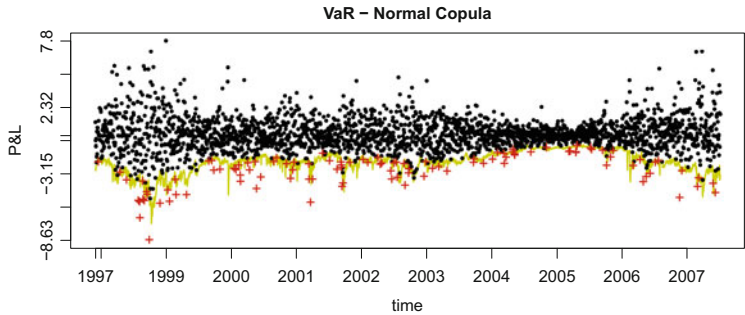
Fig. 17.13 $\widehat{VaR}(\alpha)$ (*solid line*), P&L (*dots*) and exceedances (*crosses*), $\alpha = 0.05$, $\hat{\alpha} = 0.0464$. P&L samples generated with Gaussian copula   Q SFEVaRcopulaSIM2ptv

estimated dependence parameters, we can verify that the dependence structure is represented in a similar form with all copula families, as in Fig.17.10.

Using backtesting results to compare the performance in the VaR estimation, we remark that on average the Clayton and Gaussian copulae *overestimate* the VaR. In terms of capital requirement, a financial institution computing VaR with those copulae would be requested to keep *more* capital aside than necessary to guarantee the desired confidence level.

The estimation with Gumbel copula, on another side, produced results close to the desired level. Gumbel copulae seems to represent specific data dependence structures (like lower tail dependencies, relevant to explain simultaneous losses) better than Gaussian and Clayton copulae.
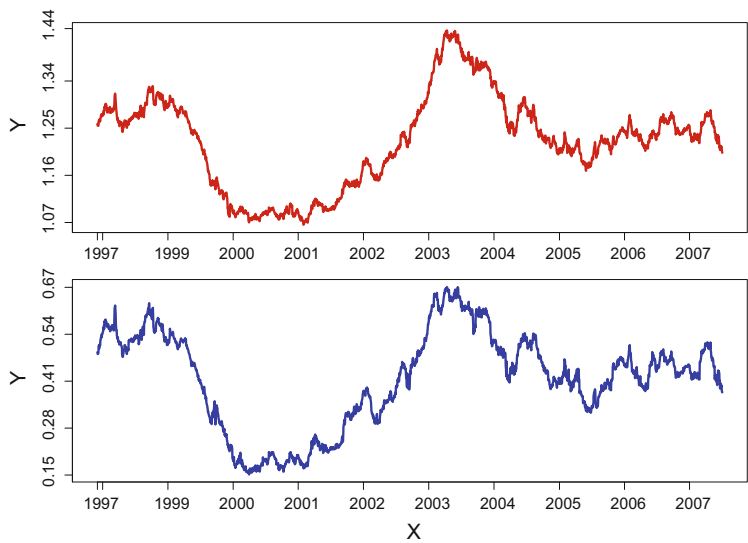
Fig. 17.14 Dependence parameter $\hat{\theta}$, estimated using the IFM method, Clayton (*upper panel*) and Gumbel (*lower panel*) copulae, moving window ($w = 250$)   **Q** SFEdynCop

Table 17.5   GARCH(1,1) parameters, three-dimensional portfolio

	$\hat{\omega}$	$\hat{\alpha}$	$\hat{\beta}$
$j = 1$	0.00	0.03	0.96
$j = 2$	0.00	0.07	0.93
$j = 3$	0.00	0.09	0.88

17.7   Recommended Literature

The term copula goes back to the works of Sklar (1959). A detailed discussion with proofs and deep mathematical treatment can be found in Joe (1997) and Nelsen (2006).

A practical introduction is given in Deutsch and Eller (1999). Embrechts, McNeil and Straumann (1999b) discuss restrictions of the copula technique and their relation to the classical correlation analysis.

Different techniques for the simulation from d pseudo random variables with joint distribution defined by a copula C and d marginal distributions are used. A standard method for Archimedean copulae (the conditional distribution method) as well as other different methods is described in Bouyé, Durrleman, Nikeghbali, Riboulet and Roncalli (2000), Devroye (1986) and Embrechts, Frey and McNeil (2005).
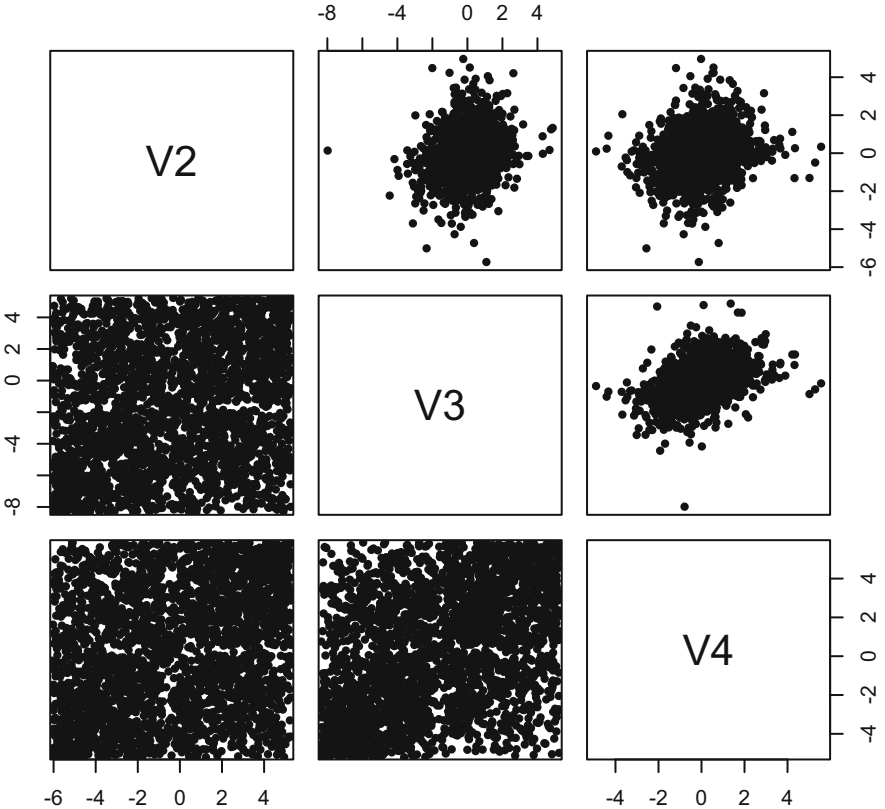
Fig. 17.15 Scatterplots from GARCH residuals (*upper triangular*) and from residuals mapped on unit square by the cdf (*lower triangular*)   SFE5dim

Table 17.6 Parameters from marginal distributions

	$\hat{\mu}$	$\hat{\sigma}$
$j = 1$	-0.0149	0.9976
$j = 2$	-0.0193	0.9987
$j = 3$	-0.0104	0.9992

Whelan (2004) provides tools for generating samples from Archimedean copulae, Savu and Trede (2010) derived the density of such copulae and Joe (1997) proves their positive quadrant dependence (see Theorem 4.4). Okhrin et al. (2013) and Okhrin, Okhrin and Schmid (n.d.) considered methods for determining the optimal structure of the HAC, provided asymptotic theory for the estimated parameters and derive theoretical properties of this copula family.
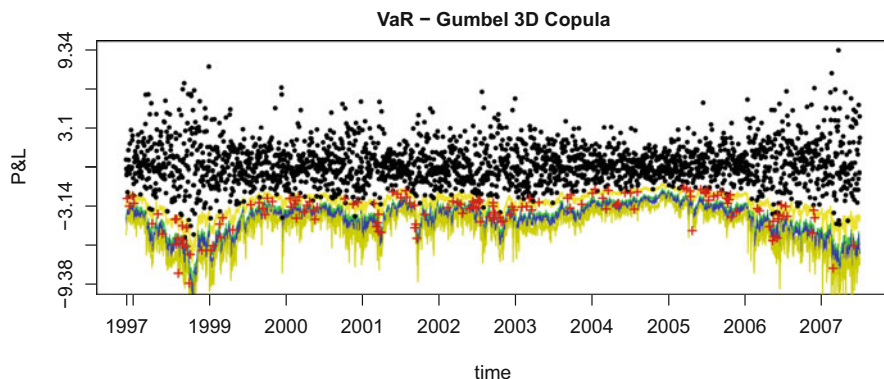
Fig. 17.16 $\widehat{VaR}(\alpha)$ and P&L (*dots*), estimated with three-dimensional simple Gumbel copula, $\alpha_1 = 0.05$ ($\hat{\alpha}_1 = 0.0612$), $\alpha_2 = 0.01$ ($\hat{\alpha}_2 = 0.0232$), $\alpha_3 = 0.005$ ($\hat{\alpha}_3 = 0.016$) and $\alpha_4 = 0.001$ ($\hat{\alpha}_4 = 0.006$)   🔲 SFEVaRHAC
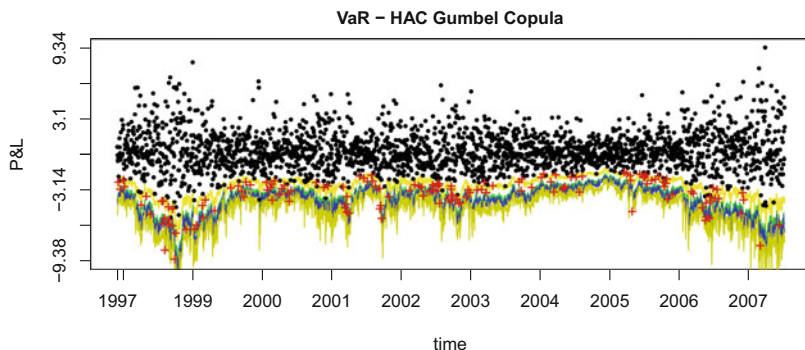
Fig. 17.17 $\widehat{VaR}(\alpha)$ and P&L (*dots*), estimated with three-dimensional HAC with Gumbel generators, $\alpha_1 = 0.05$ ($\hat{\alpha}_1 = 0.0592$), $\alpha_2 = 0.01$ ($\hat{\alpha}_2 = 0.0208$), $\alpha_3 = 0.005$ ($\hat{\alpha}_3 = 0.014$) and $\alpha_4 = 0.001$ ($\hat{\alpha}_4 = 0.004$)   🔲 SFEVaRHAC

Detailed description of the VaR estimation procedure at prescribed level α can be found in Giacomini and Härdle (2005).

Detailed discussion on the IFM procedure can be found in Joe and Xu (1996). Note, that this method does not lead to efficient estimators, however, as argued by Joe (1997) the loss in the efficiency is modest.

The theory for copula model selection tests is developed in Chen and Fan (2006a), Chen and Fan (2006b) and Chen, Fan and Tsyrennikov (2004) for static set up.

17.8   Exercises

Exercise 17.1 *Are the following functions for $a, b \in [0, 1]$ valid copula functions?*

(a) $C_1(a, b) = a^2/2 + b^2/2 - (a - b)^2/2$
(b) $C_2(a, b) = (|a| + |b| - |a - b|)/2$

Exercise 17.2 *Let X_1, X_2 be identically distributed (but not independent) random variables with cdf F. Define the random variables $U_i = 1 - F(X_i)$ for $i = 1, 2$ and the joint distribution of $(U_1, U_2)^\top$ be given with copula function C. Calculate the joint distribution of $(X_1, X_2)^\top$, i.e. $P(X_1 \leq s, X_2 \leq t)$.*

Exercise 17.3 *One method of generating random numbers from any copula is the conditional distribution method. Consider a pair of the uniform random variables (U, V) with copula C. Show that*

$$P(V \leq v | U = u) = \frac{\partial}{\partial u} C(u, v).$$

Exercise 17.4 *Let U and V be two uniform random variables whose joint distribution function is a Clayton copula C. Calculate the inverse of the conditional distribution $V | U$.*

Exercise 17.5 *Calculate coefficients of upper and lower tail dependence for Gumbel copula.*

Chapter 18
Statistics of Extreme Risks

When we model returns using a GARCH process with normally distributed inno-
vations, we have already taken into account the second *stylized fact* (see Chap. 13).
The distribution of the random returns automatically has a leptokurtosis and larger
losses occurring more frequently than under the assumption that the returns are
normally distributed. If one is interested in the 95 %-VaR of liquid assets, this
approach produces the most useful results. For the extreme risk quantiles such as
the 99 %-VaR and for riskier types of investments the risk is often underestimated
when the innovations are assumed to be normally distributed, since there is a higher
probability of particularly extreme losses than a GARCH process ε_t with normally
distributed Z_t can produce.

Thus procedures have been developed which assume that the tails of the innova-
tions' distribution are heavier. The probability of extreme values largely depends on
how slowly the probability density function $f_Z(x)$ of the innovations goes to 0 as
$|x| \to \infty$. The rate at which it diminishes must be estimated from the data. Since
extreme observations are rare, this produces a difficult estimation problem. Even
large data sets contain only limited information on the true probability of an extreme
loss (profit). In such a situation methods from extreme value statistics produce a
more realistic estimate of the risk. In this chapter a short overview of the basic ideas
and several of the latest applications are given.

18.1 Risk Measures

Measuring risk means summarizing a distribution of risk factors with a number
known as a risk measure, see McNeil (1999). The mean or standard deviation, for
instance, measure some aspects but do not provide information about the extreme
risk. The Value-at-Risk (VaR) is the standard measure for downside risk, and the
elements that define it are the maximum probable loss on a given portfolio over a

period of time, given some specific confidence level. It provides an upper limit for a loss, which is exceeded only on a small proportion of occasions. An alternative risk measure is introduced in this section—the expected shortfall (ES).

Definition 18.1 (Quantile Function)   If F is a distribution function, we call the generalized inverse

$$F^{-1}(\gamma) = \inf\{x \in \mathbb{R};\ F(x) \geq \gamma\},\ 0 < \gamma < 1,$$

the *quantile function*. It then holds that $P\{X_1 \leq F^{-1}(\gamma)\} = \gamma$, i.e. $F^{-1}(\gamma)$ is the γ-*quantile* of the distribution F.

If F is strictly monotonic increasing and continuous, then F^{-1} is the generalized inverse of F.

Definition 18.2 (Value-at-Risk and Expected Shortfall)   Let $0 < q < 1$, and let F be the distribution of the loss X of a financial investment within a given time period, for example, 1 day or 10 trading days. Typical values for q are $q = 0.95$ and $q = 0.99$.

(a) The *Value-at-Risk* (VaR) is the q-quantile

$$VaR_q(X) = x_q = F^{-1}(q).$$

(b) The *expected shortfall* (S_q) is defined as

$$S_q = \mathsf{E}\{X|X > x_q\}.$$

The most commonly used measurement today for quantifying market risk is the Value-at-Risk. However, in the future it is very likely that the expected shortfall will be at least as important.

Definition 18.3 (Coherent Risk Measure)   A *coherent risk measure* is a real-valued function $\rho : \mathbb{R} \to \mathbb{R}$ of real-valued random variables, which models the losses, with the following characteristics:

(A1) $X \geq Y$ a.s. $\Longrightarrow \rho(X) \geq \rho(Y)$ (*Monotonicity*)
(A2) $\rho(X + Y) \leq \rho(X) + \rho(Y)$ (*Subadditivity*)
(A3) $\rho(\lambda X) = \lambda \rho(X)$ for $\lambda \geq 0$ (*Positive homogeneity*)
(A4) $\rho(X + a) = \rho(X) + a$ (*Translation equivariance*)

These characteristics correspond to intuitive requirements of a market risk measurement:

(A1)   When the loss from investment X is always larger or equal to the loss from investment Y, then the risk from investment X is also larger or equal to the risk from investment Y.

(A2)   The risk of a portfolio consisting of investments in X and Y is at most as large as the sum of the individual risks (diversification of the risk).

(A3)   When an investment is multiplied, then the risk is also multiplied accordingly.

(A4)   By adding a risk free investment, i.e. a non-random investment with known losses a ($a < 0$, when the investment has fixed payments), to a portfolio, the risk changes by exactly a.

The VaR does not meet condition (A2) in certain situations. Let X and Y, for example, be i.i.d. and both can take on the value 0 or 100 with probabilities $P(X = 0) = P(Y = 0) = p$ and $P(X = 100) = P(Y = 100) = 1 - p$. Then $X + Y$ can be 0, 100 and 200 with probabilities $P(X + Y = 0) = p^2$, $P(X + Y = 100) = 2p(1 - p)$ and $P(X + Y = 200) = (1 - p)^2$, respectively. For $p^2 < q < p$ and $q < 1 - (1 - p)^2$, e.g. for $p = 0.96, q = 0.95$, it holds that

$$VaR_q(X) = VaR_q(Y) = 0, \quad \text{but} \quad VaR_q(X + Y) = 100.$$

The expected shortfall, on the other hand, is a coherent risk measure that always fulfills all four conditions. It also gives a more intuitive view of the actual risk of extreme losses than the Value-at-Risk. The VaR only depends on the probability of losses above the q-quantile x_q, but it does not say anything about whether these losses are always just a little above the threshold x_q or whether there are also losses that are much larger than x_q that need to be taken into account. In contrast the expected shortfall is the expected value of the potential losses above x_q and depends on the actual size of the losses.

The Value-at-Risk is simply a quantile and can be, for example, estimated as a sample quantile $\hat{F}_n^{-1}(q)$, where $\hat{F}_n(x)$ is the empirical distribution of a sample of negative values, i.e. losses, from the past. This particular estimator of $q \approx 1$, which is for the typical VaR-level of 0.95 and 0.99, is often too optimistic. Alternative VaR estimators, which have the possibility of reflecting extreme losses better, are the peaks-over-threshold (POT) and the Hill quantile estimators. Analogous estimators for the expected shortfall can be also derived.

18.2   Data Description

The methods of Extreme Value Theory (EVT) that are described in the following sections are illustrated with the help of programs written in Matlab software.

The data is represented by the closing stock prices of four DAX companies, namely Bayer, BMW, Siemens and Volkswagen, for the period 1 January 2002–31 December 2012, on daily basis. For each company, there are 3,360 observations, collected from Monday to Friday. For non-trading days, the prices considered were the previous ones. A number of 3,288 trading days were recorded on the Frankfurt Stock Exchange (FSE) for the period under review. The source for the

Fig. 18.1 Closing prices of stocks: Bayer, BMW, Siemens, Volkswagen. Time period:1 January 2002–31 December 2012

data is Datastream. From the initial dataset, an equally weighted portfolio is built, consisting of one of each of the four stocks. Figure 18.1 plots the closing prices of Bayer, BMW, Siemens and Volkswagen.

The daily log returns for the companies are computed according to the formula:

$$r_t = \log\left(\frac{P_t}{P_{t-1}}\right),$$

where P_t is the level of the stock price on day t.

Figure 18.2 displays the log returns graphically. The scale is the same for all companies in order to make the results comparable. The daily log returns usually fall within the boundaries of a 15 % decrease and a 15 % increase, respectively, with the exception of some outliers displayed by the log returns of the Bayer, Siemens and Volkswagen stocks. The plot of the log returns of the portfolio is displayed in Fig. 18.3. The volatility, as expected, is visibly lower than in the individual cases, reflecting diversification.

Table 18.1 provides descriptive statistics for the log returns of the individual companies and the portfolio, respectively. Considering the standard deviation as a measure for volatility, the Siemens log returns display the largest volatility, despite the fact that Bayer has larger outliers. The boundaries between minimum and maximum are narrowest and also the volatility is the lowest in the case of the equally-weighted portfolio. In terms of profitability, the average log return is positive in all cases.
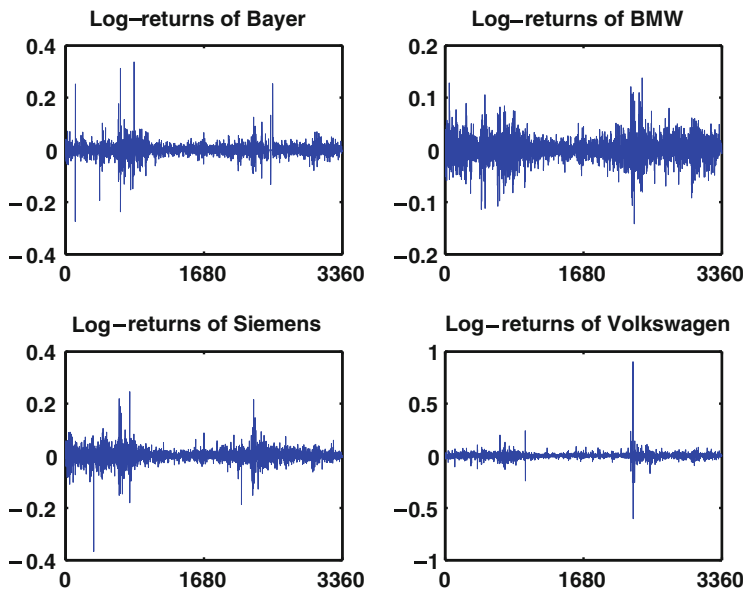
Fig. 18.2 Daily log returns for Bayer, BMW, Siemens and Volkswagen. Time period: 1 January 2002–31 December 2012   SFEcomplogreturns
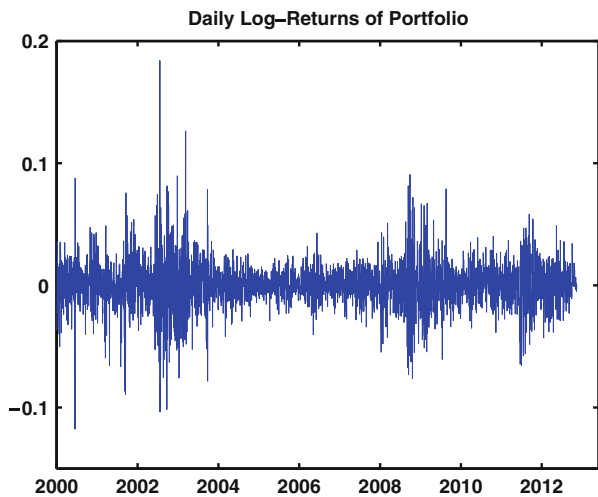
Fig. 18.3 Daily log returns of the portfolio of Bayer, BMW, Siemens and Volkswagen. Time period: 1 January 2002–31 December 2012   SFEportlogreturns

Table 18.1 Summary statistics for the log return daily data

Name	Minimum	Maximum	Mean	Median	Standard deviation
Bayer	−0.2746	0.3370	0.0001	0.0000	0.0253
BMW	−0.1416	0.1381	0.0002	0.0000	0.0226
Siemens	−0.3673	0.2460	0.0001	0.0000	0.0264
Volkswagen	−0.6031	0.9026	0.0003	0.0000	0.0339
Portfolio	−0.1176	0.1839	0.0001	0.0002	0.0185

Time period: 1 January 2002–31 December 2012

Figure 18.4 shows the plot of the probability distribution function and the quantile plot for the pdf $F(x) = \Phi(x)$ of the standard normal, for the portfolio of Bayer, BMW, Siemens and Volkswagen. The deviations from the straight line show that the data is not normally distributed.

Figure 18.5 displays the kernel density estimate of the positive and negative log returns of the portfolio, respectively. The reason for considering negative log returns is that while one is interested in the loss side of the distribution, the risk measures are expressed as positive numbers. So basically, when trying to estimate the Value-at-Risk and the expected shortfall, the losses are depicted as positive quantities, and the profits as negative ones.

18.3 Estimation Methods

In this section, the main methods used for estimating the Value-at-Risk and the expected shortfall will be presented. According to McNeil (1999), there are two main types of models that deal with extreme values.

The first group and the oldest consists of the *block maxima models*. They represent models designed for the largest observations from samples of identically distributed observations.

The second group of models is considered to be more modern and it consists of the *peaks-over-threshold* (POT) models. They concern all the large observations which exceed a certain threshold. Within the POT class of models, one may choose between two different approaches. One is fully parametric and based on the generalized Pareto distribution (GPD), while the other is semi-parametric and is centred around the Hill estimator. We regard the POT models as being generally the most useful in practice, since they are able to use the often limited data available on extreme values in a more efficient way.
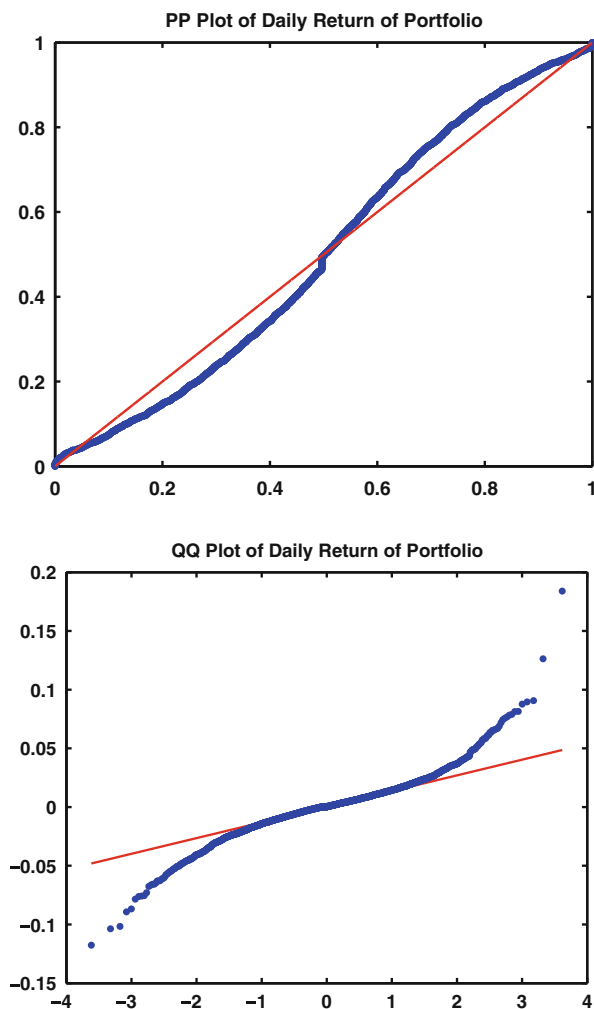
Fig. 18.4 PP plot and QQ plot of the portfolio of Bayer, BMW, Siemens and Volkswagen. Time period: 1 January 2002–31 December 2012   SFEportfolio

18.3.1   *The Block Maxima Method*

Consider the stochastic behaviour of the maximum $M_n = \max(X_1, \ldots, X_n)$ of n identically distributed random variables $X_1, \ldots, X_n$ with cumulative distribution function (cdf) $F(x)$. From a risk management perspective $X_t = -Z_t$ is the negative return at day t. The cdf of M_n is

Density Estimate of the Positive Log–Returns of the Portfolio

Density Estimate of the Negative Log–Returns of the Portfolio
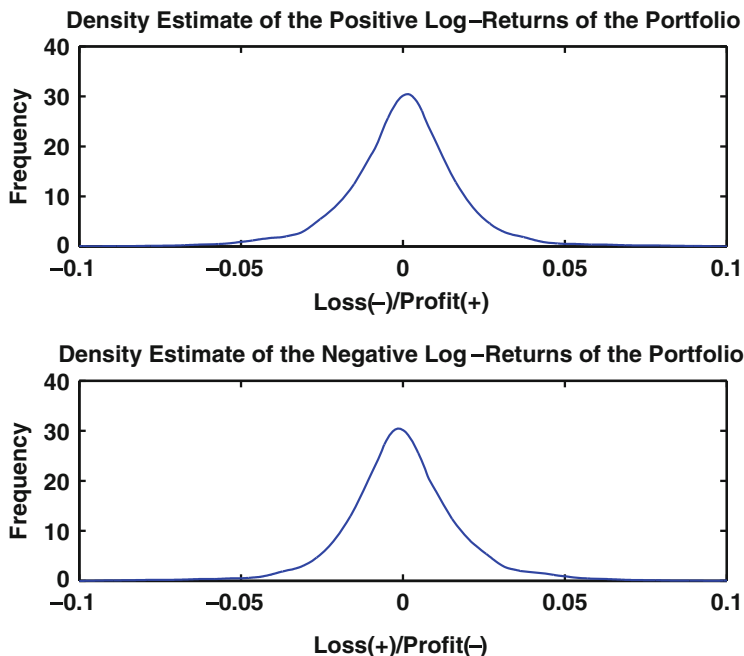
Fig. 18.5 Kernel density estimate of the positive and negative log returns of the portfolio of Bayer, BMW, Siemens and Volkswagen. Time period: 1 January 2002–31 December 2012

Q SFEdenport

$$P(M_n \leq x) = P(X_1 \leq x, \dots, X_n \leq x) = \prod_{t=1}^{n} P(X_t \leq x) = F^n(x). \qquad (18.1)$$

Only unbounded random variables X_t are considered, i.e. $F(x) < 1$ for all $x < \infty$. Obviously it holds that $F^n(x) \to 0$ for all x, when $n \to \infty$, and thus $M_n \xrightarrow{P} \infty$.

Figure 18.6 illustrates the concept of block maxima, where the block size is $n = 3$.

The maximum of n unbounded random variables may become arbitrary large. In order to achieve a non-degenerate behaviour limit, M_n has to be standardized in a suitable fashion. For an analysis of asymptotics one needs the limit law for the block maxima M_n.

Definition 18.4 (Maximum Domain of Attraction) The random variable X_t belongs to the *maximum domain of attraction* (MDA) of a non-degenerate distribution G, if for suitable sequences $c_n > 0, d_n$ it holds that:

$$\frac{M_n - d_n}{c_n} \xrightarrow{\mathcal{L}} G \quad \text{for } n \to \infty,$$

i.e. $F^n(c_n x + d_n) \to G(x)$ at all continuity points x of the cdf $G(x)$.
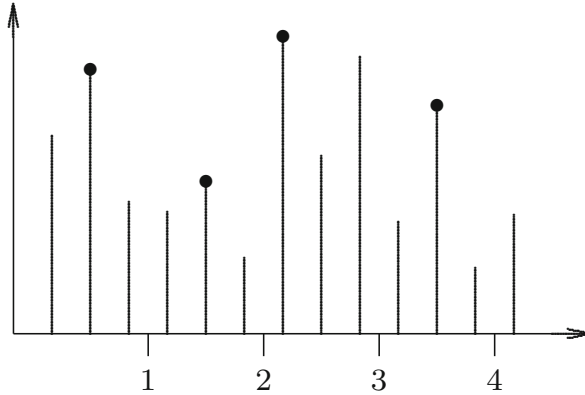
Fig. 18.6 Block-maxima

It turns out that only a few distributions G can be considered as the asymptotic limit distribution of the standardized maximum M_n. They are referred to as the *extreme value distributions*. These are the following three distribution functions:

$$\text{Fréchet}: \quad G_{1,\alpha}(x) = \exp\{-x^{-\alpha}\}, \ x \geq 0 , \ \text{for } \alpha > 0,$$

$$\text{Gumbel}: \quad G_0(x) = \exp\{-e^{-x}\}, \ x \in \mathbb{R},$$

$$\text{Weibull}: \quad G_{2,\alpha}(x) = \exp\{-|x|^{-\alpha}\}, \ x \leq 0 , \ \text{for } \alpha < 0.$$

The Fréchet distributions are concentrated on the non-negative real numbers $[0, \infty)$, while the Weibull distribution, on the other hand, on $(-\infty, 0]$. The Gumbel distributed random variables can attain any real number. Figure 18.7 displays the density function of the Gumbel distribution, the Fréchet distribution with parameter $\alpha = 2$ and the Weibull distribution with parameter $\alpha = -2$. Jenkinson and von Mises suggested a parametric representation for the three standard distributions:

Definition 18.5 (Generalized Extreme Value Distributions) The *generalized extreme value distribution* (GEV = *generalised extreme value*) with the *form parameter* $\gamma \in \mathbb{R}$ has the distribution function:

$$G_\gamma(x) = \exp\{-(1 + \gamma x)^{-1/\gamma}\}, \ 1 + \gamma x > 0 \text{ for } \gamma \neq 0$$
$$G_0(x) = \exp\{-e^{-x}\}, \ x \in \mathbb{R}$$

G_0 is the Gumbel distribution, whereas $G_\gamma, \gamma \neq 0$ is linked to the Fréchet- and Weibull distributions by the following relationships:

$$G_\gamma\left(\frac{x-1}{\gamma}\right) = G_{1,1/\gamma}(x) \text{ for } \gamma > 0,$$

Extreme Value Densities
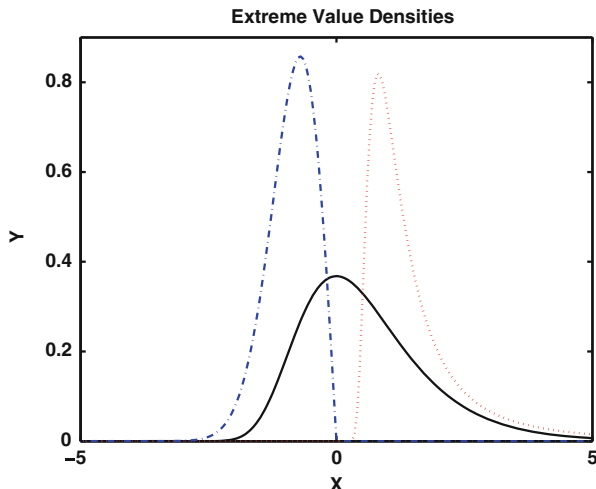
Fig. 18.7 Gumbel distribution (*solid line*), Fréchet distribution with parameter $\alpha = 2$ (*dotted line*) and Weibull distribution with parameter $\alpha = -2$ (*dash-dot line*)   Q SFEevt1

$$G_\gamma \left(-\frac{x+1}{\gamma} \right) = G_{2,-1/\gamma}(x) \text{ for } \gamma < 0.$$

This definition describes the standard form of the GEV distributions. In general the centre and the scale can be changed in order to obtain other GEV distributions: $G(x) = G_\gamma \left(\frac{x-\mu}{\sigma} \right)$ with the form parameter γ, the location parameter $\mu \in \mathbb{R}$ and the scale parameter $\sigma > 0$. For asymptotic theory this does not matter since the standardized sequences c_n, d_n can be always chosen so that the asymptotic distribution G has the standard form ($\mu = 0, \sigma = 1$). McNeil (1999) considers the GEV to be the natural limit distribution for normalized maxima.

An important result of the asymptotic distribution of the maximum M_n is the *Fisher–Tippett theorem*:

Theorem 18.1 (Fisher and Tippett (1928)) *If sequences $c_n > 0, d_n$ exist and a non-degenerate distribution G, so that*

$$\frac{M_n - d_n}{c_n} \xrightarrow{\mathcal{L}} G \quad \text{for } n \to \infty,$$

then G is a GEV distribution.

Proof As a form of clarification the basic ideas used to prove this central result are outlined. Let $t > 0$, and $[z]$ represent the integer part of z. Since $F^{[nt]}$ is the distribution function of $M_{[nt]}$, due to our assumptions on the asymptotic distribution of M_n it holds that

$$F^{[nt]}(c_{[nt]}x + d_{[nt]}) \longrightarrow G(x) \text{ for } [nt] \to \infty, \text{ i.e. } n \to \infty.$$

On the other hand it also holds that

$$F^{[nt]}(c_n x + d_n) = \{F^n(c_n x + d_n)\}^{[nt]/n} \longrightarrow G^t(x) \text{ for } n \to \infty.$$

In other words this means that

$$\frac{M_{[nt]} - d_{[nt]}}{c_{[nt]}} \xrightarrow{\mathcal{L}} G , \qquad \frac{M_{[nt]} - d_n}{c_n} \xrightarrow{\mathcal{L}} G^t$$

for $n \to \infty$. According to the Lemma, which is stated below, this is only possible when

$$\frac{c_n}{c_{[nt]}} \longrightarrow b(t) \geq 0, \qquad \frac{d_n - d_{[nt]}}{c_{[nt]}} \longrightarrow a(t)$$

and

$$G^t(x) = G\{b(t)x + a(t)\}, \ t > 0, \ x \in \mathbb{R}. \tag{18.2}$$

This relationship holds for arbitrary values t. We use it in particular for arbitrary t, s and $s \cdot t$ and obtain

$$b(st) = b(s) b(t), \ a(st) = b(t)a(s) + a(t). \tag{18.3}$$

The functional equations (18.2), (18.3) for $G(x), b(t), a(t)$ have only one solution, when G is one of the distributions $G_0, G_{1,\alpha}$ or $G_{2,\alpha}$, that is, G must be a GEV distribution.   □

Lemma 18.1 (Convergence Type Theorem)   *Let $U_1, U_2, \ldots, V, W$ be random variables, $b_n, \beta_n > 0$, $a_n, \alpha_n \in \mathbb{R}$. If*

$$\frac{U_n - a_n}{b_n} \xrightarrow{\mathcal{L}} V$$

in distribution for $n \to \infty$, then it holds that:

$$\frac{U_n - \alpha_n}{\beta_n} \xrightarrow{\mathcal{L}} W \qquad \textit{if and only if} \qquad \frac{b_n}{\beta_n} \longrightarrow b \geq 0, \quad \frac{a_n - \alpha_n}{\beta_n} \longrightarrow a \in \mathbb{R}.$$

In this case W has the same distribution as $bV + a$.

Notice that the GEV distributions are identical to the so-called *max-stable* distributions, by which for all $n \geq 1$ the maxima M_n of n i.i.d. random variables
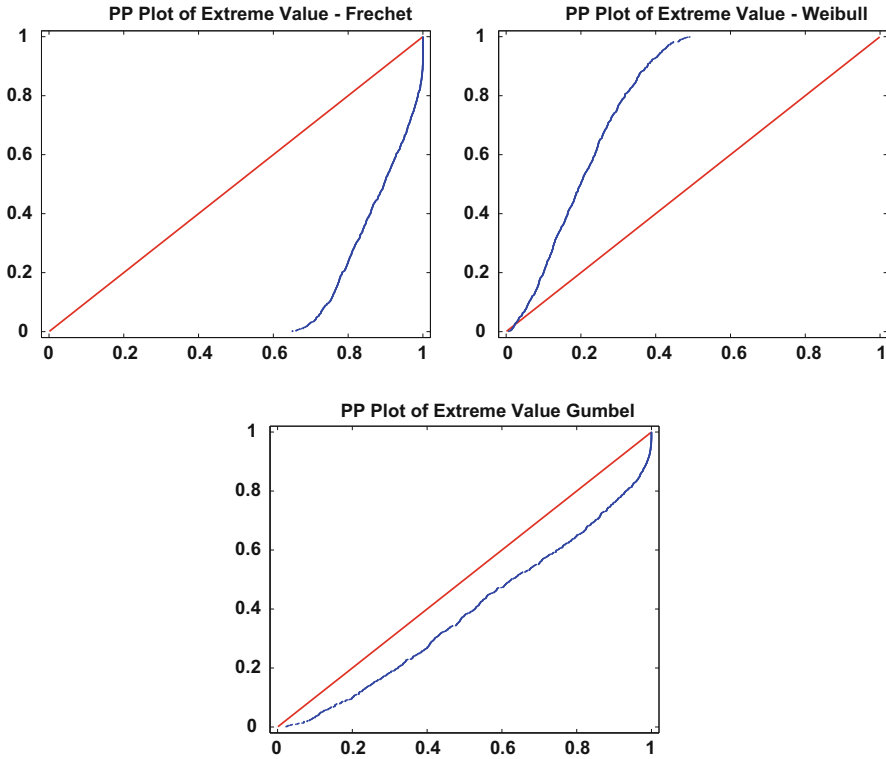
Fig. 18.8 PP plot for the normal distribution and pseudo random variables with extreme value distributions. Fréchet (*upper left*), Weibull (*upper right*) and Gumbel (*below*)   ☒ SFEevt2

$X_1, \ldots, X_n$ have the same distribution as $c_n X_1 + d_n$ for appropriately chosen $c_n > 0, d_n$.

Figure 18.8 shows the normal plot for the special case $F(x) = \Phi(x)$ with computer generated random variables that have a Gumbel distribution, Fréchet distribution with parameter $\alpha = 2$ and a Weibull distribution with parameter $\alpha = -2$, respectively. The differences with the normally distributed random variables, which would have approximately a straight line in a normal plot, can be clearly seen. Figure 18.9 compares the cdf of the pseudo random variables from the Gumbel distribution against the theoretical Gumbel distribution.

Identifying the Type of the Limit (GEV) Distribution

If the maximum of i.i.d. random variables converges in distribution after being appropriately standardized, then the question arises which of the three GEV distributions is the asymptotic distribution. The deciding factor is how fast the probability for extremely large observations decreases beyond a threshold x, when x increases. Since this exceedance probability plays an important role in EVT, some
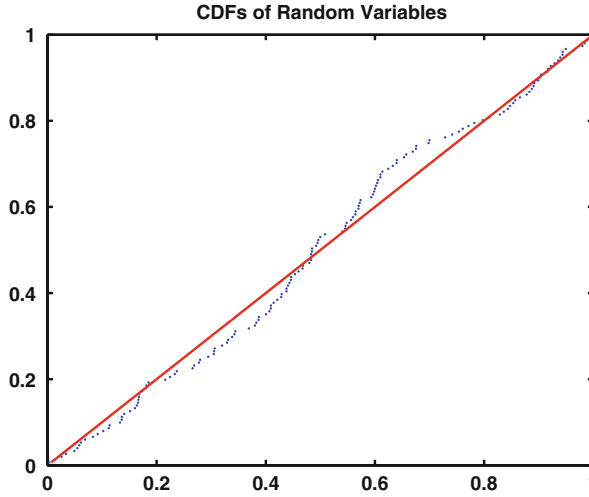
Fig. 18.9 PP plot of the pseudo random variables with Gumbel distribution against theoretical Gumbel distribution   SFEevt3

more notations are introduced:

$$\overline{F}(x) = \mathrm{P}(X_1 > x) = 1 - F(x).$$

The relationship between the exceedance probability $\overline{F}(x)$ and the distribution of the maxima M_n will become clear with the following theorem.

Theorem 18.2

(a) *For $0 \leq \tau \leq \infty$ and every sequence of real numbers $u_n, n \geq 1$, it holds for $n \to \infty$ that*

$$n\overline{F}(u_n) \to \tau \quad \text{if and only if} \quad \mathrm{P}(M_n \leq u_n) \to e^{-\tau}.$$

(b) *F belongs to the MDA of the GEV distribution G with the standardized sequences c_n, d_n exactly when $n \to \infty$*

$$n\overline{F}(c_n x + d_n) \to -\log G(x) \quad \text{for all } x \in \mathbb{R}.$$

The exceedance probability of the Fréchet distribution $G_{1,\alpha}$ behaves like $1/x^{\alpha}$ for $x \to \infty$, because the exponential function around 0 is approximately linear, i.e.

$$\overline{G}_{1,\alpha}(x) = \frac{1}{x^{\alpha}}\{1 + \mathcal{O}(1)\} \qquad \text{for} \quad x \to \infty.$$

Essentially all of the distributions that belong to the MDA of this Fréchet distribution show the same behaviour; $x^\alpha \overline{F}(x)$ is almost constant for $x \to \infty$, or more specifically, it is a slowly varying function.

Definition 18.6 (Slowly Varying Functions)   A positive measurable function L in $(0, \infty)$ is called *slowly varying*, if for all $t > 0$

$$\frac{L(tx)}{L(x)} \to 1 \qquad \text{for} \quad x \to \infty.$$

Typical slowly varying functions are, in addition to constants, logarithmic growth rates, for example $L(x) = \log(1 + x), x > 0$.

Theorem 18.3 (MDA of the Frechét Distribution)   *F belongs to the MDA of the Fréchet distribution $G_{1,\alpha}$ for some $\alpha > 0$, if and only if $x^\alpha \overline{F}(x) = L(x)$ is a slowly varying function. The random variables X_t with the distribution function F are unbounded, i.e. $F(x) < 1$ for all $x < \infty$, and it holds that*

$$\frac{M_n}{c_n} \xrightarrow{\mathcal{L}} G_{1,\alpha}$$

with $c_n = F^{-1}(1 - \frac{1}{n})$.

For the description of the standardized sequence c_n, the following notation was used: c_n is an extreme quantile of the distribution F, and it holds that $\overline{F}(c_n) = P(X_t > c_n) = 1/n$.

There is a corresponding criterion for the MDA of the Weibull distribution that can be shown using the relationship $G_{2,\alpha}(-x^{-1}) = G_{1,\alpha}(x)$, $x > 0$. Random variables, whose maxima are asymptotically Weibull distributed, are by all means bounded, i.e. a constant $c < \infty$ exists, such that $X_t \leq c$ holds with probability 1 ($P(X_t \leq c) = 1$). Therefore, in financial applications they are only interesting in special situations where using a type of hedging strategy, the loss, which can result from an investment, is limited. Below mainly cases where the losses are unlimited will be discussed; cases in which losses are limited can be treated in a similar fashion.

Fréchet distributions appear as asymptotic distributions of the maxima of those random variables whose probability of values beyond x only slowly decreases with x, whereas only bounded random variables belong to the MDA of Weibull distributions. Many known distributions such as the exponential or the normal distribution do not belong to either one of the groups. It is likely that in such cases the distribution of the appropriate standardized maxima converges to a Gumbel distribution.

Theorem 18.4 (MDA of the Gumbel Distribution)   *The distribution function F of the unbounded random variables X_t belongs to the MDA of the Gumbel distribution if measurable scaling functions $c(x), g(x) > 0$ as well as an absolute*

continuous function $e(x) > 0$ exist with $c(x) \to c > 0$, $g(x) \to 1$, $e'(x) \to 0$ for $x \to \infty$ so that for $z < \infty$

$$\overline{F}(x) = c(x) \exp\left(-\int_z^x \frac{g(y)}{e(y)} dy\right), \quad z < x < \infty.$$

In this case it holds that

$$\frac{M_n - d_n}{c_n} \overset{\mathcal{L}}{\longrightarrow} G_0$$

with $d_n = F^{-1}\left(1 - \frac{1}{n}\right)$ and $c_n = e(d_n)$.

As a function $e(x)$, the *average excess function* can be used:

$$e(x) = \frac{1}{\overline{F}(x)} \int_x^\infty \overline{F}(y)\, dy, \quad x < \infty,$$

which will be considered in more detail in the following.

The exponential distribution with parameter λ has the distribution function $F(x) = 1 - e^{-\lambda x}$, $x \geq 0$, so that $\overline{F}(x) = e^{-\lambda x}$ fulfils the conditions stipulated in the theorem with $c(x) = 1$, $g(x) = 1$, $z = 0$ and $e(x) = 1/\lambda$. The maximum M_n of n independent exponentially distributed random variables with parameter λ thus converges in distribution to the Gumbel distribution:

$$\lambda\left(M_n - \frac{1}{\lambda} \log n\right) \overset{\mathcal{L}}{\longrightarrow} G_0 \quad \text{for} \quad n \to \infty.$$

It can be shown, for example, that the normal distribution also belongs to the MDA of the Gumbel distribution. If, for example, M_n is the maximum of n independent standard normally distributed random variables, then it holds that

$$\sqrt{2 \log n}(M_n - d_n) \overset{\mathcal{L}}{\longrightarrow} G_0 \quad \text{for} \quad n \to \infty$$

$$\text{with} \quad d_n = \sqrt{2 \log n} - \frac{\log \log n + \log(4\pi)}{2\sqrt{2 \log n}}.$$

Figure 18.10 represents the PP plot of 130 tail values of negative log returns of the portfolio of Bayer, BMW, Siemens and Volkswagen against the Generalized Extreme Value distribution. The risk measures are expressed as positive numbers, even though they depict losses. The goal is to find the maxima of the distribution of negative log returns. The block maxima method is used, with a block size of $n = 20$. The data is fitted to the GEV distribution, and a parameter $\gamma = 0.4$ is estimated. Figure 18.11 represents the QQ plot of the data against the GEV distribution and the same method is used. As it results from the two plots, the empirical data fits quite well to the GEV distribution with parameter $\gamma = 0.4$.

PP plot, Generalized Extreme Value Distribution
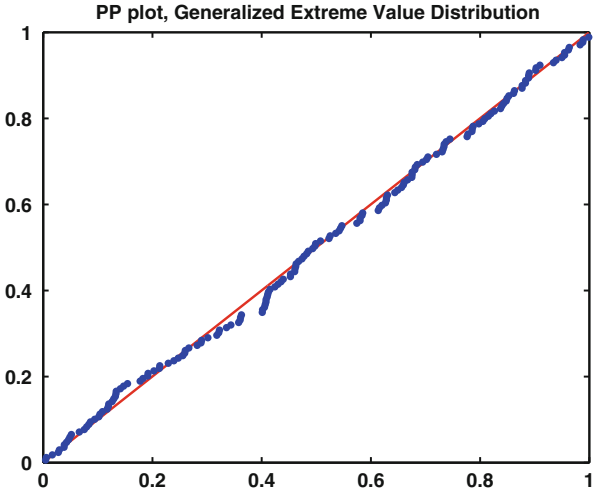
Fig. 18.10 PP plot of 130 tail values of negative log returns of the portfolio of Bayer, BMW, Siemens and Volkswagen against the generalized extreme value distribution with parameter $\gamma = 0.4$ estimated with block maxima method. Time period: 1 January 2002–31 December 2012

SFEtailGEV_pp

QQ plot, Generalized Extreme Value Distribution
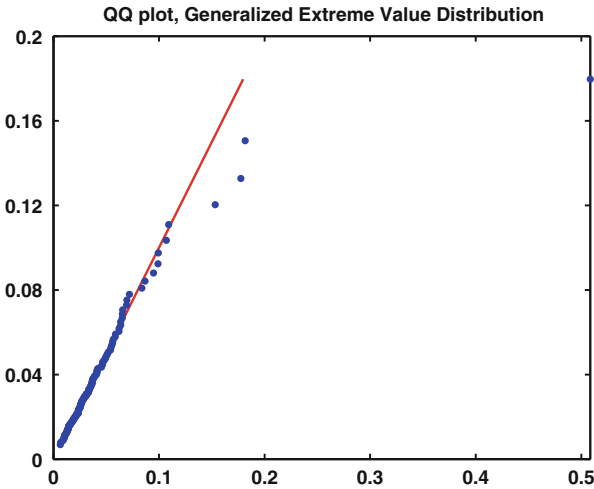
Fig. 18.11 QQ plot of 130 tail values of negative log returns of the portfolio of Bayer, BMW, Siemens and Volkswagen against the generalized extreme value distribution with parameter $\gamma = 0.4$ estimated with block maxima method. Time period: 1 January 2002–31 December 2012

SFEtailGEV_qq
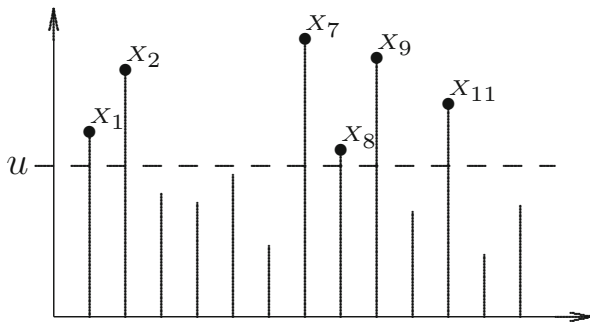
Fig. 18.12 Excesses over a threshold u

18.3.2 The Peaks-over-Threshold (POT) Method

There is a tight relationship between the asymptotic behaviour of the maxima of random variables and the distribution of the corresponding excesses which builds the foundation for an important estimation method in extreme value statistics. In general it deals with observations crossing a specified threshold u, situation which is depicted in Fig. 18.12.

The distribution F_u of the observations crossing the threshold is defined as follows:

Definition 18.7 (Excess Distribution) Let u be an arbitrary threshold and F a distribution function of an unbounded random variable X.

(a) $F_u(x) = P\{X - u \le x \mid X > u\} = \{F(u + x) - F(u)\}/\overline{F}(u)$, $0 \le x < \infty$ is called the *excess distribution* beyond the threshold u.

(b) $e(u) = E\{X - u \mid X > u\}$, $0 < u < \infty$, is the *average excess function*.

With partial integration it follows that this definition of the average excess function together with the following Theorem 18.5 agrees with:

$$e(u) = \int_u^\infty \frac{\overline{F}(y)}{\overline{F}(u)} dy.$$

If Δ_u is a random variable with the distribution function F_u, then its expectation is $E\Delta_u = e(u)$.

Theorem 18.5 (Mean Excess Function) *X is a positive, unbounded random variable with an absolutely continuous distribution function F.*

(a) The average excess function $e(u)$ identifies F exactly:

$$\overline{F}(x) = \frac{e(0)}{e(x)} \exp\left(-\int_0^x \frac{1}{e(u)} du\right), \quad x > 0.$$

(b) If F is contained in the MDA of the Fréchet distribution $G_{1,\alpha}$, then $e(u)$ is approximately linear for $u \to \infty$: $e(u) = \frac{1}{\alpha-1} u\{1 + \mathcal{O}(1)\}$.

Throughout the rest of the chapter $X, X_1, \ldots, X_n$ are unbounded, i.i.d. random variables with distribution function F.

Notation $X_{(1)} \le \ldots \le X_{(n)}$ and $X^{(1)} \ge \ldots \ge X^{(n)}$ represent the *order statistics*, that is, the data is sorted according to increasing or decreasing size. Obviously then $X_{(1)} = X^{(n)}$, $X_{(n)} = X^{(1)}$ etc.

The Generalized Pareto Approach

McNeil (1999) underlines the advantages of this style of analysis. It is important that simple parametric formulae are obtained for the measures of risk, and it is relatively easy to estimate the statistical error by means of the maximum likelihood techniques. The GPD is another probability distribution, but, in the risk management context, it is equally or more important than the Normal distribution. The Normal distribution is not suitable for addressing extreme loss, because its tails are too thin. Another member of the distributions in the MDA of the Fréchet distribution $G_{1,\alpha}$ is the *Pareto distribution* with the distribution function

$$W_{1,\alpha}(x) = 1 - \frac{1}{x^\alpha}, x \ge 1, \alpha > 0,$$

as well as all other distributions with *Pareto tails*, i.e. with

$$\overline{F}(x) = \frac{\kappa}{x^\alpha}\{1 + \mathcal{O}(1)\} \quad \text{for} \quad x \to \infty.$$

Since $\overline{F}^{-1}(\gamma)$ for $\gamma \approx 1$ behaves here like $(\kappa/\gamma)^{1/\alpha}$, c_n for $n \to \infty$ is identical to $(\kappa n)^{1/\alpha}$, and

$$\frac{M_n}{(\kappa n)^{1/\alpha}} \xrightarrow{\mathcal{L}} G_{1,\alpha} \quad \text{for } n \to \infty.$$

Definition 18.8 (Generalized Pareto Distribution)   The GPD (GP $=$ *generalized Pareto*) with parameters $\beta > 0$, γ has the distribution function

$$W_{\gamma,\beta}(x) = 1 - \left(1 + \frac{\gamma x}{\beta}\right)^{-\frac{1}{\gamma}} \quad \text{for} \quad \begin{cases} x \ge 0 & \text{if } \gamma > 0 \\ 0 \le x \le \frac{-\beta}{\gamma} & \text{if } \gamma < 0, \end{cases}$$

and

$$W_{0,\beta}(x) = 1 - e^{-\frac{x}{\beta}}, \ x \ge 0.$$

$W_\gamma(x) = W_{\gamma,1}(x)$ is called the *generalized standard Pareto distribution* or *standardized* GP distribution.
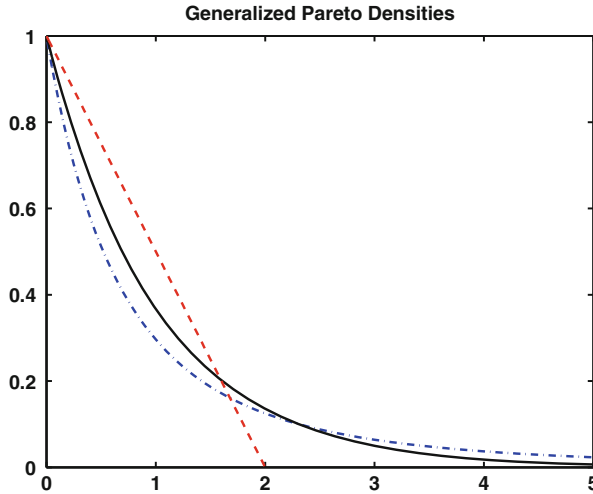
Fig. 18.13 Standard Pareto distribution ($\beta = 1$) with parameter $\gamma = 0.5$ (*dash-dot line*), 0 (*solid line*) and -0.5 (*broken line*)   SFEgpdist

Figure 18.13 shows the generalized standard Pareto distribution with parameters $\gamma = 0.5, 0$ and -0.5, respectively.

For $\gamma = 0$ the standardized GP distribution is an exponential distribution with parameter 1. For $\gamma > 0$ it is a Pareto distribution $W_{1,\alpha}$ with the parameter $\alpha = 1/\gamma$. For $\gamma < 0$ the GP distribution is also referred to as a *Beta distribution* and has the distribution function $W_{2,\alpha} = 1 - (-x)^{-\alpha}, -1 \leq x \leq 0, \alpha < 0$.

Theorem 18.6 (MDA of GEV Distribution)   *The distribution F is contained in the MDA of the GEV distribution G_γ with the form parameter $\gamma \geq 0$, exactly when for a measurable function $\beta(u) > 0$ and the GP distribution $W_{\gamma,\beta}$ it holds that:*

$$\sup_{x \geq 0} |F_u(x) - W_{\gamma,\beta(u)}(x)| \to 0 \, for \, u \to \infty.$$

A corresponding result also holds for the case when $\gamma < 0$, in which case the supremum of x must be taken for those $0 < W_{\gamma,\beta(u)}(x) < 1$.

For the GPD $F = W_{\gamma,\beta}$ it holds for every finite threshold $u > 0$

$$F_u(x) = W_{\gamma,\beta+\gamma u}(x) \quad \text{for} \quad \begin{cases} x \geq 0 & \text{if } \gamma \geq 0 \\ 0 \leq x < -\frac{\beta}{\gamma} - u & \text{if } \gamma < 0, \end{cases}$$

In this case $\beta(u) = \beta + \gamma u$.

Definition 18.9 (Empirical Average Excess Function)   Let $K_n(u) = \{j \leq n; X_j > u\}$ be the index of the observations outside of the threshold u, and let $N(u) = \#K_n(u)$ be their total number and

$$\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^n \mathbf{1}(X_j \leq x)$$

the *empirical distribution function*, $\overline{\hat{F}_n} = 1 - \hat{F}_n$.

$$e_n(u) = \int_u^\infty \overline{\hat{F}_n}(y)dy / \overline{\hat{F}_n}(u) = \frac{1}{N(u)} \sum_{j \in K_n(u)} (X_j - u)$$

$$= \frac{1}{N(u)} \sum_{j=1}^n \max\{(X_j - u), 0\}$$

is called the *empirical average excess function*.

$e_n(u)$ estimates the average excess function $e(u)$ from Sect. 18.3.1.

In the rest of this section and in the following one we will discuss estimators for extreme value characteristics, such as the exceedance probabilities $\overline{F}(x) = 1 - F(x)$ for values x or the extreme quantile $F^{-1}(q)$ for $q \approx 1$.

First, only distributions F are considered that are contained in the MDA of a GEV distribution G_γ, $\gamma \geq 0$. The corresponding random variables are thus unbounded.

Definition 18.10 (Excess over Threshold)   Let $K_n(u)$ and $N(u)$ be, as before, the index and total number of observations beyond the threshold u, respectively. The *excess* over the threshold u is defined as the random variables $Y_l, l = 1, \ldots, N(u)$, with

$$\{Y_1, \ldots, Y_{N(u)}\} = \{X_j - u; \ j \in K_n(u)\} = \{X^{(1)} - u, \ldots, X^{(N(u))} - u\}.$$

The excesses Y_l, $l \leq N(u)$ describe by how much the observations, which are larger than u, go beyond the threshold u. The POT method assumes that these excesses are the basic information source for the initial data. From the definition it immediately follows that $Y_1, \ldots, Y_{N(u)}$ are i.i.d. random variables with distribution F_u given their random total number $N(u)$, i.e. the excess distribution from Definition 18.7 is the actual distribution of the excesses. Due to Theorem 18.6 it also holds that $F_u(y) \approx W_{\gamma, \beta(u)}(y)$ for a GP distribution $W_{\gamma, \beta(u)}$ and all sufficiently large u.

The problem of estimating the exceedance probability $\overline{F}(x)$ for large x is first considered. A natural estimator is $\overline{\hat{F}}_n(x)$, and the cdf at x is replaced with the empirical distribution function. For large x, however, the empirical distribution function varies a lot because it is determined by the few extreme observations which are located around x. The effective size of the sub-sample of extreme, large observations is too small to use a pure non-parametric estimator such as the empirical distribution function. We therefore use the following relationship between the extreme exceedance probability $\overline{F}(x)$, the exceedance probability $\overline{F}(u)$ for a large, but not extremely large threshold and the excess distribution. Due to

Definition 18.7 the excess distribution is

$$\overline{F}_u(y) = P(X - u > y \mid X > u) = \overline{F}(y + u)/\overline{F}(u), \qquad \text{i.e.}$$
$$\overline{F}(x) = \overline{F}(u) \cdot \overline{F}_u(x - u), \ \ u < x < \infty. \tag{18.4}$$

For large u and using Theorem 18.6 we can approximate F_u with $W_{\gamma,\beta}$ for appropriately chosen γ, β. $F(u)$ is replaced with the empirical distribution function $\hat{F}_n(u)$ at the threshold u, for which due to the definition of $N(u)$ it holds that

$$\hat{F}_n(u) = \frac{n - N(u)}{n} = 1 - \frac{N(u)}{n}.$$

For u itself this is a useful approximation, but not for the values x, which are clearly larger than the average sized threshold u. The estimator $1 - \hat{F}_n(x)$ of $\overline{F}(x)$ for extreme x only depends on a few observations and is therefore too unreliable. For this reason the POT method uses the identity (18.4) for $\overline{F}(x)$ and replaces both factors on the right-hand side with their corresponding approximations, whereby the unknown parameter of the GPD is replaced with a suitable estimator.

Definition 18.11 (POT Estimator)   The *POT estimator* $\overline{F}^{\wedge}(x)$ for the exceedance probability $\overline{F}(x)$, for large x, is

$$\overline{F}^{\wedge}(x) = \frac{N(u)}{n} \overline{W}_{\hat{\gamma},\hat{\beta}}(x - u) = \frac{N(u)}{n} \left\{ 1 + \frac{\hat{\gamma}(x - u)}{\hat{\beta}} \right\}^{-1/\hat{\gamma}}, \ u < x < \infty,$$

whereby $\hat{\gamma}, \hat{\beta}$ are suitable estimators for γ and β, respectively.

Maximum Likelihood Estimation of $\hat{\gamma}, \hat{\beta}$

$\hat{\gamma}, \hat{\beta}$ can be, for example, calculated as maximum likelihood estimators from the excesses $Y_1, \ldots, Y_{N(u)}$. First, we consider the case where $N(u) = m$ is a constant and where $Y_1, \ldots, Y_m$ is a sample of i.i.d. random variables with the distribution $W_{\gamma,\beta}, \gamma > 0$. Thus $W_{\gamma,\beta}$ is literally a Pareto distribution and has the probability density

$$p(y) = \frac{1}{\beta} \left(1 + \frac{\gamma y}{\beta} \right)^{-1/\gamma - 1}, \ x \geq 0.$$

Therefore, the log likelihood function is

$$\ell(\gamma, \beta \mid Y_1, \ldots, Y_m) = -m \log \beta - \left(\frac{1}{\gamma} + 1 \right) \sum_{j=1}^{m} \log \left(1 + \frac{\gamma}{\beta} Y_j \right).$$

By maximizing this function with respect to γ, β we obtain the maximum likelihood (ML) estimator $\hat{\gamma}, \hat{\beta}$. Analogously we could also define the ML estimator for the parameter of the GPD using $\gamma \leq 0$.

Theorem 18.7 *For all $\gamma > -\frac{1}{2}$ it holds for $m \to \infty$*

$$\sqrt{m}\left(\hat{\gamma} - \gamma, \frac{\hat{\beta}}{\beta} - 1\right) \xrightarrow{\mathcal{L}} N_2(0, D^{-1}),$$

with $D = (1+\gamma)\begin{pmatrix} 1+\gamma & -1 \\ -1 & 2 \end{pmatrix}$, i.e. $(\hat{\gamma}, \hat{\beta})$ are asymptotically normally distributed. In addition they are asymptotically efficient estimators.

In our initial problem $m = N(u)$ was random. Here the estimators we have just defined, $\hat{\gamma}$ and $\hat{\beta}$, are the conditional ML estimators given $N(u)$. The asymptotic distribution theory is also known in this case; in order to avoid an asymptotic bias, $\overline{F}$ must fulfil an additional regularity condition. After we find an estimator for the exceedance probability and thus a cdf for large x, we immediately obtain an estimator for the extreme quantile.

Estimating the Value-at-Risk

Definition 18.12 (POT Quantile Estimator)   The *POT Quantile estimator* $\hat{x}_q$ for the q-quantile $x_q = F^{-1}(q)$ is the solution to $\overline{F}^{\wedge}(\hat{x}_q) = 1 - q$, i.e.

$$\hat{x}_q = u + \frac{\hat{\beta}}{\hat{\gamma}}\left[\left\{\frac{n}{N(u)}(1-q)\right\}^{-\hat{\gamma}} - 1\right].$$

This quantile is actually the Value-at-Risk, so

$$\widehat{VaR}_q = u + \frac{\hat{\beta}}{\hat{\gamma}}\left[\left\{\frac{n}{N(u)}(1-q)\right\}^{-\hat{\gamma}} - 1\right].$$

Comparison to the Empirical Quantile

These estimators can be compared with the usual sample quantiles. To do this a threshold value u is selected so that exactly k excesses lie beyond u, that is $N(u) = k > n(1-q)$ and thus $u = X^{(k+1)}$. The POT quantile estimator that is dependent on the choice of u respectively k is

$$\hat{x}_{q,k} = X^{(k+1)} + \frac{\hat{\beta}_k}{\hat{\gamma}_k}\left[\left\{\frac{n}{k}(1-q)\right\}^{-\hat{\gamma}_k} - 1\right],$$

where $\hat{\gamma}_k, \hat{\beta}_k$ is the ML estimator, dependent on the choice of k, for γ and β. The corresponding sample quantile is

$$\hat{x}_q^s = X^{([n(1-q)]+1)}.$$

This is in approximate agreement with $\hat{x}_{q,k}$ when the minimal value $k = [n(1 - q)] + 1$ is chosen for k. Simulation studies show that the value k_0 of k, which minimizes the mean squared error $\mathrm{MSE}(\hat{x}_{q,k}) = \mathsf{E}(\hat{x}_{q,k} - x_q)^2$, is much larger than $[n(1 - q)] + 1$, i.e. the POT estimator for x_q differs distinctly from the sample quantile $\hat{x}_q^s$ and is superior to it with respect to the mean squared error when the thresholds u or k are respectively chosen.

We are interested in the threshold u, for which the mean squared error of $\hat{x}_q$ is as small as possible. The error can be split into the variance and the squared bias of $\hat{x}_q$:

$$\mathrm{MSE}(\hat{x}_q) = \mathsf{E}(\hat{x}_q - x_q)^2 = \mathsf{Var}(\hat{x}_q) + \{\mathsf{E}(\hat{x}_q) - x_q\}^2.$$

The Mean Square Error Dilemma

Unfortunately the two components of the mean squared error move in opposite directions when the threshold u used in calculating the POT quantile estimators is varied. The following bias variance dilemma needs to be considered:

- when u is too large, there are a few excesses Y_l, $l \leq \mathrm{N}(u)$, and the estimator's variance is too large,
- when u is too small, the approximation of the excess distribution using a GPD is not good enough, and the bias $\mathsf{E}(\hat{x}_q) - x_q$ is no longer reliable.

An essential aid in selecting an appropriate threshold u is the average excess plot, which is approximately linear beyond the appropriate threshold. This has already been discussed in Theorem 18.5, when one considers the relationship between the Fréchet distribution as the asymptotic distribution of the maxima and the Pareto distribution as the asymptotic distribution of the excesses. It is supported by the following result for the Pareto and exponential distributions $W_{\gamma,\beta}, \gamma \geq 0$.

Theorem 18.8   *Let Z be a $W_{\gamma,\beta}$ distributed random variable with $0 \leq \gamma < 1$. The average excess function is linear:*

$$e(u) = \mathsf{E}\{Z - u | Z > u\} = \frac{\beta + \gamma u}{1 - \gamma}, \ \ u \geq 0, \ for \ 0 \leq \gamma < 1.$$

With the usual parametrization of the Pareto distribution $\gamma = \frac{1}{\alpha}$, i.e. the condition $\gamma < 1$ means that $\alpha > 1$ and thus $\mathsf{E}|Z| < \infty$.

This result motivates the following application in choosing the threshold: select the threshold u of the POT estimator so that the empirical average excess function

$e_n(v)$ for values $v \geq u$ is approximately linear. An appropriate u is chosen by considering the average excess plots, where it is recommended that the largest points $(X_{(k)}, e_n(X_{(k)}))$, $k \approx n$, along the right-hand edge of the plot be excluded, since their large variability for the most part distorts the optical impression.

Estimating the Expected Shortfall

The expected shortfall is closely related to the average excess function when $u = x_q$, as immediately can be seen from the definition:

$$S_q = e(x_q) + x_q.$$

Only the POT estimator for the expected shortfall S_q is considered. Since $F_u(x) \approx W_{\gamma,\beta}(x)$ for a sufficiently large threshold u, it holds from Theorem 18.5, b) with $\alpha = 1/\gamma$

$$e(v) \approx \frac{\beta + (v - u)\gamma}{1 - \gamma} \quad \text{for} \;\; v > u.$$

Therefore, for $x_q > u$ we have

$$\frac{S_q}{x_q} = 1 + \frac{e(x_q)}{x_q} \approx \frac{1}{1 - \gamma} + \frac{\beta - \gamma u}{x_q(1 - \gamma)}.$$

The *POT estimator for the expected shortfall S_q* is thus

$$\hat{S}_{q,u} = \frac{\hat{x}_q}{1 - \hat{\gamma}} + \frac{\hat{\beta} - \hat{\gamma}u}{1 - \hat{\gamma}},$$

where $\hat{x}_q$ is the POT quantile estimator.

Given that $\hat{x}_q$ is actually the Value-at-Risk, one can rewrite:

$$\hat{S}_{q,u} = \frac{\widehat{VaR}_q}{1 - \hat{\gamma}} + \frac{\hat{\beta} - \hat{\gamma}u}{1 - \hat{\gamma}}.$$

Figure 18.14 represents the PP plot of 130 tail values of negative log returns of the portfolio of Bayer, BMW, Siemens and Volkswagen against the GPD. Negative log returns are used, following the same logic from the previous method. The peaks-over-threshold method is used, with the Generalized Pareto approach. The data is fitted to the GPD, and a parameter $\gamma = 0.34$ is estimated. Figure 18.15 represents the QQ plot of the negative log returns of the portfolio against the GPD and the method used is the same. Also in the case of the POT method, the empirical data fits quite well to the GPD with parameter $\gamma = 0.34$.
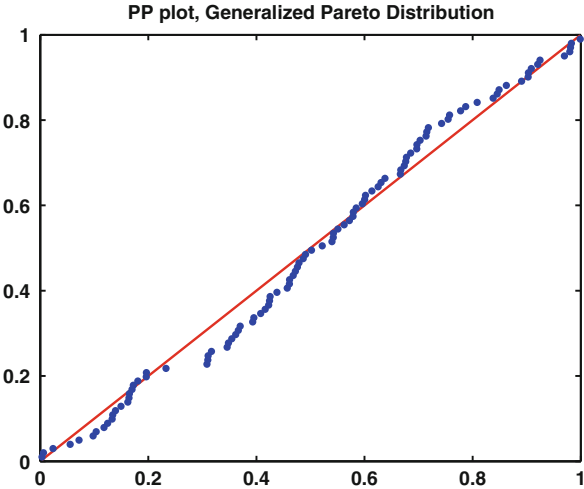
Fig. 18.14 PP plot of 130 tail values of negative log returns of the portfolio of Bayer, BMW, Siemens and Volkswagen against the generalized Pareto distribution with parameter $\gamma = 0.34$ estimated with the POT method, generalized Pareto approach. Time period: 1 January 2002–31 December 2012 $\boxed{Q}$ SFEtailGPareto_pp
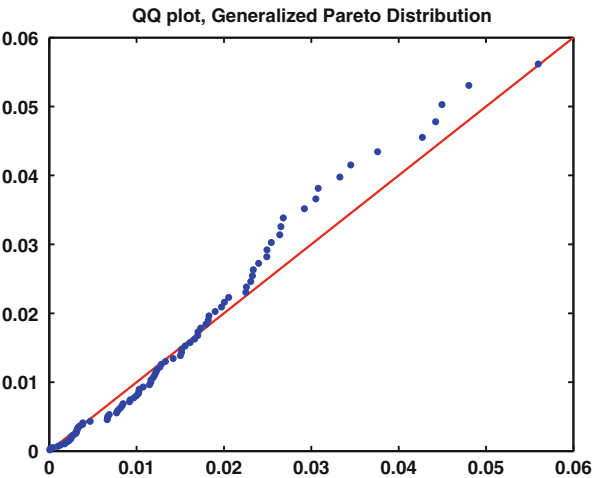
Fig. 18.15 QQ plot of 130 tail values of negative log returns of the portfolio of Bayer, BMW, Siemens and Volkswagen against the generalized Pareto distribution with parameter $\gamma = 0.34$ estimated with the POT method, generalized Pareto approach. Time period: 1 January 2002–31 December 2012 $\boxed{Q}$ SFEtailGPareto_qq

The Hill Estimator Approach

The POT method for estimating the exceedance probability and the extreme quantiles can be used on data with cdf that is in the MDA of a Gumbel or a Fréchet

distribution, as long as the expected value is finite. Even for extreme financial data, this estimator seems reasonable based on empirical evidence. A classic alternative to the generalised Pareto approach POT estimator is the Hill estimator. It is only useful for distributions with slowly decaying tails, such as those in the MDA of the Fréchet distribution, and in simulations it often performs worse in comparison to the POT estimator.

In this section it is assumed that the data $X_1, \ldots, X_n$ are i.i.d. with a distribution function F in the MDA of $G_{1,\alpha}$ for some $\alpha > 0$. Due to Theorem 18.3 this is the case when $\overline{F}(x) = x^{-\alpha} L(x)$ with a slowly varying function L. The tapering behaviour of $\overline{F}(x) = P(X_t > x)$ for increasing x is mainly determined by the so-called *tail exponents* α. The starting point of the Hill method is the following estimator for α.

Definition 18.13 (Hill Estimator) $X^{(1)} \geq X^{(2)} \geq \ldots \geq X^{(n)}$ are the order statistics in decreasing order. The *Hill estimator* $\hat{\alpha}_H$ of the tail exponents α for a suitable $k = k(n)$ is

$$\hat{\alpha}_H = \left(\frac{1}{k} \sum_{j=1}^{k} \log\, X^{(j)} - \log\, X^{(k)} \right)^{-1}.$$

The form of the estimator can be seen from the following simple special case. In general it holds that $\overline{F}(x) = L(x)/(x^\alpha)$, but here it is assumed that with a fixed $c > 0$ $L(x) = c^\alpha$ is constant. Set $V_j = \log(X_j/c)$, it holds that

$$P(V_j > v) = P(X_j > ce^v) = \overline{F}(ce^v) = \frac{c^\alpha}{(ce^v)^\alpha} = e^{-\alpha v},\ y \geq 0,$$

$V_1, \ldots, V_n$ are therefore independent exponentially distributed random variables with parameter α. As is well known it holds that $\alpha = (\mathsf{E}V_j)^{-1}$, and the ML estimator $\hat{\alpha}$ for α is $1/\overline{V}_n$, where $\overline{V}_n$ stands for the sample average of $V_1, \ldots, V_n$, thus,

$$\hat{\alpha} = \frac{1}{\overline{V}_n} = \left(\frac{1}{n} \sum_{j=1}^{n} \log(X_j/c) \right)^{-1} = \left(\frac{1}{n} \sum_{j=1}^{n} \log\, X^{(j)} - \log\, c \right)^{-1},$$

where for the last equation only the order of addition was changed. $\hat{\alpha}$ is already similar to the Hill estimator. In general it of course only holds that $\overline{F}(x) \approx \frac{c^\alpha}{x^\alpha}$ for sufficiently large x. The argument for the special case is similar for the largest observations $X^{(1)} \geq X^{(2)} \geq \ldots \geq X^{(k)} \geq u$ beyond the threshold u, so that only the k largest order statistics enter the definition of the Hill estimator.

The Hill estimator is consistent, that is it converges in probability to α when $n, k \to \infty$ such that $k/n \to 0$. Under an additional condition it can also be shown that $\sqrt{k}(\hat{\alpha}_H - \alpha) \xrightarrow{\mathcal{L}} N(0, \alpha^2)$, i.e. $\hat{\alpha}_H$ is asymptotically normally distributed.

Similar to the POT estimator when considering the Hill estimator the question regarding the choice of the threshold $u = X^{(k)}$ comes into play, since the

observations located beyond it enter the estimation. Once again we have a bias variance dilemma:

- when k is too small, only a few observations influence $\hat{\alpha}_H$, and the variance of the estimator, which is α^2/k asymptotically, becomes too large,
- when k is too large, the assumption underlying the derivation of the estimator, i.e. that $L(x)$ is approximately constant for all $x \geq X^{(k)}$ is in general not well met and the bias $\mathsf{E}\hat{\alpha}_H - \alpha$ becomes too large.

Based on the fundamentals of the Hill estimator for the tail exponents α we obtain direct estimators for the exceedance probability $\overline{F}(x)$ and for the quantiles of F. Since $\overline{F}(x) = x^{-\alpha}L(x)$ with a slowly varying function L, it holds for large $x \geq X^{(k)}$ that:

$$\frac{\overline{F}(x)}{\overline{F}(X^{(k)})} = \frac{L(x)}{L(X^{(k)})} \left(\frac{X^{(k)}}{x}\right)^{\alpha} \approx \left(\frac{X^{(k)}}{x}\right)^{\alpha}, \tag{18.5}$$

Because exactly one portion k/n of the data is larger or equal to the order statistic $X^{(k)}$, this is the $(1 - k/n)$ sample quantile. Therefore, the empirical distribution function takes on the value $1 - k/n$ at $X^{(k)}$, since it uniformly converges to the distribution function F, for sufficiently large n, a k that is not too large in comparison to n yields: $F(X^{(k)}) \approx 1 - k/n$, i.e., $\overline{F}(X^{(k)}) \approx k/n$. Substituting this into (18.5), a *Hill estimator is obtained for the exceedance probability* $\overline{F}(x)$:

$$\overline{F}_{H}^{\wedge}(x) = \frac{k}{n} \left(\frac{X^{(k)}}{x}\right)^{\hat{\alpha}_H}$$

By inverting this estimator, the *Hill quantile estimator* is obtained for the $q-$ quantile x_q with $q \approx 1$:

$$\hat{x}_{q,H} = X^{(k)} \left\{\frac{n}{k}(1-q)\right\}^{-1/\hat{\alpha}_H}$$

$$= X^{(k)} + X^{(k)} \left[\left\{\frac{n}{k}(1-q)\right\}^{-\hat{\gamma}_H} - 1\right]$$

with $\hat{\gamma}_H = 1/\hat{\alpha}_H$, where the second representation clearly shows the similarities and differences to the POT quantile estimator.

   In Value-at-Risk notation,

$$\widehat{VaR}_{q,H} = X^{(k)} + X^{(k)} \left[\left\{\frac{n}{k}(1-q)\right\}^{-\hat{\gamma}_H} - 1\right].$$

The Hill quantile estimator is also illustrated with the help of a Matlab program.
Q SFEhillquantile

18.4   Backtesting

A moving window procedure is used as follows: static windows of size w = 250 scrolling in time t for VaR estimation:

$$\{X_t\}_{t=s-w+1}^{s}$$

for $s = w, \ldots, T$.

The VaR estimation procedure generates a time series $\{\widehat{VaR}_{1-\alpha}^t\}_{t=w}^{T}$ and $\{\hat{\mu}\}_{t=w}^{T}$, $\{\hat{\sigma}\}_{t=w}^{T}$, $\{\hat{\gamma}\}_{t=w}^{T}$ of parameter estimates.

The estimated VaR values are compared with true realizations l_t of the Profit and Loss function. An exceedance occurs when l_t is smaller than the $\widehat{VaR}_{1-\alpha}^t$ (Fig. 18.16).

The ratio of the number of exceedances to the number of observations gives the exceedances ratio:

$$\hat{\alpha} = \frac{1}{T-h} \sum_{t=h+1}^{T} \mathbf{1}(l_t < \widehat{VaR}_{1-\alpha}^t)$$
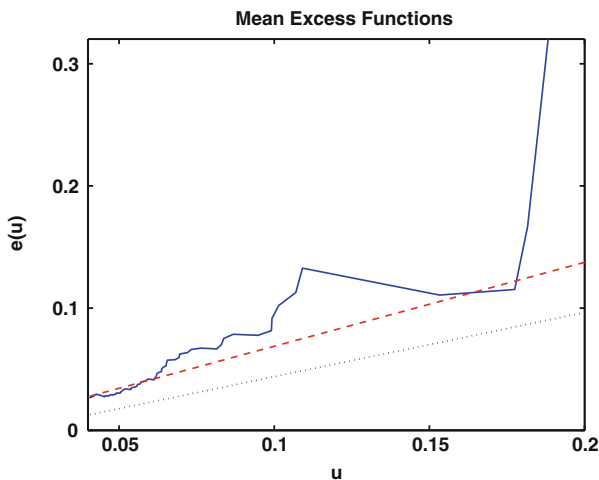
Fig. 18.16 Empirical mean excess function (*blue line*), mean excess function of a generalized Pareto distribution (*black line*) and mean excess function of a Pareto distribution with parameter estimated with Hill estimator (*red line*) for the portfolio of Bayer, BMW, Siemens and Volkswagen, negative log returns. Time period: 1 January 2002–31 December 2012   **Q** SFEMeanExcessFun

18.5 EVT for Time Series

Let Z_t, $-\infty < t < \infty$, be a *strictly stationary time series*, as defined in Definition 11.6, that is the distribution of the data and its probability structure do not change over time. Each single observation Z_t has, among other things, the same distribution function F. To compare consider the i.i.d. random variables $X_1, X_2, \ldots$ with the same distribution F. Let $M_n = \max\{Z_1, \ldots, Z_n\}$, $M_n^x = \max\{X_1, \ldots, X_n\}$ be the maxima of n values from the time series respectively from n independent observations. A simple but basic relationship for the previous sections is (18.1), i.e.

$$P(M_n^x \le y) = \{P(X_j \le y)\}^n = F^n(y),$$

where the independence of X_t is used. For dependent data this relationship does not hold and the distribution of the maximum M_n is not determined by F alone, but rather from the complete distribution of the time series. Luckily in many cases there is at least one comparable, approximate relationship (Fig. 18.17):

$$P(M_n \le y) \approx F^{n\delta}(y) \ge F^n(y) \quad \text{for large } n,$$

where $\delta \in [0, 1]$ is the so-called extremal index. In order to find an exact definition, recall Theorem 18.2 for the independent case, whereby

$$n\overline{F}(u_n) \to \tau$$

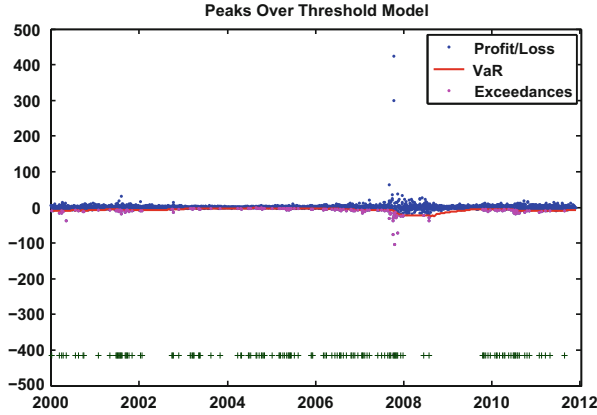if and only if $P(M_n^x \le u_n) \to e^{-\tau}$.
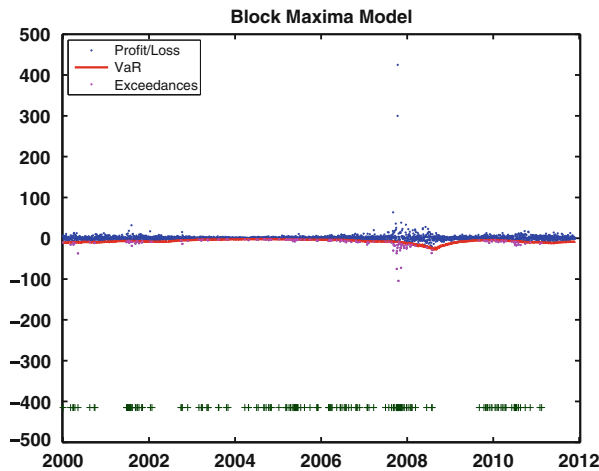
Fig. 18.17 Value-at-risk estimation at 0.05 level for the portfolio of Bayer, BMW, Siemens and Volkswagen. Time period: 1 January 2002–31 December 2012. Size of moving window 250, size of block 40. Backtesting result $\hat{\alpha} = 0.0530$ 🔍 SFEvar_block_max_backtesting
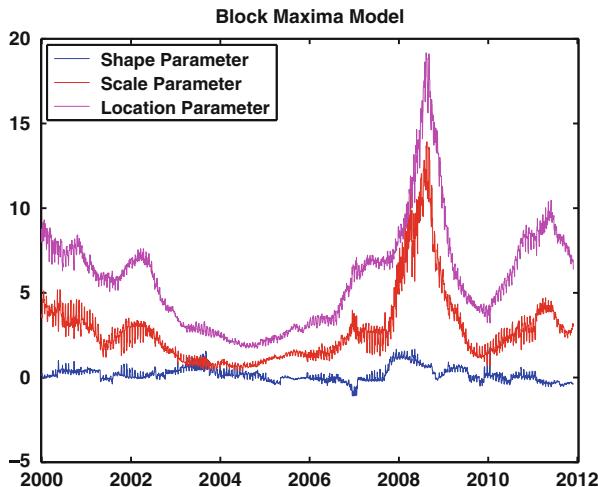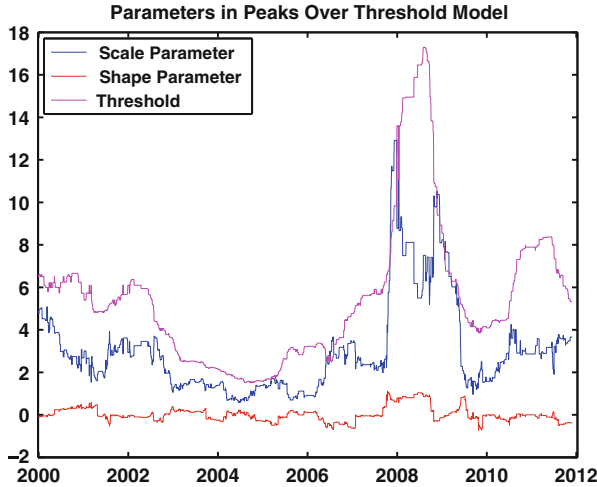
Block Maxima Model

Fig. 18.18 Parameters estimated in the block maxima model for the portfolio of Bayer, BMW, Siemens and Volkswagen. Time period: 1 January 2002–31 December 2012

🔍 SFEvar_block_max_params

Definition 18.14 (Extremal Index) $\delta \in [0, 1]$ is called the *extremal index* of the time series Z_j, $-\infty < j < \infty$, when for certain τ, u_n

$$n\overline{F}(u_n) \to \tau \quad \text{and} \quad P(M_n \leq u_n) \to e^{-\delta\tau}.$$

(If δ exists, then the value does not depend on the specific choice of τ, u_n) (Fig. 18.18).

From the definition the above claimed approximate relationship between the distribution of the maximum and the exceedance probability immediately follows:

$$P(M_n \leq u_n) \approx e^{-\delta\tau} \approx e^{-\delta n \overline{F}(u_n)} = (e^{-\overline{F}(u_n)})^{n\delta} \approx (1 - \overline{F}(u_n))^{n\delta} = F^{n\delta}(u_n),$$

when u_n is large and thus $\overline{F}(u_n) \approx 0$.

Pure white noise automatically has the extremal index $\delta = 1$, since Z_t here are independent. It is not obvious that all ARMA(p, q) processes (see Chap. 12) with normally distributed innovations also have an extremal index $\delta = 1$, its maxima thus behave like maxima from independent data. Intuitively this comes from, on the one hand, ARMA processes having an exponentially decreasing memory, i.e. the observations $Z_t, Z_{t+\tau}$ are for sufficiently large time periods τ practically independent, and, on the other hand, the probability of two extreme observations occurring within the same time interval (which is not too long) is low. These qualitative statements can be formulated as two precise criteria of time series that

Fig. 18.19 Value-at-risk estimation at 0.05 level for the portfolio of Bayer, BMW, Siemens and Volkswagen. Time period: 1 January 2002–31 December 2012. Size of moving window 250. Backtesting result $\hat{\alpha} = 0.0569$ **Q** SFEvar_pot_backtesting

have an extremal index of 1, the exact formulation of which will not be given here (Fig. 18.19).

For financial time series models the second condition is not fulfilled, because they contradict the presence of volatility clusters (see Chap. 13), i.e. the local frequency of extreme observations. The extremal index of an ARCH(1) process with parameters ω, α (see Definition 13.1) is, for example, always $\delta = \delta(\alpha) < 1$. It can be approximated for $\alpha = 0.5$, for example, $\delta \approx 0.835$.

Finally note that not every time series has an extremal index. A simple counterexample is $Z_t = A \cdot X_t$ with i.i.d. random variables X_t, which are modelled by a random factor $A > 0$ that is independent of X_t. Since the factor A is contained in all observations, even in the most distant past, this time series has no decreasing memory. If the distribution of X_t has slowly decaying tails, i.e. they belong to the MDA of a Fréchet distribution, then it can be shown that Z_t cannot have an extremal index (Fig. 18.20).

The extreme theory for time series is still developing. The Fisher–Tippett theorem, however, exists as a central result in the following modified form:

Theorem 18.9   *Let $\{Z_t\}$ be a strictly stationary time series with the distribution function F and an extremal index $\delta > 0$. Let $X_1, X_2, \ldots$ be i.i.d. with the same distribution function F. $M_n^x = \max\{X_1, \ldots, X_n\}$. Let G_γ be a general extreme value distribution. We have*

$$\mathrm{P}\left(\frac{M_n^x - d_n}{c_n} \le x\right) \to G_\gamma(x)$$

$$\text{if and only if} \quad \mathrm{P}\left(\frac{M_n - d_n}{c_n} \le x\right) \to G_\gamma^\delta(x)$$

for all x with $0 < G_\gamma(x) < 1$.

Fig. 18.20 Parameters estimated in the POT Model for the portfolio of Bayer, BMW, Siemens and Volkswagen. Time period: 1 January 2002–31 December 2012   🅠 SFE_var_pot_params

The maxima of the time series are standardized by the same series c_n, d_n and converge in distribution to the same type of asymptotic distribution as the maxima of the corresponding independent data, since G_γ^δ is itself a general extreme value distribution with the same form parameters as G_γ. For example, for $\gamma > 0$ it holds that

$$G_\gamma^\delta(x) = \exp\{-\delta(1 + \gamma x)^{-1/\gamma}\} = G_\gamma \left(\frac{x - \mu}{\sigma}\right), \ \ 1 + \gamma x > 0$$

with $\sigma = \delta^\gamma$ and $\mu = -(1 - \delta^\gamma)$, i.e. except for the location and scale parameters the distributions are identical.

Many of the techniques used in extreme value statistics, that were developed for independent data can be used on time series. To do this, however, one needs to have more data, because the effective size of the sample is only $n\delta$ instead of n. Besides that, additional problems appear: the POT method is perhaps in theory still applicable, but the excesses are no longer independent, especially when a financial time series with volatility clusters is considered. For this reason the parameters of the GPD, with which the excess distribution is approximated, cannot be estimated by simply taking the maximum of the likelihood function of independent data. One way out of this is to either use special model assumptions, with which the likelihood function of the dependent excesses can be calculated, or by using a reduction technique, with which the data is made more "independent" at the cost of the sample size. One application, for example, replaces the cluster of neighbouring excesses with a maximum value from the cluster, whereby the cluster size is so chosen that the sample size of the excesses is approximately reduced by the factor δ. Afterwards the POT estimators, which were developed for independent data, can be calculated from the reduced excesses.

Another problem is that the extremal index needs to be estimated in order to be able to use applications like the one just described. In the literature several estimation techniques are described. We will introduce only one here; one that can be described without a lot of technical preparation, the so-called *Block method*. First the time series data $Z_1, \ldots, Z_n$ is divided into b blocks, each has a length l (size $n = bl$, b, l large). Let $M_l^{(k)}$ be the maximum of the observations in the k-th block:

$$M_l^{(k)} = \max(Z_{(k-1)l+1}, \ldots, Z_{kl}), \ k = 1, \ldots, b.$$

For a large threshold value u, let $N(u) = \#\{t \leq n; \ Z_t > u\}$ be the number of observations beyond the threshold and let $B(u) = \#\{k \leq b; \ M_l^{(k)} > u\}$ be the number of blocks with at least one observation beyond the threshold u. The estimator for the extremal index is then

$$\hat{\delta} = \frac{1}{l} \frac{\log(1 - B(u)/b)}{\log(1 - N(u)/n)}.$$

Heuristically this estimator can be derived from the following three observations:

(1) From the definition of the extremal index it follows that $P(M_n \leq u) \approx F^{\delta n}(u)$, when $n, u \to \infty$, so that $n \overline{F}(u) \to \tau$. Solving for δ it follows that

$$\delta \approx \frac{\log P(M_n \leq u)}{n \log F(u)}.$$

(2) F can be estimated using the empirical distribution function $\hat{F}_n$, so that $F(u) = 1 - P(Z_t > u) \approx 1 - \frac{N(u)}{n}$.

(3) With $n = bl$ it follows that

$$P(M_n \leq u) \approx \prod_{k=1}^{b} P(M_l^{(k)} \leq u) \approx \{P(M_l^{(1)} \leq u)\}^b$$

$$\approx \left(\frac{1}{b} \sum_{k=1}^{b} \mathbf{1}(M_l^{(k)} \leq u) \right)^b = \left(1 - \frac{B(u)}{b} \right)^b.$$

By combining the three observations we have

$$\delta \approx \frac{b \ \log (1 - B(u)/b)}{n \ \log (1 - N(u)/n)} = \hat{\delta}.$$

18.6   Recommended Literature

Both of the basic theorems, Theorems 18.1 and 18.6, of this section go back to
Fisher and Tippett (1928) respectively Pickands (1975). The essential notion of
quantifying risk by coherent risk measures was introduced by Artzner et al. (1997).
Further information can be found in Artzner et al. (1999).

A comprehensive summary of the modelling and statistical analysis of extreme
results is given in the monograph from Embrechts et al. (1997). There one finds
proofs as well as detailed mathematical and practical considerations of the content
of this section and an extensive bibliography. Another actual and recommendable
book on extreme value statistics is Reiss and Thomas (1997). A more in depth
implementation of the method in the form of quantlets discussed in this last
reference, which goes beyond the selection introduced in this section, can be found
in Reiss and Thomas (2000).

A substantial problem that occurs when applying the methods of extreme value
statistics such as the POT or the Hill estimators is the choice of the threshold value u
or the corresponding number k of large order statistics. We have already mentioned
how this choice can be made with the help of graphical representations. A more
in depth discussion including the corresponding quantlets can be found in Reiss
and Thomas (2000). Polzehl and Spokoiny (2006) and Grama and Spokoiny (2003)
describe current procedures used for estimating the tail exponents, for which the
choice of u or k respectively, given the available data, can be adaptively and thus
automatically chosen.

The methods described in this chapter give estimators for the Value-at-Risk as
unconditional quantiles. Often one wishes to include financial data from the recent
past when estimating risk, for example in a GARCH(1,1) model the last observation
and the last volatility. In this case the Value-at-Risk is a conditional quantile given
the available information. One possibility of using extreme value statistics in such
cases is based on the assumptions of a specific stochastic volatility model which is
parametric as in McNeil and Frey (2000) or non-parametric as in Chap. 15.

Given the assumptions of the model a conditional volatility σ_t is estimated given
the past, which together with the data results in an estimator for the innovations
Z_t. In calculating the conditional quantile it is not assumed that the Z_t are
standard normally distributed, but instead the needed unconditional quantile of
the innovations is estimated from the estimated innovations with, for example, the
POT estimator. Alternatively one can estimate the conditional quantile also direct as
non-parametric, in which the conditional distribution function is first estimated with
a kernel estimator and then the inverse is taken. With moderately large quantiles,
for example, with a 95 % VaR, the method from Franke and Mwita (2003) gives
good results, even for innovation distributions with heavy tails and infinite variance.
For extreme quantiles such as the 99 % VaR a semi-parametric method must be
considered, as is the case with the POT method, in order to obtain useful results.
Mwita (2003) estimates first a non-parametric, medium-sized conditional quantile

and modifies this estimator through the fitting of a Pareto distribution to the extreme excesses.

Other useful references are Gilli and Këllezi (2006), Embrechts, Resnick and Samorodnitsky (1998), Longin (2003), McNeil (1999) and Tsevas and Panaretos (1998).

18.7 Exercises

Exercise 18.1 *The tail behaviour of distributions determines the size and the frequencies of extremal values. Heavy-tailed distributions, like stable (including Cauchy and Levy) tend to have more extremal values than distributions like the normal with light exponentially decreasing tails.*

Provide evidence for this statement by simulating stable and normal variates. More precisely, simulate 500 1.5-stable and normal variables and comment on the size and frequency of the outliers with same scale.

Exercise 18.2 *The maximum of n independent unbounded random variables tends in probability to infinity. The convergence to infinity may be slow or fast depending on the tail behaviour of the distributions. Consider a sequence of random variables $M_n = max(X_1, \ldots, X_n)$, $n = m, 2m, 3m, \ldots$. Plot n vs. M_n for different kinds of distributions. It is suggested that one does this exercise for standard normal and stable distribution (see Exercise 18.1).*

Exercise 18.3 *The empirical quantile is defined on the basis of order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ as $\hat{x}_q = \hat{F}_n^{-1}(q)$. Derive the asymptotic distribution of $\hat{x}_q - x_q$.*

Exercise 18.4 *The PP-Plot is a diagnostic tool for graphical inspection of the goodness of fit of hypothetical distribution F. (a) How the PP-Plot is constructed? Construct the normal PP-Plot of daily log-returns of the portfolio (Bayer, BMW, Siemens) from 2002-01-01 to 2012-12-31 to check the fit of the normal distribution. Is the normal distribution an acceptable approximation of the data?*

(b) For the given dataset of the 100 tail values of daily log-returns, estimate the parameter γ using block maxima method. Validate the fit of the GEV distribution with the estimated parameter γ using PP-plot.

(c) Repeat (b) for the Peaks over Threshold (POT) method by estimating γ for the Generalized Pareto Distribution. Use PP-Plot to check the fit of the distribution. Is the approximation better?

Exercise 18.5 *The QQ-Plot is a diagnostic tool for graphical inspection of the goodness of fit of hypothetical distribution F.*

(a) What is the advantage of QQ-Plot in comparison with PP-Plot? Construct the normal QQ-Plot of daily log-returns of the portfolio (Bayer, BMW, Siemens) from 2002-01-01 to 2012-12-31 to check the fit of the normal distribution. Is the normal distribution an acceptable approximation of the data?

(b) *For the given dataset of the 100 tail values of daily log-returns globally estimate the parameter γ using block maxima method. Validate the fit of the GEV distribution with the estimated parameter γ using QQ-plot.*

(c) *Repeat (b) for the Peaks over Threshold (POT) method by estimating a global γ for the Generalized Pareto Distribution. Use QQ-Plot to check the fit of the distribution. Is the approximation better?*

Exercise 18.6 *The mean excess function*

$$e(u) = E(X - u \mid X > u) \ 0 < u < \infty$$

determines not only the tail behaviour of the distribution but also uniquely determines F. Prove the formula

$$\overline{F}(x) = \frac{e(0)}{e(x)} \exp\left\{ -\int_0^x \frac{1}{e(u)} du \right\}, \quad x > 0$$

Simulate from Fréchet distribution with $\alpha = 2$, do the PP-plot and calculate the mean excess function. Estimate α from the mean excess function and plot the empirical mean excess function.

Exercise 18.7 *Suppose the Pareto distribution $\bar{F}(x) = P(X > x) \sim kx^{-\alpha}$ where $\alpha > 0$. It is well known that an approximation of the parameter α can be obtained s $\log \bar{F}(x) \approx \log k - \alpha \log x$. Estimate this logarithm approximation for the empirical distribution of the portfolio (Bayer, BMW, Siemens) from 2002-01-01 to 2012-12-31.*

Exercise 18.8 *Estimate the γ parameter locally with Block Maxima, POT, Hill and Regression Model of the portfolio (Bayer, BMW, Siemens) from 2002-01-01 to 2012-12-31. Assuming different types of distributions, plot the mean excess function $e(u)$ for this portfolio. What do you observe?*

Exercise 18.9 *Estimate the Value-at-Risk with Block Maxima Model and with the POT Model of the portfolio (Bayer, BMW, Siemens) from 2002-01-01 to 2012-12-31. Plot the shape and scale parameters estimates over time.*

Exercise 18.10 *Let Y_1 be a "short position" in a stock with log-normal distribution*

$$Y_1 = \pi - S$$

with $S = \exp(Z)$ where Z is normally distributed with $N(m, \sigma^2)$.

(a) *Calculate $VaR_\alpha(Y_1)$ for $\alpha \in (0, 1)$.*

(b) *Let $Y_1, Y_2, \ldots$ independent and identically distributed. Show for $\alpha \in (0, 1)$:*

$$VaR_\alpha \left(n^{-1} \sum_{i=1}^n Y_i \right) \to - E[Y_1]$$

(c) *Which parameter values α violate the convexity property given a large n?*

Exercise 18.11 *Let ρ be a normed risk measure. Show that*

(a) $\rho(x) \geq -\rho(-x)$ for all x.
(b) $\rho(\lambda x) \geq \lambda \rho(x)$ for all $\lambda \in (-\infty, 0)$ and all x.
(c) For $x \leq 0$ we have $\rho(x) \geq 0$.
(d) Let ρ be a normed monetary risk measure. Show that two of the following properties always imply the third property

    (α) Convexity
    (β) Positive homogeneity
    (γ) Subadditivity

Exercise 18.12 *Suppose that an insurance portfolio has claims X_i which are exponentially distributed $\exp(\lambda)$. Suppose that from earlier analysis one has fixed $\lambda = 10$.*

(a) Suppose now that there are $n = 100$ such claims in this portfolio and one observes values larger than 50 and 100. How likely are such extreme losses?
(b) How could you proceed with extreme value theory (EVT)? How could you find the norming constant λ and $\log(n)$? Does it converge to limit?

Exercise 18.13 *Show that the function $L(x) = \log(1 + x)$ is a slowly varying function.*
    Hint: a positive measurable function L in $(0, \infty)$ that satisfies $\lim_{x \to \infty} \frac{L(tx)}{L(x)} = 1$ for all $t > 0$ is called a slowly varying function.

Exercise 18.14 *Let $X_1, \ldots X_n$ are i.i.d. random variables with a Pareto distribution with the cdf*

$$W_{(1,\alpha)}(x) = 1 - \frac{1}{x^\alpha}, x \geq 1, \alpha > 0$$

(a) Calculate $E\{w_{(1,\alpha)}(x)\}$.
(b) What is the cdf of $\min(x_1, \ldots x_n)$?

Chapter 19
Neural Networks

A neural network is a non-linear system that converts a series of real input values $x_1, \ldots, x_p$ over several intermediary steps to one or more terminal variables $y_1, \ldots, y_q$. It represents a function $\nu : \mathbb{R}^p \to \mathbb{R}^q$:

$$(y_1, \ldots, y_q)^\top = \nu(x_1, \ldots, x_p),$$

that has a special form given by the network structure. This is graphically displayed in Fig. 19.1 in the form of a directed graph whose nodes are grouped in various levels. In the *input layer* each node represents an input variable; in the *output layer* each node represents an output variable. In between there are one or more *hidden layers,* whose nodes are neither sources nor layers of the graph. The network in Fig. 19.1 contains only one hidden layer. In additional it is a *feed forward network*, since it contains no edges that begin in a node and end in the same node or in a different node from the same or a previous layer.

Figure 19.2 displays a *feedback network*, in that there is feedback among the nodes of the two hidden layers. In the following we will concentrate on the feed forward network.

Neural networks are used in financial statistics to represent functions, which, for example, can represent the default probability of a credit, the forecast of an exchange rate or the volatility of a stock. Here the emphasis is on non-parametric applications, which in comparison to the local smoothing function discussed in Chap. 15 require advanced modelling and can be quite involved to calculate. On the other hand it is still practical when numerous variables need to be considered in forecasts or quantifying risk, i.e. when the dimension p of the function arguments is large.
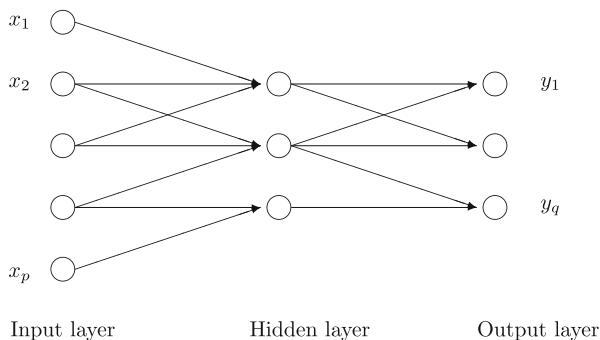
Fig. 19.1 Neural feed forward network with a hidden layer
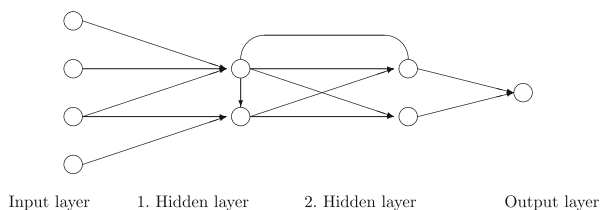
Fig. 19.2 Neural feedback network with two hidden layers

19.1 From Perceptron to Non-linear Neuron

The perceptron is a simple mathematical model of how a nerve cell functions in receiving signals from sense cells and other nerve cells (the input variables) and from this sends a signal to the next nerve cell or remains inactive. In spite of all of the disadvantages the perceptron is very influential on the way of thinking with respect to neural networks, so that it is a good starting point for the discussion of components from which neural networks are constructed. The perceptron works in two steps:

- the input variables $x_1, \ldots, x_p$ are multiplied and added with *weights* $w_1, \ldots, w_p$,
- a *threshold operation* is applied to the result.

$x = (x_1, \ldots, x_p)^\top$, $w = (w_1, \ldots, w_p)^\top$ represent the input vector, and weight vector respectively, and for a given b let $\psi(u) = \mathbf{1}(u > b)$ be the corresponding threshold function. The output variables $y = \psi(w^\top x)$ of the perceptron is 1 (the nerve cell "fires"), when the sum of the weighted input signals lies above the threshold and is 0 otherwise (the nerve cells remain inactive). The effect of the perceptron depends on the weights $w_1, \ldots, w_p$ and the threshold value b. An equivalent representation can be obtained by including the constant $x_0 \overset{\text{def}}{=} 1$ as an additional input variable, which has the weight $w_0 = -b$ and a threshold value of 0 is chosen since then
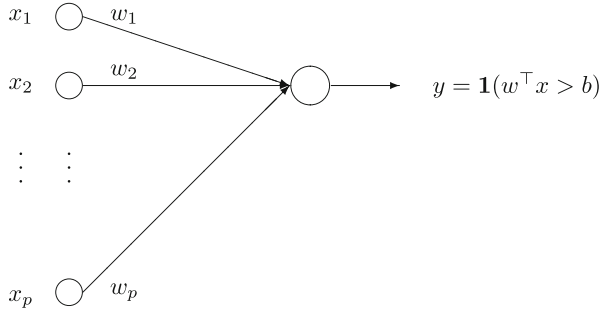
Fig. 19.3 The perceptron

$$\mathbf{1}\left(\sum_{i=1}^{p} w_i x_i > b\right) = \mathbf{1}\left(\sum_{i=0}^{p} w_i x_i > 0\right).$$

This representation is often more comfortable since with the system parameters that can be freely chosen one does not have to differentiate between weights and threshold values.

A perceptron can be trained to solve classification problems of the following type: Given are objects which belong to one of two classes, C_0 or C_1. Decisions are made based on observations of the object $x_1, \ldots, x_p$, whether it belongs to C_0 or C_1 (Fig. 19.3).

The perceptron characterized by the weights $w_0, \ldots, w_p$ classifies an object as belonging to C_0 respectively C_1 when the output variable $y = y(x_1, \ldots, x_p)$ is 0 respectively 1. So that the classification problem "may be" solved, the weights $w_0, \ldots, w_p$ must be "learned". To do this there is a *training set*

$$(x^{(1)}, z^{(1)}), \ldots, (x^{(T)}, z^{(T)})$$

from T input vectors

$$x^{(t)} = (x_1^{(t)}, \ldots, x_p^{(t)})^\top$$

available whose correct classification

$$z^{(1)}, \ldots, z^{(T)} \in \{0, 1\}$$

is known. With the help from *learning rules* suitable weights $\hat{w}_0, \ldots, \hat{w}_p$ are determined from the training set.

In statistical terms the problem is to estimate the parameters of the perceptron from the data $(x^{(t)}, z^{(t)})$, $t = 1, \ldots, T$. A *learning rule* is an estimation method which produces estimates $\hat{w}_0, \ldots, \hat{w}_p$.

A learning rule is, for example, the Delta or *Widrow–Hoff learning rule*: The input vectors $x^{(t)}, t = 1, \ldots, T$, are used consecutively as input variables of the perceptron and the output variables $y^{(t)}$, $t = 1, \ldots, T$, are compared to the correct classification $z^{(t)}$, $t = 1, \ldots, T$. If in one step $y^{(t)} = z^{(t)}$, then the weights remain unchanged. If on the other hand $y^{(t)} \neq z^{(t)}$, then the weight vector $w = (w_0, \ldots, w_p)^\top$ is adjusted in the following manner:

$$w_{new} = w + \eta(z^{(t)} - y^{(t)}) x^{(t)}$$

η is a small relaxation factor which must eventually slowly approach zero in order to ensure convergence of the learning algorithm. The initial value of w is arbitrarily given or randomly chosen, for example, uniformly distributed over $[0, 1]^{p+1}$.

The learning does not end when all of the input vectors are presented in the network, but rather after $x^{(T)}$ has been entered, $x^{(1)}$ is used again as the next input variable. The training set is tested multiple times until the network of all objects in the training set has been correctly identified or until a given quality criterion for measuring the error in classification is low enough.

The weights $w_0, \ldots, w_p$ can be identified up to a positive scale factor, i.e. for $\alpha > 0$, $\alpha w_0, \ldots, \alpha w_p$ lead to the same classification. By applying the learning rule, for example, the Widrow–Hoff, it is possible that $||w||$ will continuously increase; this can lead to numerical problems. In order to prohibit this, one uses the so-called *weight decay* technique, i.e. a modified learning rule in which $||w||$ remains stable.

Example 19.1 (Learning the OR-Function) Let $p = 2$ and $x_1, x_2 \in \{0, 1\}$. The classifications that needs to be learned is the logical OR:

$$z = 1, \text{ if } x_1 = 1 \text{ or } x_2 = 1,$$
$$z = 0, \text{ if } x_1 = 0 \text{ and } x_2 = 0.$$

The following input vectors, including the first coordinate $x_0 = 1$

$$x^{(1)} = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \quad x^{(2)} = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix}, \quad x^{(3)} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \quad x^{(4)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$$

are used as the training set with the correct classification $z^{(1)} = z^{(2)} = z^{(3)} = 1$, $z^{(4)} = 0$. The perceptron with the weights w_0, w_1, w_2 classifies an object as 1 if and only if

$$w_0 x_0 + w_1 x_1 + w_2 x_2 > 0 ,$$

and as 0 otherwise. For the starting vector we use $w = (0, 0, 0)^\top$, and we set $\eta = 1$. The individual steps of the Widrow–Hoff learning take the following form:

1. $w^{(1)}$ gives $y^{(1)} = 0 \neq z^{(1)}$. The weights are changed:
   $w_{new} = (0, 0, 0)^\top + (1 - 0)(1, 1, 0)^\top = (1, 1, 0)^\top$
2. $x^{(2)}$ is correctly classified with the weight vector.
3. $x^{(3)}$ is correctly classified with the weight vector.
4. For $x^{(4)}$ is $w^\top x^{(4)} = 1$, so that the weights are again changed:
   $w_{new} = (1, 1, 0)^\top + (0 - 1)(1, 0, 0)^\top = (0, 1, 0)^\top$
5. $x^{(1)}$ is now used as input and is correctly classified.
6. Since $w^\top x^{(2)} = 0$:
   $w_{new} = (0, 1, 0)^\top + (1 - 0)(1, 0, 1)^\top = (1, 1, 1)^\top$
7. Since $w^\top x^{(3)} = 3 > 0$, $x^{(3)}$ is correctly classified.
8. $x^{(4)}$ is incorrectly classified so that
   $w_{new} = (1, 1, 1)^\top + (0 - 1)(1, 0, 0)^\top = (0, 1, 1)^\top$

Thus the procedure ends since the perceptron has correctly identified all the input vectors in the training set with these weights. The perceptron learned the OR function over the set $\{0, 1\}^2$.

One distinguishes different types of learning for neural networks:

Supervised Learning: Compare the network outputs $y = y(x_1, \ldots, x_p)$ with the correct $z = z(x_1, \ldots, x_p)$. When $y \neq z$, the weights are changed according to the learning rule.

Reinforcement Learning: From every network output $y = y(x_1, \ldots, x_p)$ one discovers, whether it is "correct" or "incorrect"—in the latter case though one does not know the correct value. When y is "incorrect", the weights are changed according to the learning rule.

Unsupervised Learning: There is no feedback while learning. Similar to the cluster analysis, random errors are filtered from the data with the help of redundant information.

For $y \in \{0, 1\}$ supervised and reinforcement learning are the same. Included in this type is the Widrow–Hoff learning rule for the perceptron.

The perceptron cannot learn all of the desired classifications. The classical counter example is the logical argument XOR = "exclusive or":

$$z = 1, \text{ if either } x_1 = 1 \text{ or } x_2 = 1,$$
$$z = 0, \text{ if } x_1 = x_2 = 0 \text{ or } x_1 = x_2 = 1.$$

A perceptron with weights w_0, w_1, w_2 corresponds to a hyperplane $w_0 + w_1 x_1 + w_2 x_2 = 0$ in $\mathbb{R}^2$ space of the inputs $(x_1, x_2)^\top$, which separates the set using the perceptron of 0 classified objects from those classified as 1. It is not hard to see that no hyperplane exists for "exclusive or" where inputs should be classified as 1 $\binom{1}{0}, \binom{0}{1}$ can be separated from those to be classified as 0 $\binom{0}{0}, \binom{1}{1}$.

Definition 19.1 (Linearly Separable) For $p \geq 1$ to subsets $\mathcal{X}_0, \mathcal{X}_1 \subseteq \mathbb{R}^p$ are called *linearly separable* if $w \in \mathbb{R}^p$, $w_0 \in \mathbb{R}$ exists with

$$w_0 + w^\top x > 0 \text{ for } x \in \mathcal{X}_1,$$
$$w_0 + w^\top x \leq 0 \text{ for } x \in \mathcal{X}_0.$$

The perceptron with p input variables $x_1, \ldots, x_p$ (with respect to the constant $x_0 \stackrel{\text{def}}{=}$ 1) can learn the classification exactly that is consistent with the linearly separable sets of inputs.

If no perfect classification is possible through a perceptron, then one can at least try to find a "good" classification, that is, to determine the weights $w_0, \ldots, w_p$ so that a measurement for the amount of incorrectly identified classifications can be minimized. An example of such an application is given by the *least squares (LS) classification:*

Assuming that the training set $(x^{(1)}, z^{(1)}), \ldots, (x^{(T)}, z^{(T)})$ is given. Determine for some given weight w_0 the weights $w_1, \ldots, w_p$ so that

$$Q(w) = Q(w_1, \ldots, w_p) = \sum_{i=1}^{T} (z^{(i)} - y^{(i)})^2 = \min !$$

$$\text{with } y^{(i)} = \mathbf{1}(w_0 + w^\top x^{(i)} > 0), \ w = (w_1, \ldots, w_p)^\top$$

w_0 can be arbitrarily chosen since the weights $w_0, \ldots, w_p$ described above are only determined up to a scale factor. In the case of the perceptron, which takes on a binary classification, $Q(w)$ is simply the number of incorrectly defined classifications. The form mentioned above can also be directly applied to other problems. The attainable minimum of $Q(w)$ is exactly 0 (perfect classification of the training set) when both sets

$$\mathcal{X}_0^{(T)} = \{x^{(i)}, \ i \leq T; \ z^{(i)} = 0\}, \ \mathcal{X}_1^{(T)} = \{x^{(i)}, \ i \leq T; \ z^{(i)} = 1\}$$

are linearly separable.

The Widrow–Hoff learning rule solves the LS classification problem; there are, however, a series of other learning rules or estimation methods which can also solve the problem. The perceptron has proven to be too inflexible for many applications. Therefore, one considers general forms of *neurons* as components used to build a neuron network:

Let $x = (x_1, \ldots, x_p)^\top$, $w = (w_1, \ldots, w_p)^\top$ be input and weight vectors, respectively. For $\beta, \beta_0 \in \mathbb{R}$

$$\psi_\beta(t) = \left\{ 1 + \exp\left(-\frac{t + \beta}{\beta_0} \right) \right\}^{-1}$$
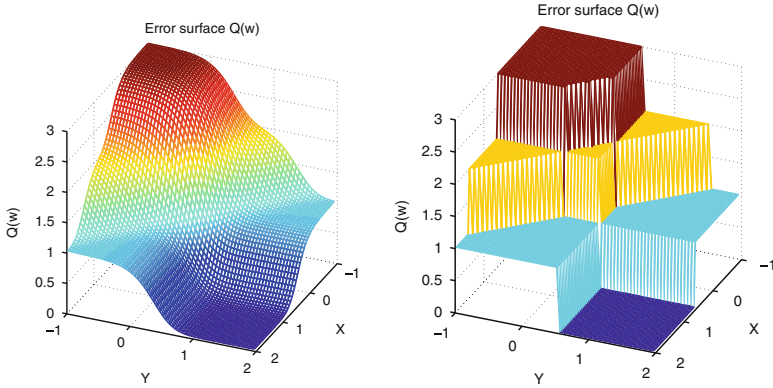
Fig. 19.4 Error surface of $Q(w)$ given weight $w = (w_1, w_2)^\top$ with transform function: threshold function (*left*) and sigmoid function (*right*)   SFEerrorsurf
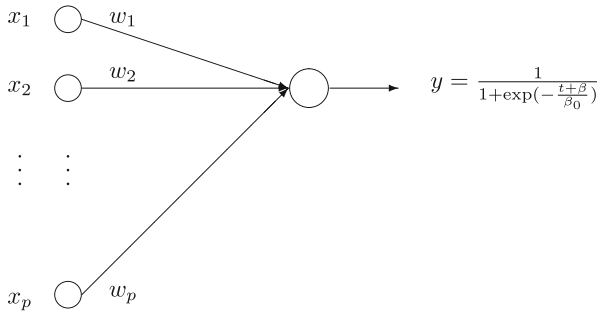
Fig. 19.5 Neuron with a sigmoid transformation function

is the logistic function, which due to its form is often referred to as "the" *sigmoid function*. One can also use other functions with sigmoid forms, for example, the density function of a normal distribution. The output variable of the neuron is $y = \psi_\beta(w^\top x)$.

For $\beta_0 \to 0+$ $\psi_\beta(t)$ approaches a threshold function:

$$\psi_\beta(t) \longrightarrow \mathbf{1}(t + \beta > 0) \qquad \text{for } \beta_0 \longrightarrow 0+ ,$$

so that the preceptron is a boundary of the neuron with a logistic activation function. An example of $Q(w)$ for neurons with threshold function and sigmoid function as activation function is shown in Fig. 19.4. The corresponding method is presented in Fig. 19.5.

The parameter β_0 is often not explicitly chosen, since it can be integrated as a scale factor in the other parameters $w_1, \ldots, w_p,$ β of the neurons. If one also sets $w_0 = \beta$ and $x_0 \stackrel{\text{def}}{=} 1$, then the output variables can also be written in the form:
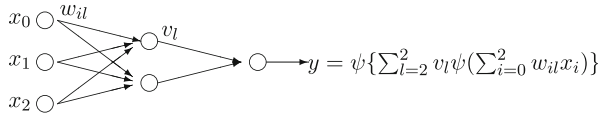
Fig. 19.6 Multiple layer perceptron with a hidden layer
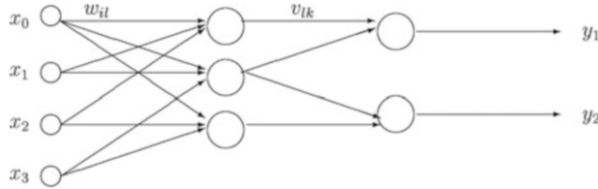
Fig. 19.7 Multiple layer perceptron with two hidden layers

$$y = \psi(w_0 + w^\top x) = \psi(\sum_{k=0}^{p} w_k x_k) \text{ with } \psi(t) = \frac{1}{1 + e^{-t}}.$$

By combining multiple neurons with sigmoid or—in the limit case—threshold activation functions with a feed forward network one obtains a so-called *multiple layer perceptron* (MLP) neural network. Figure 19.6 shows such a neural network with two input variables with respect to the constant $x_0 \stackrel{\text{def}}{=} 1$, two sigmoid neurons in the hidden layer that are connected by another sigmoid neuron to the output variables, where $\psi(t) = \{1 + e^{-t}\}^{-1}$ as above.

Neural networks can also be constructed with multiple hidden layers that give multiple output variables. The connections do not have to be complete, i.e. edges between the nodes of consecutive layers may be missing or equivalently several weights can be set to 0. Instead of the logical function or similar sigmoid functions, threshold functions may also appear in some neurons. Another probability is the so-called *radial basis functions* (RBF). To the former belongs the density of the standard normal distribution and similar symmetrical kernel functions. In this situation one no longer speaks of an MLP network, but of an RBF network.

Figure 19.7 shows an incomplete neural network with two output variables. The weights $w_{13}, w_{22}, w_{31}, v_{12}$ and v_{31} are set to 0, and the corresponding edges are not displayed in the network graphs. The output variable y_1 is, for example

$$y_1 = v_{11}\psi(w_{01} + w_{11}x_1 + w_{21}x_2) + v_{21}\psi(w_{02} + w_{12}x_1 + w_{32}x_3),$$

a linear combination of the results of the two upper neurons of the hidden layers.

Until now we have only discussed those cases that are most often handled in the literature, where a neuron has an effect on the linear combination of variables from the previous layer. Occasionally the case where the output of a neural of the form $\psi\left(\Pi_{i=1}^{p} w_i x_i\right)$ respectively $\psi\left(\max_{i=1,\dots,p} x_i\right)$ is considered.

Neural networks of MLP types can be used for classification problems as well as for regression and forecast problems. In order to find an adequate network for each problem, the weights have to be learned through a training set, i.e. the network parameters are estimated from the data. Since we are restricting ourselves to the case of supervised learning, this means that $(x^{(1)}, z^{(1)}), \ldots, (x^{(T)}, z^{(T)})$ are given for the training set. The $x^{(i)} \in \mathbb{R}^p$ are input vectors, the $z^{(i)} \in \mathbb{R}^q$ are the corresponding desired output values from the network. The vectors $z^{(i)}$ are compared to the actual output vectors $y^{(i)} \in \mathbb{R}^q$ of the network. The weights are determined so that the deviations between $z^{(i)}$ and $y^{(i)}$ are small. An example of this is the *least squares (LS) application* already mentioned in the analysis of the perceptron:

Assuming that the training set $(x^{(1)}, z^{(1)}), \ldots, (x^{(T)}, z^{(T)})$ is given. The weights $w_{0l}, l = 1, \ldots, r, x_0 \overset{\text{def}}{=} 1$ are given, where r is the number of neurons in the first hidden layer. The weights of all the other edges in the network (between the input layer, the hidden layers and the output layer) are determined so that

$$\sum_{k=1}^{T} ||z^{(k)} - y^{(k)}||^2 = \min!$$

In the network given in Fig. 19.7 the minimization is carried out with respect to the weights $w_{11}, w_{12}, w_{21}, w_{23}, w_{32}, w_{33}, v_{11}, v_{21}, v_{22}, v_{32}$. As for the perceptron the weights w_{01}, w_{02}, w_{03} can be set in order to avoid the arbitrary choice of scale factors.

Instead of the LS method, other loss functions can also be minimized, for example, weighted quadratic distances or, above all in classification, the Kullback–Leibler distance:

$$\sum_{k=1}^{T} \sum_{i} \left\{ z_i^{(k)} \log \frac{z_i^{(k)}}{y_i^{(k)}} + (1 - z_i^{(k)}) \log \frac{1 - z_i^{(k)}}{1 - y_i^{(k)}} \right\} = \min!$$

Since only the $y_i^{(k)}$ depend on the weights, it is equivalent to minimise the cross-entropy between z_i and y_i, which are both contained in $(0, 1)$:

$$-\sum_{k=1}^{T} \sum_{i} \left\{ z_i^{(k)} \log y_i^{(k)} + (1 - z_i^{(k)}) \log(1 - y_i^{(k)}) \right\} = \min!$$

19.2 Back Propagation

The most well-known method with which the feed forward network learns its weights from the training set is the back propagation. The basic idea is none other than a numerical method to solve the (non-linear) least squares problem that saves on memory, at the cost however, of possible slower convergence and numerical instabilities.

To illustrate, consider a neural network with an output variable y (i.e. $q = 1$) and a hidden layer with only one neuron:

$$y = \psi(w_0 + w^\top x).$$

ψ can be a logistic function, or some other transformation function. The training set is $(x^{(1)}, z^{(1)}), \dots, (x^{(T)}, z^{(T)})$. The weight w_0 is held constant in order to avoid the arbitrary scale factor. The function to be minimized

$$Q(w) = \sum_{k=1}^{T} (z^{(k)} - y^{(k)})^2$$

is thus only dependent on the weights $w_1, \dots, w_p$ of the input variables.

An elementary numerical method for minimizing Q is the *decreasing gradient method*. Given a weight $w(N)$ one calculates the next approximation by moving a small step in the direction of the steepest decline of Q:

$$w(N + 1) = w(N) - \eta \operatorname{grad} Q\{w(N)\},$$

$$\operatorname{grad} Q(w) = -\sum_{k=1}^{T} 2(z^{(k)} - y^{(k)})\psi'(w^\top x^{(k)})x^{(k)}.$$

To accelerate the convergence, the small constant $\eta > 0$ can also converge to 0 during the iteration process. Figure 19.8 shows the path of optimizing $Q(w)$
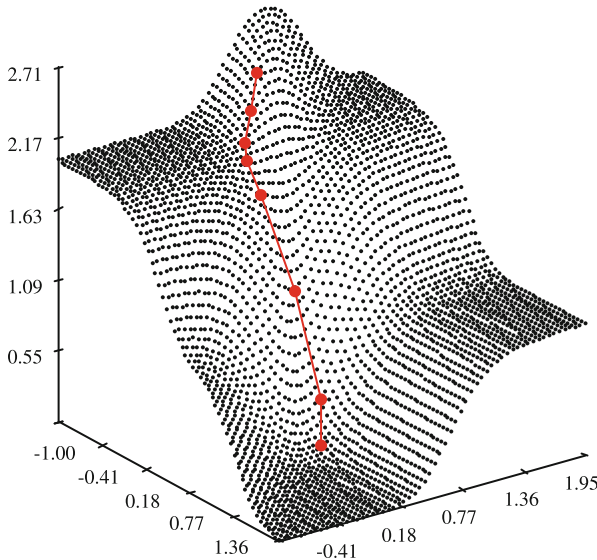
Fig. 19.8 Gradients descent proceedings with goal vector $z = (0, 1, 0, 1)^\top$ ▣ SFEdescgrad

evaluated in i-steps at $w_1, \ldots, w_i$, where each w is corrected according to the back propagation rule.

With the decreasing gradient method the quality of the weight $w(N)$, that is the actual network, is evaluated simultaneously using all the data in the training set. The network is applied to all $x^{(1)}, \ldots, x^{(T)}$, and only after this the weight vector changed.

Back propagation is also a form of the decreasing gradient method with the difference that the network is repeatedly applied to the single $x^{(k)}$ and after every single step the weights are already changed in the direction of the steepest decline of the function $Q_k(w) = (z^{(k)} - y^{(k)})^2$:

$$w(N + 1) = w(N) - \eta \operatorname{grad} Q_k\{w(N)\} , \; k = 1, \ldots, T,$$
$$\operatorname{grad} Q_k(w) = -2(z^{(k)} - y^{(k)})\psi'(w^\top x^{(k)})x^{(k)}.$$

If in this process the training set has been gone through once, the iteration starts again from the beginning. T steps in the back propagation correspond then roughly to one step in the decreasing gradient method. Also by the back propagation algorithm it may be necessary to allow η to converge slowly to 0.

The Widrow–Hoff learning rule is in principle a back propagation algorithm. The threshold function $\psi(t) = \mathbf{1}(w_0 + t > 0)$ is non-differentiable, but after the presentation of $x^{(k)}$ the weights are changed in the direction of the steepest decline of $Q_k(w)$, i.e. in the direction of $x^{(k)}$ for $z^{(k)} = 1$, $y^{(k)} = 0$ and in the direction of $-x^{(k)}$ for $z^{(k)} = 0$, $y^{(k)} = 1$. By correct classifications the weights here remain unaltered.

Naturally one can apply every numerical algorithm that can calculate the minimum of a non-linear function $Q(w)$ to determine the weights of a neural network. By some applications, for example, the conjugate gradient method has proven to be the fastest and most reliable method. All of these algorithms have the danger of landing in a local minimum of $Q(w)$. In the literature on neural networks it is occasionally claimed that with the combination of training, the networks, such as local minima do not occur. Based on experience of statistics with maximum likelihood estimators of large dimensional parameters, this is to be expected since the training of neural networks for applications of regression analysis, for example, can be interpreted under the appropriate normality assumptions as the maximum likelihood estimation technique.

19.3   Neural Networks in Non-parametric Regression Analysis

Neural networks of type MLP describe a mapping of the input variables $x \in \mathbb{R}^p$ onto the output variables $y \in \mathbb{R}^q$. We will restrict ourselves in this section to the case where the network has only one hidden layer and the output variable is univariate ($q = 1$). Then $y \in \mathbb{R}$ as a function of x has the form

$$y = v_0 + \sum_{h=1}^{H} v_h \psi \left(w_{0h} + \sum_{j=1}^{p} w_{jh} x_j \right) \overset{\text{def}}{=} v_H(x; \vartheta) \tag{19.1}$$

where H is the number of neurons in the hidden layer and ψ is the given transformation function. The parameter vector

$$\vartheta = (w_{01}, \ldots, w_{p1}, w_{02}, \ldots, w_{pH}, v_0, \ldots, v_H)^{\top} \in \mathbb{R}^{(p+1)H+H+1}$$

contains all the weights of the network. This network with one hidden layer already has a universal approximation property: every measurable function $m : \mathbb{R}^p \to \mathbb{R}$ can be approximated as accurately as one wishes by the function $v_H(x, \vartheta)$ when ψ is a monotone increasing function with a bounded range. More precisely, the following result holds, Hornik, Stinchcombe and White (1989):

Theorem 19.1 *Let* $\psi : \mathbb{R} \to [0, 1]$ *be monotone increasing with* $\lim_{u \to -\infty} \psi(u) = 0$, $\lim_{u \to \infty} \psi(u) = 1$, *and let* $J = \{v_H(x; \vartheta); \ H \geq 1, \vartheta \in \mathbb{R}^{(p+1)H+H+1}\}$ *be the set which is mapped by a MLP function with a hidden layer from* $\mathbb{R}^p$ *to* $\mathbb{R}$.

(a) For every Borel measurable function $f : \mathbb{R}^p \to \mathbb{R}$ *there a series* $v_n \in J$, $n \geq 1$ *exists, with* $\mu(x; |f(x) - v_n(x)| > \varepsilon) \longrightarrow 0$ *for* $n \to \infty$, $\varepsilon > 0$, *where* μ *is an arbitrary probability measure of the Borel-σ-Algebra from* $\mathbb{R}^p$.

(b) For every increasing function $f : \mathbb{R}^p \to \mathbb{R}$ *there a series* $v_n \in J$, $n \geq 1$ *exists, with* $\sup_{x \in C} |f(x) - v_n(x)| \longrightarrow 0$ *for* $n \to \infty$, *where* C *is an arbitrary compact subset of* $\mathbb{R}^p$.

The range of ψ can be set to any bounded interval, not only $[0, 1]$, without changing the validity of the approximation properties.

The weight vector ϑ is not uniquely determined by the network function v_H. If, for example, the transformation function is asymmetric around 0, i.e. $\psi(-u) = -\psi(u)$, then $v_H(x; \vartheta)$ does not change when

(a) the neurons of the hidden layer are interchanged, which corresponds to a substitution of the coordinates of ϑ, or when

(b) all input weights $w_{0h}, \ldots, w_{ph}$ and the output weight v_h of the neural are multiplied by -1.

In order to avoid this ambiguity we will restrict the parameter set to a fundamental set in the sense of Rüeger and Ossen (1997), which for every network function $v_H(x; \vartheta)$ contains exactly one corresponding parameter vector ϑ. In the case of asymmetric transformation functions we restrict ourselves, for example, to weight vectors with $v_1 \geq v_2 \geq \ldots \geq v_H \geq 0$. In order to simplify the following considerations we also assume that ϑ is contained in a sufficiently large compact subset $\Theta_H \subset \mathbb{R}^{(p+1)H+H+1}$ of a fundamental range.

Due to their universal approximation properties neural networks are a suitable tool in constructing non-parametric estimators for regression functions. For this we consider the following heteroscedastic regression model:

$$Z_t = f(X_t) + \varepsilon_t , \quad t = 1, \ldots, n,$$

where $X_1, \ldots, X_n$ are independent, identically distributed d-variate random variables with a density of $p(x)$, $x \in \mathbb{R}^d$. The residuals $\varepsilon_1, \ldots, \varepsilon_n$ are independent, real valued random variables with

$$\mathsf{E}(\varepsilon_t | X_t = x) = 0, \quad \mathsf{E}(\varepsilon_t^2 | X_t = x) = s_\varepsilon^2(x) < \infty.$$

We assume that the conditional mean $f(x)$ and the conditional variance $s_\varepsilon^2(x)$ of Z_t are, given $X_t = x$, continuous functions bounded to $\mathbb{R}^d$. In order to estimate the regression function f, we fit a neural network with a hidden layer and a sufficiently large number, H, of neurons to the input variables $X_1, \ldots, X_n$ and the values $Z_1, \ldots, Z_n$, i.e. for given H we determine the non-linear least squares estimator $\hat{\vartheta}_n = \mathrm{argmin}_{\vartheta \in \Theta_H} D_n(\vartheta)$ with

$$\hat{D}_n(\vartheta) = \frac{1}{n} \sum_{t=1}^{n} \{Z_t - \nu_H(X_t; \vartheta)\}^2 .$$

Under appropriate conditions $\hat{\vartheta}_n$ converges in probability for $n \to \infty$ and a constant H to the parameter vector $\vartheta_0 \in \Theta_H$, which corresponds to the best approximation of $f(x)$ by a function of type $\nu_H(x; \vartheta)$, $\vartheta \in \Theta_H$:

$$\vartheta_0 = \mathrm{argmin}_{\vartheta \in \Theta_H} D(\vartheta) \quad \text{with} \quad D(\vartheta) = \mathsf{E}\{f(X_t) - \nu_H(X_t; \vartheta)\}^2.$$

Under somewhat stronger assumptions the asymptotic normality of $\hat{\vartheta}_n$ and thus of the estimator $\hat{f}_H(x) = \nu_H(x; \hat{\vartheta}_n)$ also follows for the regression function $f(x)$.

The estimation error $\hat{\vartheta}_n - \vartheta_0$ can be divided into two asymptotically independent sub-components: $\hat{\vartheta}_n - \vartheta_0 = (\hat{\vartheta}_n - \vartheta_n) + (\vartheta_n - \vartheta_0)$, where the value

$$\vartheta_n = \mathrm{argmin}_{\vartheta \in \Theta_H} \frac{1}{n} \sum_{t=1}^{n} \{f(X_t) - \nu_H(X_t; \vartheta)\}^2$$

minimizes the sample version of $D(\vartheta)$, Franke and Neumann (2000):

Theorem 19.2 *Let ψ be bounded and twice differentiable with a bounded derivative. Suppose that $D(\vartheta)$ has a unique global minimum ϑ_0 in the interior of Θ_H, and the Hesse matrix $\nabla^2 D(\vartheta_0)$ of D at ϑ_0 is positive definite. In addition to the above mentioned conditions for the regression model it holds that*

$$0 < \delta \leq s_\varepsilon^2(x) \leq \Delta < \infty \text{ for all } x,$$
$$\mathsf{E}(|\varepsilon_t|^\gamma | X_t = x) \leq C_\gamma < \infty \text{ for all } x, \gamma \geq 1$$

with suitable constants $\delta, \Delta, C_n, \gamma \geq 1$. Then it holds for $n \to \infty$:

$$\sqrt{n} \begin{pmatrix} \hat{\vartheta}_n - \vartheta_n \\ \vartheta_n - \vartheta_0 \end{pmatrix} \xrightarrow{\mathcal{L}} N\left(0, \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}\right)$$

with covariance matrices

$$\Sigma_i = \{\nabla^2 D(\vartheta_0)\}^{-1} B_i(\vartheta_0) \{\nabla^2 D(\vartheta_0)\}^{-1}, i = 1, 2,$$

$$B_1(\vartheta) = 4 \int s_\varepsilon^2(x) \, \nabla v_H(x; \vartheta) \nabla v_H(x; \vartheta)^\top \, p(x) dx$$

$$B_2(\vartheta) = 4 \int \{f(x) - v_H(x; \vartheta)\}^2 \nabla v_H(x; \vartheta) \nabla v_H(x; \vartheta)^\top p(x) dx$$

where ∇v_H represents the gradient of the network function with respect to the parameter ϑ.

From the theorem it immediately follows that $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)$ is asymptotically $N(0, \ \Sigma_1 + \Sigma_2)$ distributed. Σ_1 here stands for the variability of the estimator $\hat{\vartheta}_n$ caused by the observational error ε_t. Σ_2 represents the proportion of asymptotic variability that is caused by the mis-specification of the regression function, i.e. from the fact that $f(x)$ is of the form $v_H(x; \vartheta)$ for a given H and no ϑ. In the case that it is correctly specified, where $f(x) = v_H(x; \vartheta_0)$, this covariance component disappears, since $B_2(\vartheta_0) = 0$ and $\Sigma_2 = 0$.

Σ_1, Σ_2 can be estimated as usual with the sample covariance matrices. In order to construct tests and confidence intervals for $f(x)$ a couple of alternatives to the asymptotic distribution are available: Bootstrap, or in the case of heteroscedasticity, the Wild Bootstrap method, Franke and Neumann (2000).

Theorem 19.2 is based on the theoretical value of the least squares estimator $\hat{\vartheta}_n$, which in practice must be numerically determined. Let $\tilde{\vartheta}_n$ be such a numerical approximation of $\hat{\vartheta}_n$. The quality of the resulting estimator $\tilde{\vartheta}_n$ can depend on the numerical method used. White (1989b) showed in particular that the back propagation algorithm leads, under certain assumptions, to an asymptotically inefficient estimator $\tilde{\vartheta}_n$, i.e. the asymptotic covariance matrix of $\sqrt{n}(\tilde{\vartheta}_n - \vartheta_0)$ is larger than that of $\sqrt{n}(\hat{\vartheta}_n - \vartheta_0)$ in the sense that the difference of the two matrices is positive definite. Nevertheless White also showed that by joining a single global minimization step, the estimator calculated from the back propagation can be modified so that for $n \to \infty$ it is as efficient as the theoretical least squares estimator $\hat{\vartheta}_n$.

Until now we have held the number of neurons H in the hidden layer of the network and thus the dimension of the parameter vector ϑ constant. The estimator based on the network, $\hat{f}_H(x) = v_H(x; \hat{\vartheta}_n)$ converges to $v_H(x; \vartheta_0)$, so that in general the bias $\mathsf{E}\{\hat{f}_H(x)\} - m(x)$ for $n \to \infty$ does not disappear, but rather

converges to $v_H(x; \vartheta_0) - f(x)$. With standard arguments it directly follows from Theorem 19.2 that:

Corollary 19.1 *Under the assumptions from Theorem 19.2 it holds for $n \to \infty$ that*

$$\sqrt{n} \left\{ v_H(x; \hat{\vartheta}_n) - v_H(x; \vartheta_0) \right\} \xrightarrow{\mathcal{L}} N(0, \sigma_\infty^2)$$

with $\sigma_\infty^2 = \nabla v_H(x; \vartheta_0)^\top (\Sigma_1 + \Sigma_2) \nabla v_H(x; \vartheta_0)$.

In order to obtain a consistent estimator for $f(x)$, the number of neurons H, which by the non-parametric estimator $\hat{f}_H(x)$ plays the role of a smoothing parameter, must increase with n. Due to the universal approximation properties of the neural network $v_H(x; \vartheta_0)$ thus converges to $f(x)$, so that the bias disappears asymptotically. Since with an increasing H the dimension of the parameter vector ϑ increases, H should not approach ∞ too quickly, in order to ensure that the variance of $\hat{f}_H(x)$ continues to converge to 0. In choosing H in practice one uses a typical dilemma for non-parametric statistics, the bias variance dilemma: a small H results in a smooth estimation function $\hat{f}_H$ with smaller variance and larger bias, whereas a large H leads to a smaller bias but a larger variability of a then less smoothing estimator $\hat{f}_H$.

White (1990) showed in a corresponding framework that the regression estimator $\hat{f}_H(x)$ based on the neural network converges in probability to $f(x)$ and thus is consistent when $n \to \infty$, $H \to \infty$ at a slower rate.

From this it follows that neural networks with a free choice of H neurons in the hidden layer provide useful non-parametric function estimators in regression, and as we will discuss in the next section, in time series analysis. They have the advantage that the approximating function $v_H(x; \vartheta)$ of the form (19.1) is a combination of the neurons, which are composed of only a given non-linear transformation of an affine-linear combination of the variables $x = (x_1, \ldots, x_d)^\top$. This makes the numerical calculation of the least squares estimator for ϑ possible even when the dimension d of the input variables and the number H of neurons are large and thus the dimension $(d + 1)H + H + 1$ of the parameter vector is very large. In contrast to the local smoothing technique introduced in Chap. 15, the neural networks can also be applied as estimators of functions in large dimensional spaces. One reason for this is the non-locality of the function estimator $\hat{f}_H(x)$. This estimator does not dependent only on the observations (X_t, Z_t) with a small norm $||X_t - x||$ and thus in practice it is not as strongly afflicted by the imprecation of dimensionality, i.e. even for large n there is a smaller local density of the observation X_t in large dimensional spaces.

Theoretically it is sufficient to consider neural networks of type MLP with one hidden layer. In practice, however, one can sometimes achieve a comparably good fit to the data with a relatively more parsimonious parameterization by creating

multiple hidden layers. A network function with two hidden layers made up of H and G neurons respectively has, for example, the following form

$$
v(x; \vartheta) = v_0 + \sum_{g=1}^{G} v_g \psi \left\{ w'_{0g} + \sum_{h=1}^{H} w'_{hg} \psi \left(w_{0h} + \sum_{j=1}^{d} w_{jh} x_j \right) \right\},
$$

where ϑ represents the vector of all the weights v_g, w'_{hg}, w_{jh}. Such a function with small H, G can produce a more parsimonious parameterized approximation of the regression function $f(x)$ than a network function with only one hidden layer made up of a large number of neurons.

In addition the number of parameters to be estimated can be further reduced when several connections in the neural network are cut, i.e. when the corresponding weights are set to zero from the very beginning. The large flexibility that the neural network offers when approximating regression functions creates problems when creating the model, since one has to decide on a network structure and thus ask:

1. How many hidden layers does the network have?
2. How many neurons does each hidden layer have?
3. Which nodes (inputs, hidden neurons, outputs) of the network should be connect, i.e. which weights should be set to zero from the very beginning?

Through this process one is looking for a network which makes it possible to have a network function $v(x; \vartheta)$ that is parsimoniously parameterized and at the same time for a suitable ϑ that is a sufficiently good approximation of the regression function $f(x)$.

Similar to the classical linear regression analysis there are a comprehensive number of instruments available for specifying a network structure consistent with the data. For simplicity we will concentrate on the feed forward network with only one hidden layer made up of H neurons.

(a) *Repeated Significance Tests:* As with the stepwise construction of a linear regression model we start with a simple network assuming that one additional neuron with the number H and v_H output weights has been added. Whether in doing this the quality of the fit of the network has significantly improved is determined by testing the hypothesis $H_0 : v_H = 0$ against the alternative $H_1 : v_H \neq 0$. Since under H_0 the input weights $w_{0H}, \ldots, w_{pH}$ of the neurons in question are not identifiable, i.e. they have no influence on the value of the network function v_H, this is not a standard testing problem. White (1989a), Teräsvirta, Lin and Granger (1993) have developed Lagrange multiplier tests that are suitable for testing the significance of an additional neuron. Going in the other direction it is also possible to start with a complex network with large H assumed neurons and successively removing them until the related test rejects the hypothesis $H_0 : v_H = 0$. To reduce the number of parameters it makes sense to cut individual input connections, i.e. to set the corresponding weight to zero. For the test of the hypothesis $H_0 : w_{jh} = 0$ against the alternative

$H_1 : w_{jh} \neq 0$ classical Wald Tests can be applied due to the asymptotical results such as Theorem 19.2 (see for example Anders (1997) for applications in financial statistics).

(b) *Cross Validation and Validation:* The resulting cross validation is usually eliminated due to the extreme amount of calculations to determine the order of the model, i.e. first of all the number H of neurons in the hidden layer. In order to calculate the leave-one-out estimator for the model parameters one must fit the neural network to the corresponding sample that has been reduced by one observation a total of n times, and this must be done for every network structure under consideration. A related and more known procedure from the application of neural networks in the regression and time series analysis is to take a portion of the data away from the sample in order to measure the quality of the model based on this so-called validation set. In addition to the data (X_t, Z_t), $t = 1, \ldots, n$, used to calculate the least squares estimator $\hat{\vartheta}_n$ a second independent subsample (X_t, Z_t), $t = n + 1, \ldots, n + M$, is available. By minimizing measurements of fit, such as,

$$V(H) = \frac{1}{M} \sum_{t=n+1}^{n+M} \left\{ Z_t - v_H(X_t; \hat{\vartheta}_n) \right\}^2$$

the order of the model H and the quality of the incomplete network structure can be determined, in which individual input weights have been set to zero.

(c) *Network Information Criteria:* To compare the network structures some well-known applications for determining order, such as the Akaike Information Criterion (AIC), can be used. The application from Murata, Yoskizawa and Amari (1994) called the *Network Information Criterion* (NIC) is specialized for the case of neural networks. Here it is implicitly assumed that the residuals ε_t are normally distributed with a common variance σ_ε^2.

19.4 Forecasts of Financial Time Series with Neural Networks

To forecast the future development of financial time series an autoregressive model is particularly suitable. The value of the time series at date $t + 1$ is a function of many infinite observations from the past, in addition to an innovation independent of the past:

$$Z_{t+1} = f(Z_t, \ldots, Z_{t-p+1}) + \varepsilon_{t+1}, \quad -\infty < t < \infty, \tag{19.2}$$

where ε_t, $-\infty < t < \infty$, is independently and identically distributed with $\mathsf{E}(\varepsilon_t) = 0$, $\mathsf{Var}(\varepsilon_t) = \sigma_\varepsilon^2 < \infty$. The analogy of this formula for this non-linear autoregressive model of order p (NLAR(p)) to the regression model considered in a previous section is obvious, where the p-variate random vector $(Z_t, \ldots, Z_{t-p+1})^\top$ takes the place of the d-variate independent variable X_t. The autoregression function

$f : \mathbb{R}^p \rightarrow \mathbb{R}$ in this model immediately gives the best forecast for Z_{t+1} given the value of the time series up to date t:

$$\hat{Z}^0_{t+1|t} = f(Z_t, \ldots, Z_{t-p+1}).$$

Since f is in general not known, it seems obvious in view of the last section to approximate the autoregression function with a neural network when observations of the times series $Z_1, \ldots, Z_{n+1}$ are available. For training the network, i.e. for estimating the network weights, the vector $(Z_t, \ldots, Z_{t-p+1})^\top$ is used as input values and as output values Z_{t+1} for $t = p, \ldots, n$, is used. We will restrict ourselves for simplicity to the MLP with one hidden layer. $\hat{\vartheta}_n$ again represents the least squares estimator for the weight vector:

$$\hat{\vartheta}_n = \operatorname*{argmin}_{\vartheta \in \Theta_H} \frac{1}{n-p+1} \sum_{t=p}^{n} \left\{ Z_{t+1} - \nu_H \left(Z_t, \ldots, Z_{t-p+1}; \vartheta \right) \right\}^2$$

where ν_H is defined as in the previous section. We thus obtain a non-parametric forecast based on a neural network for Z_{t+1}:

$$\hat{Z}_{t+1|t} = \nu_H(Z_t, \ldots, Z_{t-p+1}; \hat{\vartheta}_n).$$

The result of this procedure is illustrated in Fig. 19.9: it shows the forecasting of the exchange rate time series JPY/USD using neural networks considering three periods of time dependency.
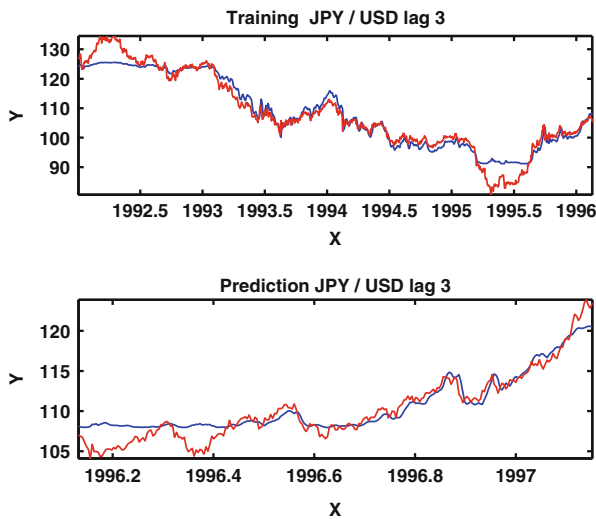
Fig. 19.9 Approximation of exchange rate JPY/USD (*red*) through RBF neural network (*blue*): Training set (*above*) and forecasts (*below*)   SFEnnjpyusd

The asymptotic normality of the parameters and of the function estimators and the consistency of $\nu_H(\cdot; \hat{\vartheta}_n)$ as an estimator of f for an increasing H remain robust even in the event where the stochastic process $\{Z_t, \ -\infty < t < \infty\}$ is α-mixing with exponentially decreasing mixing coefficients, White (1989b) and White (1990). Franke, Kreiss, Mammen and Neumann (2003) have formulated conditions for the case where $p = 1$ for the autoregression function f and for the distribution of the innovations ε_t, which guarantee for the NLAR(1) process the strongest β-mix properties with exponentially decreasing coefficients. Next to technical details it is essential that

$$\lim_{|x| \to \infty} |f(x)/x| < 1$$

is fulfilled, because it is sufficient for the innovation distribution that the density does not vanish anywhere. The last condition can be considerably weakened.

The conditions on the autoregression function is comparatively weak and obvious when one considers the stationarity conditions $|\alpha| < 1$ for linear AR(1) processes $Z_{t+1} = \alpha Z_t + \varepsilon_{t+1}$, where $f(x) = \alpha x$. Accordingly also for NLAR(p) process of large order ($p > 1$) it is sufficient to use weaker conditions on f, which above all guarantees stationarity in order to make the neural network a useful tool as a non-parametric estimator of f.

For the practical forecast one not only wants to use the last values in the time series, but also economic data available at time t such as exchange rates, index values, oil prices or the non-linear transformation of prices. To do this the non-linear autoregressive process with exogenous components of order p (NLARX(p)) process is suitable:

$$Z_{t+1} = f(Z_t, \dots, Z_{t-p+1}, X_t) + \varepsilon_t, \ -\infty < t < \infty, \tag{19.3}$$

where the innovations $\varepsilon_t, \ -\infty < t < \infty$, are again independently and identically distributed with $\mathsf{E}(\varepsilon_t) = 0$, $\mathsf{Var}(\varepsilon_t) = \sigma_\varepsilon^2 < \infty$, and X_t is the value of a d-variate stochastic process that contains all external information available at date t, which is used in the forecast.

The practical application of the forecast on financial time series with neural networks is illustrated with a pilot study that was carried out in cooperation with Commerzbank AG, Franke (1999). The goal was to develop a trading strategy for a portfolio made up of 28 of the most important stocks from the Dutch CBS-Index. We will restrict ourselves here to the buy-and-hold strategy with a time horizon of a quarter of a year (60 trading days), i.e. the portfolio is created at the beginning of a quarter and then held for 3 months with no alterations. At the end of the 3 months the value of the portfolio should be as large as possible.

As a basis for the trading strategy a 3-month forecast of the stocks is used. S_t represents the price of one of the 28 stocks. To model the time series S_t we use a NLARX process of the form (19.3); the system function f is approximated with a network function $\nu_H(S_t, A_t, X_t; \vartheta)$. Here A_t is a vector made up of constant non-

linear transformations of $S_t, \ldots, S_{t-p+1}$ that were taken from the technical market analysis, for example, a moving average, momentum or Bollinger-intervals, Müller and Nietzer ([1993]), Welcker ([1994]). The random vector X_t represents the chosen market data such as index prices, exchange rates, international interest rates, etc. As is expected with a forecast horizon of 60 units of time into the future, the actual forecasts of the stock prices in 60 days,

$$\hat{S}_{t+60|t} = v_H(S_t, A_t, X_t; \hat{\vartheta}_n),$$

are not very reliable. For making the decision as to whether a stock should be included in the portfolio or not, the general trend of the price developments is most important instead of the actual price of the stock at the end of the holding period. To realize this aspect in formulating the portfolio, it should be considered whether based on the network based forecast, $\hat{S}_{t+60|t}$, the price is expected to increase considerably (more than 5 %), decrease considerably (more than 5 %) or whether it is essentially expected to stay at the same level. The network based portfolio is composed of those stocks (with relative proportions that are taken from the stock's corresponding weight in the CBS Index) for which $(\hat{S}_{t+60|t} - S_t)/S_t > 0.05$. Here the *same* network function $v_H(S_t, A_t, X_t; \vartheta)$ is used for all 28 stocks taken into consideration whose price dependent arguments S_t actually take on the stock specific values.

In choosing a suitable network and in estimating the network weight vector ϑ the data from 1993 to 1995 is used. In choosing the network structure a statistical model selection technique and the experience of experts were used. The resulting network is a multiple layered perceptron with one hidden layer made up of $H = 3$ neurons. The input vector (S_t, A_t, X_t) has the dimension 25, so that a parameter vector $\vartheta \in \mathbb{R}^{82}$ needed to be estimated.

To check the quality of the network based trading strategy, it is applied to the data from 1996. At the beginning of every quarter a portfolio made up of 28 stocks is created based on the network based forecast. At the end of the quarter the percentage increase in value is considered. As a comparison the increase in value of a portfolio exactly replicating the CBS Index is considered. Since in the years considered the market was of the most part in an increasing phase, it is known from experience that it is hard to beat an index. As Table 19.1 shows, the network portfolio achieved a higher percentage increase in value in every quarter than the index portfolio, that is in the quarters, such as the first and fourth, where the index had substantially increased, as well as in the quarters, such as the second, where the index had

Table 19.1 Quarterly returns of a network portfolio and the index portfolio in 1996

| | Quarterly returns | | | |
	I	II	III	IV
Network portfolio	0.147	0.024	0.062	0.130
Index portfolio	0.109	−0.004	0.058	0.115

minimally decreased. Nevertheless the results need to be interpreted with a bit of caution. Even in the training phase (1993–1995) the CBS Index tended to increase, so that the network was able to specialize in a trend forecast in a generally increasing market. Presumably one would need to use a different network as a basis for the trading strategy, when the market fluctuates within a long-term lateral motion or when the index dramatically decreases.

19.5   Quantifying Risk with Neural Networks

The previous chapters introduced the most popular measurements of risk, volatility and Value-at-Risk. Both are most often defined as conditional standard deviations or as conditional quantiles respectively, based on a given historical information set. As with other non-parametric methods the neural network can also be used to estimate these measurements of risk. The advantage of the neural network based volatility and VaR estimators lies in the fact that the information used for estimating the risk can be represented by a large dimensional data vector without hurting the practicality of the method. It is possible, for example, to estimate the conditional 5% quantile of the return process of a stock from the DAX given the individual returns of all of the DAX stocks and additional macroeconomic data such as interest rates, exchange rates, etc. In the following section we briefly outline the necessary procedure.

As in (15.1) we assume a model of the form

$$Z_{t+1} = f(Z_t, \ldots, Z_{t-p+1}, X_t) + s(Z_t, \ldots, Z_{t-p+1}, X_t)\,\xi_{t+1} \tag{19.4}$$

to estimate the volatility, where ξ_t are independent, identically distributed random variables with $\mathsf{E}(\xi_t) = 0$, $\mathsf{E}(\xi_t^2) = 1$. $X_t \in \mathbb{R}^d$ represents, as in the previous section, the exogenous information available at date t which we will use in estimating the risk of the time series Z_t. The time series given by (19.4) is a non-linear AR(p) ARCH(p) process with exogenous components.

To simplify we use $Z_t(p) = (Z_t, \ldots, Z_{t-p+1})^\top \in \mathbb{R}^p$. It then holds for $z \in \mathbb{R}^p$, $x \in \mathbb{R}^d$ that

$$\mathsf{E}[Z_{t+1}|Z_t(p) = z,\ X_t = x] = f(z, x)$$
$$\mathsf{Var}[Z_{t+1}|Z_t(p) = z,\ X_t = x] = s^2(z, x)$$
$$= \mathsf{E}[Z_{t+1}^2|Z_t(p) = z,\ X_t = x] - f^2(z, x).$$

The conditional expectation function $f(z, x)$ is approximated as in the previous section by a neural network function $\nu_H(z, x; \vartheta)$ of the form (19.1). With the non-linear least squares estimator $\hat{\vartheta}_n$ we obtain for ϑ an estimator for f :

$$\hat{f}_H(z, x) = \nu_H(z, x; \hat{\vartheta}_n).$$

Analogously we could estimate the conditional mean

$$\mathsf{E}[Z_{t+1}^2 | Z_t(p) = z, \ X_t = x] = g(z, x)$$

by approximating the function with a neural network with output function $v_G(z, x; \delta)$ and estimate its parameter δ with a least squares estimator $\hat{\delta}$ within a sufficiently large compact subset $\Delta_G \subset \mathbb{R}^{(p+d+1)G+G+1}$, such as Θ_H, chosen from a fundamental range:

$$\hat{\delta}_n = \underset{\delta \in \Delta_G}{\operatorname{argmin}} \frac{1}{n - p + 1} \sum_{t=p}^{n} \left\{ Z_{t+1}^2 - v_G(Z_t(p), X_t; \delta) \right\}^2,$$

$$\hat{g}_G(z, x) = v_G(z, x; \hat{\delta}_n).$$

As an estimator for the conditional volatility we immediately obtain:

$$\hat{s}_{H,G}^2(z, x) = \hat{g}_G(z, x) - \hat{f}_H^2(z, x).$$

This estimator is in general guaranteed to be positive only for $G = H$. In order to avoid this restriction one can follow the procedure used by Fan and Yao (1998), who have studied a similar problem for the kernel estimator of the conditional variance in a heteroscedastic regression model. Using this application the residuals

$$\varepsilon_{t+1} = Z_{t+1} - f(Z_t(p), X_t) = s(Z_t(p), X_t) \xi_{t+1}$$

are approximated by the sample residuals

$$\hat{\varepsilon}_{t+1} = Z_{t+1} - \hat{f}_H(Z_t(p), X_t), \ t = p, \dots, n,$$

Since the ξ_{t+1} has mean 0 and variance 1,

$$\mathsf{E}[\varepsilon_{t+1}^2 | Z_t(p) = z, \ X_t = x] = s^2(z, x).$$

We could approximate this function directly with a neural network with G neurons and the output function $v_G(z, x; \delta)$, whose parameter are estimated by

$$\hat{\delta}_n = \underset{\delta \in \Delta_G}{\operatorname{argmin}} \frac{1}{n - p + 1} \sum_{t=p}^{n} \left\{ \hat{\varepsilon}_{t+1}^2 - v_G(Z_t(p), X_t; \delta) \right\}^2.$$

The resulting estimators for the conditional volatility, which through the $\hat{\varepsilon}_t$ is also dependent on H, is then

$$\hat{s}_{H,G}(z, x) = v_G(z, x; \hat{\delta}_n).$$
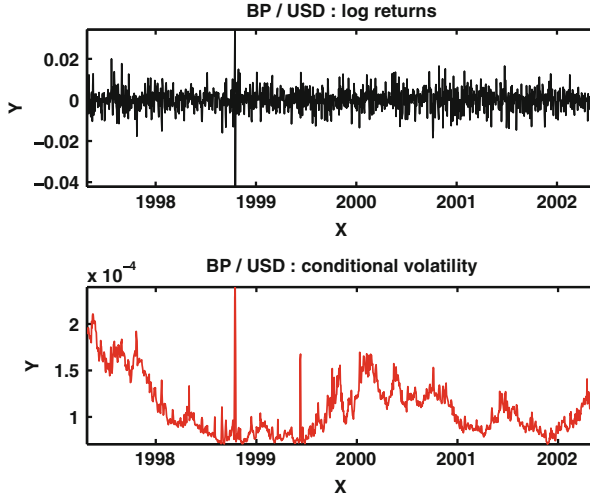
Fig. 19.10 Log-returns of exchange rate BP/USD and the estimated conditional variances by RBF neutral network   Q SFEnnarch

Figure 19.10 shows the conditional volatilities estimated from the log returns of the exchange rate time series BP/USD together with some financial indicators using the procedure described above (three periods are considered as time dependency and RDF networks are used).

It is for arbitrary G, H automatically non-negative. Since the number of neurons essentially determines the smoothness of the network function, it can make sense when approximating f and s^2 to choose different networks with $H \neq G$ neurons when it is believed that the smoothness of both functions is quite different from each other.

When the distribution of the innovations ξ_t is additionally specified in the model (19.4), we immediately obtain together with the estimators of f and s^2 an estimator of the conditional Value-at-Risk. If the distribution of ξ_t is, for example, N(0, 1), then the conditional distribution of Z_{t+1} given the information $Z_t(p)$ and X_t at date t is also a normal distribution with mean $f(Z_t(p), X_t)$ and variance $s^2(Z_t(p), X_t)$. If q_α° is the α quantile of the standard normal distribution, then the VaR process $\{Z_t\}$, i.e. the conditional α quantile of Z_{t+1} given $Z_t(p), X_t$ is:

$$VaR_{t+1} = f(Z_t(p), X_t) + s(Z_t(p), X_t)q_\alpha^\circ.$$

An estimator for this conditional Value-at-Risk based on a neural network can be obtained by replacing f and s with the appropriate estimator:

$$\widehat{VaR}_{t+1} = \hat{f}_H(Z_t(p), X_t) + \hat{s}^2_{H,G}(Z_t(p), X_t)q_\alpha^\circ. \tag{19.5}$$

In doing this we can replace the standard normal distribution with another distribution, for example, with a standardized t-distribution with mean 0 and variance 1. q_α° is then the corresponding α quantile of the innovation distribution, i.e. the distribution of ξ_t.

The estimator (19.5) for the Value-at-Risk assumes that Z_t is a non-linear ARX-ARCHX process of the form (19.4). Above all, however, it has the disadvantage of depending on the critical assumption of a specific distribution of ξ_t. Above all the above-mentioned procedure, in assuming a stochastic volatility model from the standard normal distribution, has recently been criticized in financial statistics due to certain empirical findings. The thickness of the tails of a distribution of a financial time series appears at times to be so pronounced that in order to adequately model it, even the distribution of the innovations must be assumed to be leptokurtic. Due to the simplicity of the representation a t-distribution with only a few degrees of freedom is often considered. In order to avoid the arbitrariness in the choice of the distribution of the innovations, it is possible to estimate the conditional quantile directly without relying on a model of the form (19.4). This application goes back to the regression quantile from Koenker and Bassett and has been applied by Abberger (1997) to time series in connection with kernel estimation.

We assume that Z_t is a stationary time series. As in Chap. 18 P_{t+1} represents the forecast distribution, i.e. the conditional distribution of Z_{t+1} given $Z_t(p)$, X_t. With F_{t+1} we depict the corresponding conditional distribution function

$$F_{t+1}(y|z, x) = P(Z_{t+1} \leq y | Z_t(p) = z, X_t = x)$$

for $y \in \mathbb{R}$, $z \in \mathbb{R}^p$, $x \in \mathbb{R}^d$. $q_\alpha(z, x)$ is the conditional α quantile, i.e. the solution to the equation $F_{t+1}(q_\alpha(z, x)|z, x) = \alpha$. The conditional quantile function $q_\alpha(z, x)$ solves the minimization problem

$$\mathsf{E}\{\alpha(Z_{t+1} - q)^+ + (1 - \alpha)(Z_{t+1} - q)^- | Z_t(p) = z, \ X_t = x\} = \min_{q \in \mathbb{R}}! \qquad (19.6)$$

where $y^+ = y \cdot \mathbf{1}(y \geq 0)$ and $y^- = |y| \cdot \mathbf{1}(y \leq 0)$ represent the positive and negative parts of $y \in \mathbb{R}$. In order to estimate the quantile function directly with a neural network with H neurons, we approximate $q_\alpha(z, x)$ with a network function $v_H(z, x; \gamma)$ of the form (19.1), whose weight parameter γ lies in a fundamental range $\Gamma_H \subset \mathbb{R}^{(p+d+1)H+H+1}$. γ is estimated, however, not with the least squares method, but with the minimization of the corresponding sample values from (19.6):

$$\hat{\gamma}_n = \operatorname*{argmin}_{\gamma \in \Gamma_H} \frac{1}{n - p + 1} \sum_{t=p}^{n} [\alpha\{Z_{t+1} - v_H(Z_t(q), X_t)\}^+$$

$$+ (1 - \alpha)\{Z_{t+1} - v_H(Z_t(q), X_t)\}^-].$$

As an estimator for the quantile function we obtain

$$\hat{q}_{H\alpha}(z, x) = v_H(z, x; \hat{\gamma}_n)$$

and with this the estimator for the conditional Value-at-Risk given $Z_t, \ldots, Z_{t-p+1}$, X_t

$$\widehat{VaR}_{t+1} = \hat{q}_{H\alpha}(Z_t, \ldots, Z_{t-p+1}, X_t).$$

White has shown that under suitable assumptions the function estimators $\hat{q}_{H\alpha}(z, x)$ converge in probability to $q(z, x)$ when the sample observations $n \to \infty$ and when at the same time the number of neurons $H \to \infty$ at a suitable rate.

19.6   Recommended Literature

One for mathematicians, statisticians and economists as well, an accessible introduction to the area of the neural network is, e.g. Refenes (1995a). Haykin (1999) offers a comprehensive and effective overview about different forms and applications of neural network. Anders (1997) introduces neural networks from econometrical and statistical view and discusses applications from the finance mathematical areas such as option pricing and insolvency prediction. Ripley (1996) discusses the application of neural network to classification problems in detail and puts them in the context of the classic discriminant analysis. Numerous practical applications of neural network in the finance area are introduced in Rehkugler and Zimmermann (1994), Refenes (1995b), Bol, Nakhaeizadeh and Vollmer (1996) and Franke (2000). The application described in the previous section, calculating the Value-at-Risk by adaptation of a non-linear ARCHX process based on DAX stocks is described in Franke and Diagne (2006).

Chapter 20
Volatility Risk of Option Portfolios

In this chapter we analyse the principal factors in the dynamic structure of implied volatility *at the money* (ATM). The data used are daily *Volatility*-DAX (VDAX) values. By using a principal component analysis we consider a method of modelling the risk of option portfolios on the basis of "Maximum Loss".

There is a close connection between the value of an option and the volatility process of the financial underlying. Assuming that the price process follows a geometric Brownian motion we have derived the Black–Scholes formula (BS) for pricing European options in Chap. 6. With this formula the option price is, at any given time point, a function of the volatility parameters when the following values are given: τ (time to maturity in years), K (strike price), r (risk free, long-run interest rate) and S (the spot price of the financial underlying).

Alternatively, one can describe the observed market price of an option at a specific time point with the help of the BS formula using the so-called "implied" volatility (see Chap. 6). In doing this, one typically finds a U-shaped form for the resulting surface of the volatility over different times to maturity and strike prices. This phenomenon is also referred to as the "Volatility Smile". Figure 20.1 illustrates the typical form of a volatility surface using DAX options. Shown is the implied volatility as a function of the moneyness and the remaining time to maturity τ. Here the term *moneyness* $\frac{S}{K}$ refers to the ratio of the actual price S of the financial underlying and the strike price K of the respective option. It should be noted that options are only traded on the market on a discrete price basis and a discrete time to maturity. In determining the volatility surface, as in Chap. 15, a smoothing technique needs to be applied.

By observing the volatility surface over time, distinct changes in the location and structure become obvious. Identifying the temporal dynamics is of central importance for a number of financially oriented applications. This is of particular importance for the risk management of option portfolios. To determine the volatility's"dynamics, an application of principal component of analysis is quite suitable.
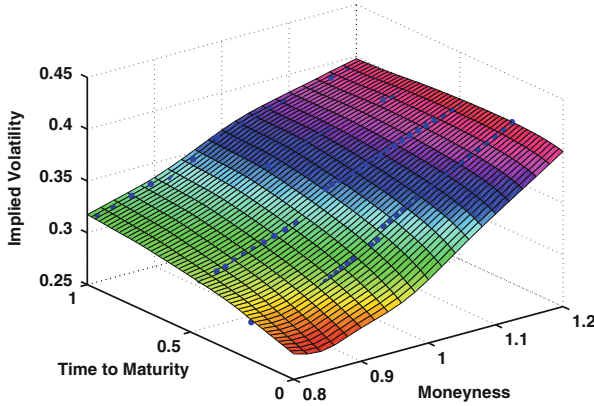
Fig. 20.1 Implied volatility surface of the DAX option on 18 July 1998   SFEVolSurfPlot

The total temporal structure can be sufficiently represented by a small number of principal components so that the dimensions of the factor space for the purpose of risk analysis can be significantly reduced.

20.1   Description of the Data

DAX options belong to the most frequently traded derivatives of the German/Swiss derivative market "EUREX". On every trading day one can find a significant number of liquid time series with varying strike prices and maturities (K, τ) on the market, which, in principle, can be used to calculate implied volatilities. In view of the often limited data processing capacities, an updated calculation of numerous volatilities and partial derivatives of an extensive option portfolio is still not feasible. Even with the appropriate available information the isolated consideration of each implied volatility as a separate source of risk is problematic, since it results in an unstructured or "uneven" volatility surface. If one were to use generated volatilities in calibrating option prices for respective risk models, it could lead to serious specification errors and significantly deteriorate the results of the corresponding trading and hedging strategies. As a result of principal component analysis a "smooth" volatility surface, in contrast to the one outlined above, can be generated with a manageable amount of information. This allows for a better calibration of the model and a more precise estimate of portfolio sensitivities.

For our study of the dynamics of implied volatility we use the volatility index (VDAX) made available by the German Stock Exchange (Deutsche Börse AG) and the closing prices of the corresponding VDAX sub-indices. These indices reflect the implied volatility of the DAX options "ATM" for times to maturity from one to 24

months. The corresponding values are determined by applying the Black–Scholes formula (6.23) using prices observed on the market:

$$C(S, \tau) = e^{(b-r)\tau} S \Phi(y + \sigma\sqrt{\tau}) - e^{-r\tau} K \Phi(y),$$

where Φ is the distribution function of the standard normal distribution and

$$y = \frac{\log(\frac{S}{K}) + (b - \frac{1}{2}\sigma^2)\tau}{\sigma\sqrt{\tau}},$$

The only parameter from the BS formula that cannot be immediately observed on the market is the actual volatility σ of the price process. In principle the volatility of the process can be estimated from historical financial market data, see Sect. 6.4.5, however, it is commonly known that, the assumption of the BS model, that the financial underlying has a geometric Brownian motion, is in reality only approximately fulfilled. Alternatively the BS formula is also used in order to calculate the σ value as the implied volatility for a given market price of a specific option. This does not mean that the market participant should accept the assumption of the Black–Scholes method. On the contrary they use the BS formula as a convenient possibility to quote and price options with these parameters.

Given the observed implied volatilities from varying times to maturity τ at a specific time point and from a strike price K, the expectations of the market participants with respect to the future actual volatility of the underlying financial instrument can be estimated. In doing so one must remember that the implied volatility of the BS model does not directly apply to the actual variance of the price's process. Although the implied BS volatility reflects a market expectation, the theoretical relationship between it and the actual volatility can only be determined using specific assumptions, see Schönbucher (1999), Härdle and Hafner (2000).

Implied volatility for ATM-DAX options are calculated for various lengths of maturity by the German Stock Exchange AG. A detailed description of how the VDAX and its sub-indices are calculated can be found in Redelberger (1994). Since 18 March 1996 maturities of 1, 2, 3, 6, 9, 12, 18 and 24 months have been considered in the calculation. On this date the trading of so-called "Long Term Options", i.e. trading of options with maturities of over 12 months, was added to the EUREX. Using closing prices the German Stock Exchange AG calculates a total of eight VDAX sub-indices for the maturities mentioned above for every trading day. These sub-indices reflect the volatility of the respective DAX option "ATM". The time to maturity structure for DAX options that are "ATM" can be determined for every trading day using the VDAX indices. Figure 20.2 illustrates a typical development of the structure, which shows strong changes in the positioning and form of the structure over time.

The analysis carried out here is not only restricted to specific maturities of liquid options, which are represented by the first four VDAX sub-indices. On the contrary, we include all eight sub-indices in the analysis for the following reasons:

Term Structure

First of all a brisk trade in even the "most distant" option contracts (i.e. the contracts with a remaining time of more than 1 year) takes place on numerous trading days, so that excluding the pertaining sub-indices from the analysis would result in a loss of information. VDAX sub-indices for long maturities have been calculated by the German Stock Exchange since 18 March 1996. After 19 December 1997 the quality of the data available to us declined considerably. In addition to the daily often unchanged prices, the entries corresponding to the removed sub-indices were usually missing. Given this we have restricted our analysis to the period from 18 March 1996 to 19 December 1997.

Including relatively non-liquid DAX options with long maturities appears to make sense for another reason: for our analysis we require constant option maturities, since the daily shortening of the time to maturity can lead to enormous biases in the analysis results with data that has not been corrected. This especially holds for options with a very short time to maturity. Thus we find it utterly necessary to use interpolated volatilities with corresponding constant time to maturities of the underlying option. Referring back to the calculation of the VDAX used by the German Stock Exchange AG we use the following linear interpolation:

For a fixed time to maturity of $\tau_1^* = 30, \tau_2^* = 60, \tau_3^* = 90, \tau_4^* = 180, \tau_5^* = 270, \tau_6^* = 360, \tau_7^* = 540, \tau_8^* = 720$ calendar days we calculate daily volatility indices $\hat{\sigma}_{I,t}(\tau_j^*), j = 1, \ldots, 8$, using the VDAX sub-indices with the next shorter respectively longer maturity $\hat{\sigma}_{I,t}(\tau_j^-)$ and $\hat{\sigma}_{I,t}(\tau_j^+)$ with

$$\hat{\sigma}_{I,t}(\tau_j^*) = \hat{\sigma}_{I,t}(\tau_j^-) \left\{ 1 - \frac{\tau_j^* - \tau_j^-}{\tau_j^+ - \tau_j^-} \right\} + \hat{\sigma}_{I,t}(\tau_j^+) \left\{ \frac{\tau_j^* - \tau_j^-}{\tau_j^+ - \tau_j^-} \right\}. \qquad (20.1)$$

This way, we obtain 8 volatility time series each with constant maturities. Every time series represents a weighted average of two consecutive VDAX sub-indices and is based on $n = 441$ daily observations of the implied DAX volatilities "ATM".

20.2   Principal Component Analysis of the VDAX's Dynamics

We will first check the data with the help of the "Augmented Dickey-Fuller" Tests (ADF-Test—see (11.46)) for stationarity. The null hypothesis of a unit root for the individual VDAX sub-indices $\hat{\sigma}_I(\tau_j^*)$ cannot be rejected at the 90 % significance level. Obviously due to this result the first differences $x_{jt} = \Delta[\hat{\sigma}_{I,t}(\tau_j^*)] = \hat{\sigma}_{I,t+1}(\tau_j^*) - \hat{\sigma}_{I,t}(\tau_j^*), t = 1, \ldots, n-1$, of the implied volatility indices will be used for further analysis. Additional ADF tests support the assumption of stationarity for the first differences. 🝆 SFEAdfKpss

Let $\bar{x}_j$ be the respective sample mean of the first differences x_{jt}. Table 20.1 contains the empirical covariance matrix $\hat{\Omega}$ used as an estimator for the 8×8 matrix Ω of the covariance $\mathsf{Cov}(x_{it}, x_{jt}), i, j = 1, \ldots, 8$. With help of the Jordan decomposition we obtain $\hat{\Omega} = \hat{\Gamma}\hat{\Lambda}\hat{\Gamma}^\top$. The diagonal matrix $\hat{\Lambda}$ contains the eigenvalues $\hat{\lambda}_k, k = 1, \ldots, 8$ of $\hat{\Omega}$, $\hat{\Gamma}$ are the eigenvectors. Time series of the principal components can be obtained with the help of $Y = X_C\hat{\Gamma}$, where X_C represents the 440×8 matrix of the centred first differences $x_{jt}^c = x_{jt} - \bar{x}_j, j = 1, \ldots, 8, t = 1, \ldots, 440$. The 440×8 matrix $Y = (Y_1, \ldots, Y_8), Y_j = (y_{1j}, y_{2j}, \ldots, y_{440,j})^\top$ contains the principal components.

How accurately the first l principal components have already determined the process of the centred first differences can be measured using the proportion of variance φ_l with respect to the total variance of the data. The proportion of explained variance corresponds to the relative proportion of the corresponding eigenvalue, i.e.

Table 20.1 Empirical covariance matrix $\hat{\Omega}$ of the first differences (all values have been multiplied by 10^5). The variances are given with bold values   🝆 SFEVolaCov

Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	Sub 7	Sub 8
20.8	9.06	6.66	6.84	4.29	2.48	2.11	1.38
9.06	**9.86**	6.67	4.44	3.21	1.72	1.11	0.92
6.66	6.67	**6.43**	3.87	2.63	1.49	1.01	0.53
6.84	4.44	3.87	**4.23**	2.66	1.39	1.38	0.68
4.29	3.21	2.63	2.66	**2.62**	1.03	1.02	0.51
2.48	1.72	1.49	1.39	1.03	**2.19**	0.63	0.33
2.11	1.11	1.01	1.38	1.02	0.63	**1.76**	0.43
1.38	0.92	0.53	0.68	0.51	0.33	0.43	**1.52**

Table 20.2 Explained sample variance using principal components in percentage

 SFEVolaPCA

Principal component	Explaining proportion of variance	Cumulative proportion
1	70.05	70.05
2	13.06	83.12
3	5.57	88.69
4	3.11	91.80
5	3.06	94.86
6	2.12	96.97
7	1.93	98.90
8	1.10	100.00

$$\varphi_l = \frac{\sum_{k=1}^{l} \lambda_k}{\sum_{k=1}^{8} \lambda_k} = \frac{\sum_{k=1}^{l} \mathsf{Var}(y_{tk})}{\sum_{k=1}^{8} \mathsf{Var}(y_{tk})}, \tag{20.2}$$

where $\lambda_k, k = 1, \ldots, 8$ are the eigenvalues of the true covariance matrix Ω. An estimator for φ_l is

$$\hat{\varphi}_l = \frac{\sum_{k=1}^{l} \hat{\lambda}_k}{\sum_{k=1}^{8} \hat{\lambda}_k}.$$

In Table 20.2 the individual proportions of the variance $\hat{\lambda}_l / \sum_{k=1}^{8} \hat{\lambda}_k$ as well as the cumulative variance from the l decomposed proportions from the principal components, $\hat{\varphi}_l$, are displayed. It is obvious that the first principal component already describes 70 % of the total variance of the underlying data. With the second principal component an additional 13 % of the total variance within the observed time period can be explained. Together 83 % of the variance of the analysed first differences of our VDAX sub-indices can be explained with the help of the first and second principal components. Obviously the explaining power of the principal components significantly declines from the third principal component onwards.

By displaying the eigenvalues in a graph, a form with a strong curvature at the second principal component is shown. In accordance with the well known "elbow" criterion, using the first two principal components with an explanation power of over 80 % of the total variance is considered to be sufficient in describing the data set. The remaining variance can be interpreted for analytical purposes as the effect of an unsystematic error term. Figure 20.3 contains the factor loading of the first two principal components. Based on the orthogonality of the components the loading factors can be estimated using the least squares regression of the individual equations

$$x_{jt}^c = \sum_{l=1}^{2} b_{jl} y_{lt} + \varepsilon_t, \tag{20.3}$$

Here ε_t is an independent error term.
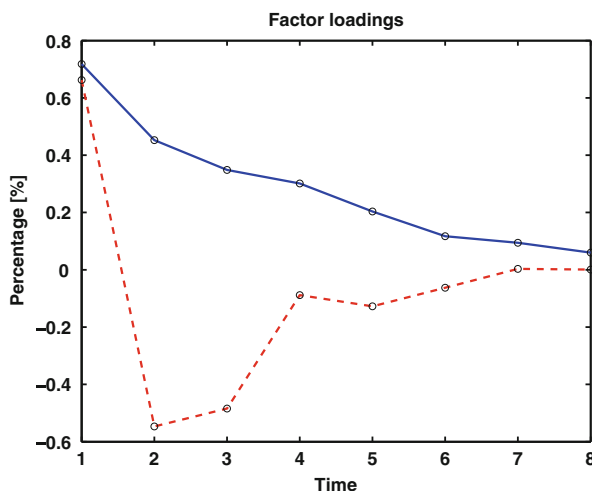
Fig. 20.3 Factor loadings of the first and second principal components   ⬛ SFEPCA

Based on the factor loadings it is clear that a shock to the first factor would affect the implied volatility of all times to maturity considered in a similar way, or would cause a non-parallel shift in the maturities' structure. A shock to the second principal component, on the other hand, causes a tilt of the structure curve: while at short times to maturity it causes a positive change, the longer time to maturities are influenced negatively. The absolute size of the effect of a shock decreases in both factors with the time to maturity.

20.3   Stability Analysis of the VDAX's Dynamics

In order to sensibly apply the principal factors in measuring the risk of portfolios we have to study their stability over time. When the principal components and the factor loadings change significantly over time, a global analysis would not be suitable to illustrate the future variance of the implied volatility nor to judge the risks of the option portfolios with sufficient accuracy.

Our procedure considers two aspects: first whether the random portion in daily data is possibly significantly higher than in weekly data. A possible cause for this is the non-synchronous trading caused by frequent realizations of the quotes in the liquid contracts with a short time to maturity and sparsely available prices in the long running contracts. In order to distinguish the possible influences of this effect, we run our analysis analogously based on weekly data. By sufficient stability in the principal components the use of daily, and respectively, weekly data should lead to similar results.

Table 20.3 Explained portion of the variance (in percentage) in different sub-periods

Principal components	1	2	3	4	5	6	7	8
Weekly data								
18.03.96–19.12.97	73.85	91.59	95.09	97.38	98.80	99.61	99.88	100
Daily data								
18.03.96–19.12.97	70.05	83.12	88.69	91.80	94.86	96.97	98.90	100
Sub-period 1								
18.03.96–05.02.97	83.36	91.84	94.65	96.39	97.76	98.78	99.52	100
Sub-period 2								
05.02.97–19.12.97	68.22	82.21	87.99	91.35	94.64	96.93	98.86	100

For the second aspect we divide our data into two non-overlapping periods of equal length. Each sub-period contains $m = 220$ daily observations of the process of the differences. For each sub-period we run a principal component analysis as described above and compare the respective sizes of the eigenvalues $\hat{\lambda}_k^i, k = 1, 2$, in both sub-periods $i = 1, 2$.

As already mentioned the effect of non-synchronous trading that appears in daily data can be eliminated by using weekly data. From Table 20.3 it emerges that the explanatory power of the first principal component is slightly higher in weekly data. This is not surprising given the expected size of the error terms proportion in daily data. Overall the explanatory proportions of the variance have similar values when using weekly data. This supports the stability of the analysis method used here w.r.t. the bias due to non-synchronous trading in daily data.

From Table 20.3 it emerges that the proportion of the variance explained by the first two principal components declines in the second sub-period. Based on this a stability test is necessary: a two-sided confidence interval for the difference of the eigenvalues from both sub-periods is

$$\log \hat{\lambda}_k^1 - 2q_\alpha \sqrt{\frac{2}{m-1}} \leq \log \hat{\lambda}_k^2 \leq \log \hat{\lambda}_k^1 + 2q_\alpha \sqrt{\frac{2}{m-1}}, \qquad (20.4)$$

where q_α represents the α quantile of a standard normal distribution, see Härdle and Simar (2012). From this it follows that

$$\mid \log \hat{\lambda}_k^1 - \log \hat{\lambda}_k^2 \mid \geq 2q_\alpha \sqrt{\frac{2}{m-1}} \qquad (20.5)$$

is a second test for $H_0 : \lambda_k^1 = \lambda_k^2$. Under the null hypothesis the respective eigenvalues are the same in both periods. The null hypothesis is rejected when the inequality is fulfilled for a corresponding critical value q. This would indicate an instability of the principal components over time.

Critical values for rejecting the null hypothesis are 0.313 (probability of error 10 %), 0.373 (probability of error 5 %) and 0.490 (probability of error 1 %). The differences of the estimated eigenvalues are 0.667 and 1.183. Both differences are significantly larger than zero with an error probability of 1 %. These results prove that the determining factors of the volatility dynamics change over time. By the determination of the risk of option portfolios it therefore appears necessary to use an adaptive method of the principal components. Here the estimation is periodically done over a moving time window and the length of the time window is adaptively set, see Cizek, Härdle and Weron (2011).

20.4 Measure of the Implied Volatility's Risk

The market value P_t of a portfolio consisting of w different options is dependent on changes of the risk free interest rate r_t, the prices S_t of the financial underlying, the time to maturity τ and the individual implied volatilities σ_I. Changes in the portfolio value can be analytically approximated using the following Taylor approximation, where it is assumed that the options are all based on the same underlying.

$$\Delta P_t = \sum_{u=1}^{w} \left\{ \frac{\partial V_{ut}}{\partial \sigma_I} \Delta \sigma_{I,t} + \frac{\partial V_{ut}}{\partial t} \Delta t + \frac{\partial V_{ut}}{\partial r} \Delta r_t \right.$$

$$\left. + \frac{\partial V_{ut}}{\partial S} \Delta S_t + \frac{1}{2} \frac{\partial^2 V_{ut}}{\partial S^2} (\Delta S_t)^2 \right\}$$

$$\Delta P_t = \sum_{u=1}^{w} \left\{ \frac{\partial V_{ut}}{\partial \sigma_I} \Delta \sigma_{I,t} + \Theta_u \Delta t + \rho_u \Delta r_t + \Delta_u \Delta S_t + \frac{1}{2} \Gamma_u (\Delta S_t)^2 \right\}$$

Here V_{ut} describes the price of the u-th option with a time to maturity τ_u at date t, and $\Theta_u, \rho_u, \Delta_u, \Gamma_u$ are the characteristic values described in Sect. 6.4 of the u-th option. In practice option traders often insert "Vega" positions directly. In doing so they create portfolios whose profit and loss profile can be determined by the changes in the implied volatilities of the respective options, see Taleb (1997). Portfolios of this kind are called (Δ, Γ) and Θ neutral. The sensitivity of the option price under consideration to the changes in the volatilities is measured by the variable $\mathcal{V}$ ("Vega"—see (6.33)).

A well-known strategy in utilizing the forecasted changes in the maturity structure of implied volatilities consists of buying and selling so-called "Straddles" with varying maturities. A straddle is constructed by simultaneously buying ("Long Straddle") or selling ("Short Straddle") the same number of ATM Call and Put options with the same time to maturity. If a trader expects a relatively strong increase in the implied volatility in the short maturities and a relatively weaker increase in the longer maturities, then he will buy straddles with a short time to maturity and sell

longer maturity straddles at a suitable ratio. The resulting option portfolio is (Δ, Γ) neutral and over a short time frame Θ neutral, i.e. it is insensitive with respect to losing value over time. The Taylor series given above can thus be reduced to:

$$\Delta P_t \approx \sum_{u=1}^{w} \left\{ \frac{\partial V_{ut}}{\partial \sigma_I} \Delta \sigma_{I,t}^{(t)} \right\} \tag{20.6}$$

The first differences of the implied volatilities can now be given as linear combinations of the principal components. By substituting the volatility indices $\sigma_{I,t}$, which are temporally next to the actual implied volatility $\hat{\sigma}_I(\tau_u^*)$, one obtains the following representation given (20.3):

$$\Delta P_t \approx \sum_{u=1}^{w} \left\{ \frac{\partial V_{ut}}{\partial \sigma_I} \left(\sum_{l=1}^{2} b_{jl} y_{lt} \right) \right\} \tag{20.7}$$

The number of principal components used in the previous expression can be reduced to the first two without any significant loss of information.

The following "Maximum Loss" (ML) concept describes the probability distribution of a short-term change in the portfolio value dependent on changes in the value of the underlying factors. The change in value of a (Δ, Γ) neutral option portfolio is substantially determined by the changes in the implied volatilities of the options contained in the portfolio. To determine the "Maximum Loss" it is necessary to have an adequate exact representation of the future distribution of the changes to the volatility of the options with varying time to maturity.

The "Maximum Loss" is defined as the largest possible loss of a portfolio that can occur over a specific factor space A_t and over a specific holding period τ. The factor space A_t is determined by a closed set with $P(A_t) = \alpha$. Here α is set to 99 or 99.9 %. The ML definition resembles at first sight the "Value-at-Risk" Definition (see Chap. 16). There is, however, an important difference between the two concepts: in calculating the "Value-at-Risk" the distribution of the returns of the given portfolio must be known, whereas the ML is defined directly over the factor space and thus has an additional degree of freedom, see Studer (1995).

In our analysis we have divided the maturity structure of the implied volatilities into two principal components, which explain a considerable portion of the variability of the structure curve. Thus the first two principal components represent the risk factors used in the ML model. The profit and loss profile of each portfolio held is determined by the corresponding changes in the risk factors using a suitable valuation model. In order to obtain this, a valuation of the underlying portfolios must theoretically occur for every point in the factor space. In the practical application the factor space is probed over a sufficiently small grid of discrete data points $y_1^z (z = 1, \ldots, N_1)$, during which the other risk factor is held constant in each case. Due to the orthogonality properties of the principal components, the profit and loss function $PL()$ is additive with $PL(y_1^{z1}, y_2^{z2}) = PL(y_1^{z1}) + PL(y_2^{z2})$.

Under the assumption of multivariate, normally distributed principal components confidence intervals can be constructed for the "Maximum Loss" over the total density

$$\varphi_2(y) = \frac{1}{(2\pi)\sqrt{\det \Lambda_2}} \exp\left(-\frac{1}{2}y^\top \Lambda_2^{-1} y\right), \tag{20.8}$$

with $y = (y_1, y_2)^\top$. Here the matrix Λ_2 represents the 2×2 diagonal matrix of the eigenvalues λ_k, $k = 1, 2$. The random variable $y^\top \Lambda_2^{-1} y = X_1^2 + X_2^2$ has a Chi-square distribution. The confidence interval for an existing portfolio is then $A_t = \{y; y^\top \Lambda_2^{-1} y \leq c_\alpha\}$, c_α, where c_α is the α quantile of a random variable with a Chi-square distribution and 2 degrees of freedom.

20.5 Recommended Literature

The presentation of this chapter closely follows the work of Fengler, Härdle and Schmidt (2002). The principal components analysis is applied to the changes of implied volatilities for fixed ranges of days to maturity by Skiadopoulos, Hodges and Clewlow (1998) who find two principal components can already sufficiently explain the dynamics of smiles. The conditions to ensure the absence of arbitrage in the volatility models are derived by Schönbucher (1999). Furthermore, Härdle and Hafner (2000) show that the prices of out-of-the-money options strongly depend on volatility features such as asymmetry.

Cizek et al. (2011) develop an adaptive method of estimation which does not use any information about the time homogeneity of the observed process. It can be used to estimate the principal components. For the effect of the implied volatilities changes on the dynamic hedging of exotic and complex options we refer to Taleb (1997).

20.6 Exercises

Exercise 20.1 *On 1 July 2005, the closing price of DAX was* 4617.07. *One observes the following call options with strike $K = 4,600$ and maturities in year $\tau_1 = 0.2109$, $\tau_2 = 0.4602$, $\tau_3 = 0.7095$. The prices of these options are respectively $C_1 = 119.4$, $C_2 = 194.3$, $C_3 = 256.9$. Assume that C_2 is not observed. In order to approximate this price one may use linear interpolation of options C_1 and C_3. The interpolation can be performed in prices C and in implied volatilities $\widehat{\sigma}$.*

Compare the two interpolation methods and check which gives the closest approximation to the true price. Interpolate also the variance and compare the results. (The interest rate is 2.1 %.)

Exercise 20.2 *On 1 July 2005, the closing price of DAX was 4,617.07. One observes the following call options with strikes K_1 = 4,000, K_2 = 4,200, K_3 = 4,500 and maturity in year τ = 0.2109. The prices of these options are respectively C_1 = 640.6, C_2 = 448.7, C_3 = 188.5. Assume that C_2 is not observed. In order to approximate this price one may use linear interpolation of options C_1 and C_3. The interpolation can be performed in prices C or in implied volatilities $\widehat{\sigma}$.*

Compare these interpolation methods and check which gives the closest approximation to the true price. Use interest rate $r = 2.1\%$.

Exercise 20.3 (Stickiness) *Let the current underlying price be $S_0 = 100$, maturity $\tau = 0.25$ years and interest rate $r = 2\%$. Assume that implied volatility is given as function of strike price $f(K) = 0.000167K^2 - 0.03645K + 2.08$.*

Plot call option prices as a function of strikes for $K \in (85, 115)$. Assume that the underlying price moves to $S_1 = 105$. The implied volatility function may be fixed to the strike prices (sticky strike) or moneyness K/S_1 (sticky moneyness). Plot call option prices with two different stickiness assumptions. Compare the relative difference of both approaches.

Exercise 20.4 *A risk reversal strategy is defined as a long position in an out-of-the-money put and a short position in an out-of-the-money call (or vice versa). Consider the risk reversal strategy of long put with strike 85 and short call with strike 115 for maturity $\tau = 0.25$ years. Let the current underlying price be $S_0 = 100$, and interest rate $r = 2\%$. Compare the prices of the risk reversal for the following implied volatility curves given as a function of strike price:*

 (i) $f_1(K) = 0.000167K^2 - 0.03645K + 2.080$.
 (ii) $f_2(K) = 0.000167K^2 - 0.03645K + 2.090$.
 (iii) $f_3(K) = 0.000167K^2 - 0.03517K + 1.952$.

Exercise 20.5 *A calendar spread strategy is defined as a position in two options with same strike but different maturity. Consider a calendar spread for at-the-money short calls with maturity $\tau_1 = 0.25$ and at-the-money long call with maturity $\tau_2 = 1$ year. Let the current underlying price be $S_0 = 100$, and interest rate $r = 2\%$. Compare the prices of the calendar spread for the following implied volatility curves given as a function of maturity.*

 (i) $f_1(\tau) = 0.15\tau + 0.05$.
 (ii) $f_2(\tau) = 0.15\tau + 0.06$.
 (iii) $f_3(\tau) = 0.1\tau + 0.075$.

Exercise 20.6 *In order to price the options for the strikes outside the observed range an extrapolation has to be used. Consider the IV data given in Table 20.4.*

Let the observation 1 be a validation observation. Using the constant extrapolation on observation 2, the linear extrapolation on observations 2, 3 and the quadratic extrapolation on observations 2, 3, 4 applied to implied volatilities obtain an estimate of the call price for observation 1. Compare the result with the actual

Table 20.4 Observed strikes, implied volatilities and call prices

Observation	K	σ	C_t
1	4,000	0.1840	640.6
2	4,100	0.1714	543.8
3	4,200	0.1595	448.7
4	4,500	0.1275	188.5

price. How would the result differ if instead of extrapolating in IVs the true price would be used? For your calculation use spot $S_0 = 4{,}617.07$, interest rate $r = 2.1\%$ and maturity $\tau = 0.2109$.

Chapter 21
Non-parametric Estimators for the Probability of Default

The estimation of the probability of default based on information on the individual customer or the company is an important part of credit screening, i.e., judging the credit standing. It is essential for the establishment of a rating or for measuring credit risk to estimate the probability that a company will end in financial difficulties within a given period, for example, 1 year. Also, here non-parametric applications prove to be flexible tools in estimating the desired default probability without arbitrary assumptions. In this chapter we will give a brief overview of the various approaches for non- and semiparametric estimates of conditional probabilities.

21.1 Logistic Regression

In order to judge the credit standing of a customer a series of data are in general available. For consumer credit there are, for example, in Müller (2000): *level of credit, age of the customer, duration of credit* as well as information on whether the customer is *unemployed* or not and whether there were problems in the past with *repaying loans*. For the insolvency prognoses for a small company relevant information would, for example, be in Anders (1997): *age of the business, sales development* from the recent past, *educational degree of the entrepreneur, type of business* and information on *liability*.

Some influential values are quantitative, such as credit volume and sales development. Others are qualitative in nature and must be transformed into numbers for estimating the default probability. For dichotomic characteristics (unemployed, employed, limited liability, unlimited liability) indicator variables are set with values of 0 and 1. For characteristics with $d > 2$ possibilities and for categorical values $d - 1$ dummy variables are introduced, which also take on the value of 0 or 1. Coding the characteristics numerically the *type of business* and three clarifying variables *trade, processed business, other* are considered for which two Dummy variables,

Z_1, Z_2, are used where $Z_1 = 1$ ($Z_2 = 1$) if and only if the type of business is *trade (processed business)*. When $Z_1 = Z_2 = 0$, the firm considered belongs to one of the *other* types of business, for example, *services*. The case $Z_1 = Z_2 = 1$ cannot occur.

If the values of the qualitative characteristics are hierarchically ordered, then it is possible to represent them with an integer valued random variable. The personal impression of the processor in the bank of the economic situation of a company: *very good, good, satisfactory, poor, very poor* can, for example, be transformed into a number scale: *1, 2, 3, 4, 5.* Here one must be certain that every monotone transformation, i.e., where the order remains consistent, produces a different numerical code that can be used with the same justification. Instead of *1, 2, 3, 4, 5* one could also use *0, 1, 3, 6, 10* for instance. Using parametric applications such as the logistic regression one should specify the arbitrary setting of a numerical scale for the hierarchical characteristics. Through a monotone transformation of the scale better estimates can eventually be obtained for the default probabilities. Adequately flexible non-parametric and semi-parametric applications, in contrast, automatically choose a suitable scale.

In order to estimate the default probability of a credit, given the information available at the time the decision is made, we assume a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ is independent, identically distributed. $X_j \in \mathbb{R}^d$ stands for the information available at the time the credit is issued to the jth customer, where qualitative characteristics are already transformed into numerical values as described above. $Y_j \in \{0, 1\}$ is the indicator variable of the credit: it has a value of 0 when the loan can be paid back without any problems and 1 when the credit partially or completely defaulted. The default probability that is to be estimated is the conditional probability that $Y_j = 1$, given $X_j = x$:

$$\pi(x) = \mathrm{P}(Y_j = 1 | X_j = x), \ x \in \mathcal{X},$$

where $\mathcal{X} \subset \mathbb{R}^d$ represents the value space of X_j.

Since $\pi(x)$ only takes on the values between 0 and 1 given that it is a probability, linear regression models cannot be used for the function estimator. The class of generalized linear models (GLM) can, however, be used to estimate the probabilities. Here it is assumed that

$$\pi(x) = G\left(\beta_0 + \sum_{i=1}^{d} x_i \beta_i\right) = G(\beta_0 + \beta^\top x).$$

$G : \mathbb{R} \to [0, 1]$ is a known function that only takes on a value between 0 and 1, the real valued parameters $\beta_0, \ldots, \beta_d$ are unknown and need to be estimated. For the special case that G is chosen to be a logistic function ψ:

$$G(t) = \psi(t) = \frac{1}{1 + e^{-t}},$$

we obtain the model of the *logistic regression*: Given $X_1, \ldots, X_n$, the credit indicators $Y_1, \ldots, Y_n$ are independent Bernoulli random variables with parameters $\psi(\beta_0 + \beta^\top X_1), \ldots, \psi(\beta_0 + \beta^\top X_n)$. The conditional likelihood function is thus

$$L(\beta_0, \ldots, \beta_d) = \prod_{j=1}^{n} \left[\psi(\beta_0 + \beta^\top X_j)^{Y_j} \{1 - \psi(\beta_0 + \beta^\top X_j)\}^{(1-Y_j)} \right].$$

Since Y_j only takes on a value between 0 and 1, the corresponding conditional log-likelihood function is

$$\log L(\beta_0, \ldots, \beta_d) = \sum_{j=1}^{n} [Y_j \log \psi(\beta_0 + \beta^\top X_j) + (1 - Y_j) \log\{1 - \psi(\beta_0 + \beta^\top X_j)\}].$$

Through maximizing L or $\log L$ one obtains the maximum likelihood estimator $\hat{\beta}_0, \ldots, \hat{\beta}_d$ of $\beta_0, \ldots, \beta_d$ and thus the maximum likelihood estimator for the default probability in the logistic regression model:

$$\hat{\pi}(x) = \psi(\hat{\beta}_0 + \hat{\beta}^\top x).$$

21.2   Semi-parametric Model for Credit Rating

The logistic regression model for the estimate of the conditional probability suffers under the same restrictions as the linear regression model when estimating the general functions. In order to avoid the dependence on the special parametric form of the model and to gain more flexibility in the function estimation it is recommended that $\pi(x)$ is estimated non-parametrically, for example, with the LP-method given in (15.4) and (15.7). In doing this, however, it is not guaranteed that the function estimator will lie between 0 and 1. In order to enforce this possible, as was carried out in the previous section, we transform the value space of the estimated function to the interval [0,1] using a given function G:

$$\pi(x) = G\{m(x)\}$$

where $m(x)$ is an arbitrary real valued function that can be estimated nonparametrically. For the estimate of the default probabilities the local smoothing methods are less suitable for two reasons. First of all x is often high dimensional in the application, for example, after adding the necessary dummy variables in the example considered by Müller and Rönz (2000) it has a dimension of 61, that even by larger local neighbourhoods of x of the random sample, over which the estimation occurs, there are either too few observations or too large to produce a reliable estimate of $m(x)$. This problem can be solved by restricting ourselves to additive models

$$\pi(x) = G\left\{ \sum_{i=1}^{d} m_i(x_i) \right\},$$

where $m_1(u), \ldots, m_d(u)$ are arbitrary functions of the one-dimensional variable u. It is however more critical that many of the coordinates of x take on a value of 0 or 1 in the credit rating, since they represent, from the very beginning, dichotomic characteristics or have been added as dummy variables for the unordered qualitative characteristics. Local smoothing functions would be suitable based on their underlying philosophy, but mainly for estimating functions with continuous arguments.

A combination of non-parametric and parametric applications offers the possibility of using the flexibility of the non-parametric method by credit rating, Müller and Rönz (2000). In doing so the influential variables are not combined in a random vector X_j, but are separated into two random vectors $X_j \in \mathbb{R}^p, Z_j \in \mathbb{R}^q$. The coordinates of Z_j represent several chosen exclusive quantitative characteristics and eventual hierarchical qualitative characteristics with sufficiently accurate subdivided value spaces. All remaining characteristics, especially the dichotomic and the dummy variables of unordered qualitative characteristics, are combined in X_j. In order to estimate the default probability we consider a generalized partial linear model (GPLM = generalized partial linear model):

$$P(Y_j = 1 | X_j = x, Z_j = z) = \pi(x, z) = G\{\beta^\top x + m(z)\}.$$

G is again a known function with values between 0 and 1, for example, the logistic function ψ. $\beta_1, \ldots, \beta_p$ are unknown parameters, m is an arbitrary, unknown function that can contain an additive constant and thus can make an additional parameter β_0 superfluous. In an extensive case study Müller (2000) has shown that the additional flexibility from the non-parametric part $m(z)$ of the model results in a better estimate of the default probability than a pure parametric logistic regression.

There are various algorithms for estimating β and $m(z)$, for example the profile likelihood method from Severini and Wong (1992) and Severini and Staniswallis (1994) or the back-fitting method from Hastie and Tibshirani (1990). Essentially they use the fact that for the known function $m(z)$ of the parameter vector β can be estimated through maximization of the log-likelihood function analog to the logistic regression

$$\log L(\beta) = \sum_{j=1}^{n} (Y_j \log G\{\beta^\top X_j + m(Z_j)\} + (1 - Y_j) \log[1 - G\{\beta^\top X_j + m(Z_j)\}])$$

and for known β the function $m(z)$ can be estimated with local smoothing analog to the LP-Method (15.4), (15.7). Both of these optimization problems are combined in an iterative numerical algorithm.

Example 21.1 As an example we consider the rating of consumer credit already referred to above that Müller (2000) had carried out with a GPLM method. The data represent a part of the extensive random sample, which is described in detail by Fahrmeir and Tutz (1994). We use a total of $n = 564$ observations, in which 24.3 % of the cases have a problem with repaying the credit ($Y_j = 1$). From the five influential variables considered, two are dichotomic; they indicate whether the customer is unemployed or not (X_{j1}) and whether the customer has had credit problems in the past or not (X_{j2}). The remaining three variables are quantitative: the duration of the credit (X_{j3} with values between 4 and 72 months), the level of the credit (between DEM 338 and DEM 15653) and the age of the customer (between 19 and 75 years). We will take the logarithm of the last two variables and transform them linearly so that they take on a value in the interval [0, 1]. The data points, as can be seen in Fig. 21.1, are dispersed comparatively homogenous over a part of the plane, which makes the local smoothing easier. These transformed variables are called Z_{j1} and Z_{j2}. We fit a GPLM

$$P(Y_j = 1 \mid X_{j1} = x_1, X_{j2} = x_2, X_{j3} = x_3, Z_{j1} = z_1, Z_{j2} = z_2)$$

$$= \psi \left\{ \sum_{k=1}^{3} \beta_k x_k + m(z_1, z_2) \right\}$$
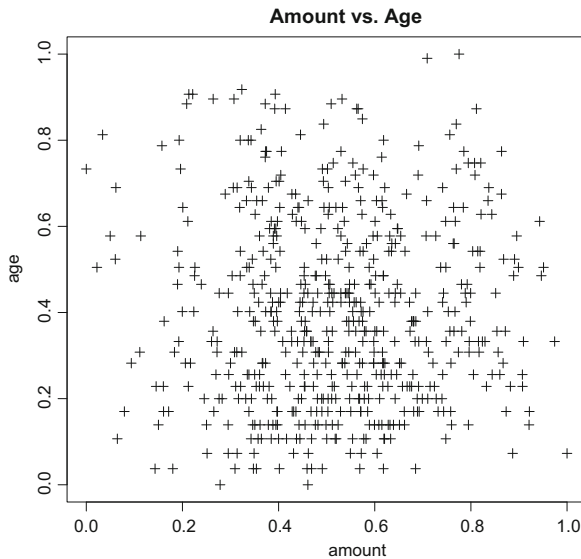
Fig. 21.1 The scatter plot of the transferred variables: level of credit and age of the customers
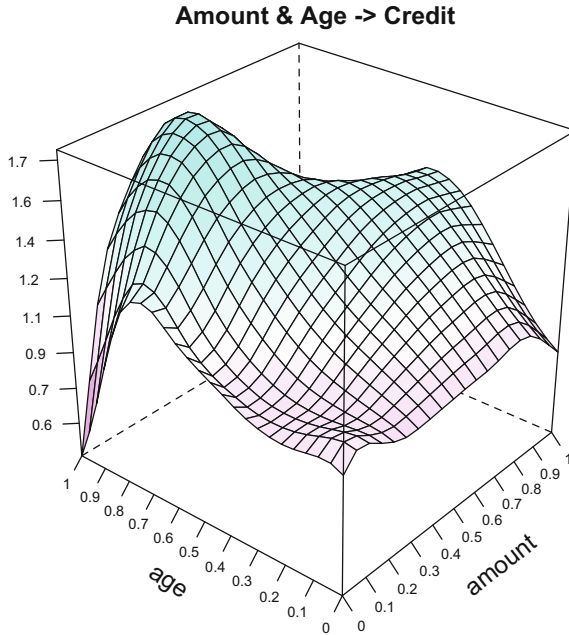
SFEgplm

Fig. 21.2 The estimated function with respect to level of credit and age of the customers
⬛ SFEgplm

to the data and obtain the estimates (the corresponding standard deviation is given in parentheses)

$$\beta_1 = 0.965\,(0.249), \ \beta_2 = 0.746\,(0.237), \ \beta_3 = -0.0498\,(0.0115).$$

The probability of default of the credit increases when the customer is unemployed or if the customer has had repayment problems in the past; this however decreases with the duration of the credit. The dependence on the transformed credit levels and ages are non-parametrically estimated. From Fig. 21.2 it is obvious that the estimated function $\hat{m}(z_1, z_2)$ is clearly non-linear with a maximum by the average value of the credit level and age. The decrease in the probability of default by high levels of credit can be explained by the fact that the random sample contains only those credits that have actually been given and that the processor was essentially reluctant to grant high levels of credit when the customer appeared to be unreliable. This effect, which is caused by credit ratings from the past, occurs on a regular basis in credit assessment, even if a systematic, model based method is not used, which excludes the credit screening of extreme risks from the very beginning and thus mean that these ratings no longer appear in the data. This must therefore be considered when interpreting and applying a model.

21.3   Credit Ratings with Neural Networks

As with non-parametric fitting of financial time series models to data, the neural network also provides an alternative to local smoothing, as with the LP method, in estimating default probabilities. The logistic regression function $\pi(x) = \psi(\beta_0 + \beta^\top x)$ is nothing more than a function defined by a neural network with only one neuron in a hidden layer, when the logistic function ψ is chosen as a transfer function. Through the combination of several neurons in one or more hidden layers default probabilities can be estimated, as with the non-parametric regression analysis, with flexibility. In order to obtain estimates between 0 and 1, it is necessary to represent the function $v_H(x; \delta)$ given by (19.1), for example, with a function G over the interval [0,1]. We restrict ourselves to one hidden layer with H neurons and choose $G = \psi$, so that the default probability given by the neuron network has the form

$$\pi_H(x; \vartheta) = \psi \left\{ v_0 + \sum_{h=1}^{H} v_h \, \psi \left(w_{oh} + \sum_{j=1}^{d} w_{ih} x_i \right) \right\},$$

where ϑ once again represents the parameter vector built from v_h and w_{ih}, $0 \leq i \leq d$, $0 \leq h \leq H$. To estimate the network weights from the data we will not use the least squares method, which makes sense for the regression model with normally distributed residuals, but instead we will maximize the log-likelihood function

$$\log L(\vartheta) = \sum_{j=1}^{n} [Y_j \log \pi_H(X_j; \vartheta) + (1 - Y_j) \log\{1 - \pi_H(X_j; \vartheta)\}]$$

following the procedure used in the logistic regression. By substituting $\hat{\vartheta}_n$ in the estimator we obtain an estimator for the default probability

$$\hat{\pi}(x) = \pi_H(x; \hat{\vartheta}_n).$$

In order to obtain an especially simple model with fewer parameters, Anders (1997) trivially modified the method for the insolvency prognoses for small and middle sized firms and assumed a default probability of the form

$$\pi_H^l(x; \vartheta) = \psi \left\{ \beta^\top x + v_0 + \sum_{h=1}^{H} v_h \psi \left(w_{oh} + \sum_{i=1}^{d} w_{ih} x_i \right) \right\},$$

which has obvious similarities to the general partial linear model, besides the fact that here a part of or all of the influential variables, i.e., the coordinates of x, can appear in linear as well as in non-parametric portions. The linear term $\beta^\top x$ can be interpreted as the value of an additional neuron whose transfer function is not the logistic function $\psi(t)$, but the identity $f(t) \stackrel{\text{def}}{=} t$. Estimating the network

from the application of a model selection technique, utilized to find the insolvency probability, is surprisingly easy. In addition to a linear term it contains only one single neuron ($H = 1$). From the six input variables only four contribute to the linear part (age of the business, sales development, indicator for limited liability, dummy variable for processed business); this means the other two coefficients β_i are 0, and only three (Dummy variables for processed business and for trade, indicator variable for educational degree of entrepreneur) contribute to the sigmoid part, meaning that the corresponding weights w_{i1} are 0. With this simple model using a validation data set, which is not used to estimate the parameters, a ratio of the correct identifications of 83.3 % was obtained for the insolvencies and of 63.3 % for the solvent companies.

Chapter 22
Credit Risk Management and Credit Derivatives

Credit risk management is an important issue in banking. In this chapter we give an overview of the models for calculating the default risk exposure of a credit portfolio. The primary goal of these models is to help credit analysts define whether a loan should be issued, which risk premia is appropriate and how much capital should be directed to the loss reserve account. In our presentation we follow Bluhm, Overbeck and Wagner (2010), Li (1998). In the remaining section an important example of a credit risk derivative is shown. A collateralized debt obligation (CDO) is a financial instrument that enables securitization of a large portfolio of assets (i.e. mortgages, auto loans, credit card debt, corporate debt, or credit default swaps (CDS)). This financial product became extremely popular on the brink of the Global Financial Crisis (2007–2009) and therefore captured mainstream media attention, mainly in a negative light. Constructed mostly as a collateral mortgages derivative, CDOs were then next traded on the global financial markets and became (very often highly rated) a part of an institutional investor's portfolio. After the US housing bubble burst, i.e. that instruments contributed to the spillover effect and thus to the globalization of the crisis. In this chapter we focus on the structure of the CDO and present a modelling and pricing technique.

22.1 Basic Concepts

Definition 22.1 (Loss Function) The loss fraction in case of default is called loss given default (LGD). The exposure at default in a considered time period is abbreviated to EAD. The loss of an obligor is thus defined by the following loss function:

$$\tilde{L} = EAD \times LGD \times L$$

with $L = \mathbf{1}(D)$. Here D stands for the default event of an obligor in a given time period (e.g. 1 year). $\mathrm{P}(D)$ is the probability of the event D.

Definition 22.2 (Expected Loss) The expected loss (EL) is defined as:

$$EL = \mathsf{E}(\tilde{L}) = EAD \times LGD \times \mathrm{P}(D).$$

For the sake of simplicity we assume here (and thereafter) that EAD and LGD are deterministic, therefore implying their independence from the default event. However modifying these assumptions may lead to more realistic models.

To calculate *EL* we need to find default probabilities which could be inferred either from credit ratings, market prices of defaultable bonds or credit derivatives. According to the first approach we use data on default frequencies for different rating classes to perform a mapping from the ratings space into the default probabilities space.

The expected loss *EL* defines the necessary loss reserve that a bank must hold as insurance against a default. In addition to the expected loss, the bank should have a cushion to cover unexpected losses.

Definition 22.3 (Unexpected Loss) The unexpected loss (UL) is defined as:

$$UL = \sqrt{\mathsf{Var}(\tilde{L})} = \sqrt{\mathsf{Var}(EAD \times LGD \times L)}$$

with $\mathsf{Var}(L) = \mathrm{P}(D)\{1 - \mathrm{P}(D)\}$.

So far we have considered the loss estimates for a single obligor. Now assume we have a credit portfolio consisting of m loans.

Definition 22.4 (Portfolio Loss) The expected portfolio loss is defined by the following random variable:

$$\tilde{L}_{PF} = \sum_{i=1}^{m} \tilde{L}_i = \sum_{i=1}^{m} EAD_i \times LGD_i \times L_i$$

with $L_i = \mathbf{1}(D_i)$.

Analogously to the single obligor case we can calculate EL_{PF} and UL_{PF}:

$$EL_{PF} = \sum_{i=1}^{m} EL_i = \sum_{i=1}^{m} EAD_i \times LGD_i \times \mathrm{P}(D_i)$$

$$UL_{PF} = \sqrt{\sum_{i,j=1}^{m} EAD_i \times EAD_j \times LGD_i \times LGD_j \times \mathsf{Cov}(L_i, L_j)}.$$

It is possible to rewrite the covariance term as follows:

$$\mathsf{Cov}(L_i, L_j) = \sqrt{\mathsf{Var}(L_i) \times \mathsf{Var}(L_j) \times \rho_{ij}}.$$

We now obviously face the problem of the unknown default correlations ρ_{ij}. One could bear down this pop-up problem by assuming that loss variables are uncorrelated. This severely contradicts empirical observations: defaults are likely to happen jointly. In fact dependency between obligors becomes the main driver of credit risk and the key issue in credit modelling. So it all boils down to understanding joint defaults. This is even more important in pricing collateral debt obligations (CDO), see Sect. 22.7. We therefore also introduce statistical models which calibrate default correlations.

22.2   The Bernoulli Model

In the preceding section we implicitly introduced the Bernoulli loss variable defined as $L_i \sim B(1; p_i)$, with L_i being the default variable of obligor i, i.e. loss is generated with probability p_i and not generated with probability $(1 - p_i)$. The fundamental idea in the modelling of joint defaults is the randomization of the involved default probabilities. While in our previous analysis we considered extracted market data or rating agencies default probabilities, now we assume that the loss probabilities are random variables that follow some distribution F within $[0, 1]^m$: $P = (P_1, \ldots, P_m) \sim F$.

We assume that Bernoulli loss variables $L_1, \ldots, L_m$ are independent conditional on a realization $p = (p_1, \ldots, p_m)$ of P. The joint distribution of the loss function is then:

$$\mathsf{P}(L_1 = l_1, \ldots, L_m = l_m) = \int_{[0,1]^m} \prod_{i=1}^{m} p_i^{l_i} (1 - p_i)^{1-l_i} \, dF(p_1, \ldots, p_m), \quad (22.1)$$

where $l_i \in \{0, 1\}$. The first and second moments of the single losses L_i are:

$$\mathsf{E}(L_i) = \mathsf{E}(P_i), \quad \mathsf{Var}(L_i) = \mathsf{E}(P_i)\{1 - \mathsf{E}(P_i)\}$$

The covariance of single losses is given by:

$$\mathsf{Cov}(L_i, L_j) = \mathsf{E}(L_i, L_j) - \mathsf{E}(L_i)\,\mathsf{E}(L_j) = \mathsf{Cov}(P_i, P_j) \quad (22.2)$$

The correlation for two counterparties' default is:

$$\mathsf{Corr}(L_i, L_j) = \frac{\mathsf{Cov}(P_i, P_j)}{\sqrt{\mathsf{E}(P_i)\,\{1 - \mathsf{E}(P_i)\}} \sqrt{\mathsf{E}(P_j)\,\{1 - \mathsf{E}[P_j]\}}}. \quad (22.3)$$

Thus we have succeeded in expressing the unknown default correlations in terms of covariances of the F distribution. Later in this chapter we will illustrate how to obtain an appropriate specification for the distribution of default probabilities and consequently solve the default correlations.

A major simplification is possible if one assumes an equal, random default probability P_i for all obligors. Such a simplification makes sense for, e.g., portfolios with loans of comparable size and with similar risk characteristics. In this case (22.1) simplifies to

$$P(L_1 = l_1, \ldots, L_m = l_m) = \int_0^1 p^k (1 - p)^{m-k} \, dF(p) \qquad (22.4)$$

where $k = \sum_{i=1}^m l_i$ is the number of defaults in the credit portfolio. Note that EL equals:

$$\overline{p} = \int_0^1 p \, dF(p) \qquad (22.5)$$

Therefore the default correlation between two different counterparties equals:

$$\rho_{ij} = \text{Corr}(L_i, L_j) =$$

$$\frac{P(L_i = 1, L_j = 1) - \overline{p}^2}{\overline{p}(1 - \overline{p})} = \frac{\int_0^1 p^2 dF(p) - \overline{p}^2}{\overline{p}(1 - \overline{p})}. \qquad (22.6)$$

Formula (22.6) shows that higher volatility of P corresponds to the default correlation. Since the numerator of (22.6) equals $\text{Var}(P) \geq 0$ the default correlation in the Bernoulli model is always positive and cannot mimic negative default correlation.

22.3   The Poisson Model

Another common approach to joint default modelling is an assumption of the Poisson-distributed loss variable L_i with intensity Λ_i. This means that $L_i \sim Pois(\lambda_i)$, $p_i = P(L_i \geq 1)$, $L_i \in \{0, 1, 2, \ldots\}$ modelling the fact that multiple defaults of obligor i may occur. It is worth mentioning here that this distribution rather deserves to be called the Bortkiewicz distribution since it was Ladislaus von Bortkiewicz, who really brought this distribution into the standard curriculum universale. Analogously to the Bernoulli mixture model we not only assume the loss variable vector L but also the intensity vector $\Lambda = (\Lambda_1, \ldots, \Lambda_m)$ to be random: $\Lambda \sim F$ within $[0, \infty)^m$. Also assuming that $L_1, \ldots L_m$ (conditional on a realization of Λ) are independent. The joint distribution of L_i is given:

$$\mathsf{P}(L_1 = l_1, \ldots, L_m = l_m)$$

$$= \int_{[0,\infty)^m} e^{-(\lambda_1 + \ldots + \lambda_m)} \prod_{i=1}^m \frac{\lambda_i^{l_i}}{l_i!} \, dF(\lambda_1, \ldots, \lambda_m), \qquad (22.7)$$

Similar to the Bernoulli case, we have for $i = 1, \ldots, m$:

$$\mathsf{E}(L_i) = \mathsf{E}(\Lambda_i)$$

$$\mathsf{Var}(L_i) = \mathsf{Var}\{\mathsf{E}(L_i|\Lambda)\} + \mathsf{E}\{\mathsf{Var}(L_i|\Lambda)\} = \mathsf{Var}(\Lambda_i) + \mathsf{E}(\Lambda_i). \qquad (22.8)$$

The correlation is then given:

$$\mathsf{Corr}(L_i, L_j) = \frac{\mathsf{Cov}(\Lambda_i, \Lambda_j)}{\sqrt{\mathsf{Var}(\Lambda_i) + \mathsf{E}[\Lambda_i]}\sqrt{\mathsf{Var}(\Lambda_j) + \mathsf{E}[\Lambda_j]}}. \qquad (22.9)$$

As in the Bernoulli Model we can express the default correlation through the covariances of the intensity vector distribution F. For approximately uniform portfolios we take a single distribution for all obligors. The analogue of (22.6) is then:

$$\mathsf{Corr}(L_i, L_j) = \frac{\mathsf{Var}(\Lambda)}{\mathsf{Var}(\Lambda) + \mathsf{E}(\Lambda)}. \qquad (22.10)$$

A more intuitive interpretation is given by the dispersion of Λ. The dispersion of a random variable is its variance to mean ratio, e.g. $D(\Lambda) \stackrel{\text{def}}{=} \frac{\mathsf{Var}(\Lambda)}{\mathsf{E}(\Lambda)}$, for the Poisson distribution it is equal to 1. Using dispersion we can rewrite formula (22.10):

$$\mathsf{Corr}(L_i, L_j) = \frac{D(\Lambda)}{D(\Lambda) + 1}. \qquad (22.11)$$

Therefore we conclude: an increase in dispersion will increase the mixture effect, which strengthens the dependence between obligor's defaults.

22.3.1   Bernoulli vs. Poisson

Comparing Bernoulli with Poisson distribution of the default risk, we see that a higher default correlation in Bernoulli distribution than in Poisson distribution always exists. In other words even in cases where the mean of Bernoulli matches the Poisson distribution, the Poisson variance will always exceed the variance of Bernoulli. The higher default correlations result in fatter tails of the corresponding loss distributions.

22.4   The Industrial Models

22.4.1   CreditMetrics™ and KMV Models

Two well-known factor models are *CreditMetrics^{TM}* and KMV. Both models belong to the Bernoulli class and imply only two possible outcomes—default or survive. Default of an obligor i occurs if the value of the obligor's assets $A_T^{(i)}$ in a valuation horizon T falls below a threshold value C_i, often interpreted as the value of the obligor's liabilities.

$$L_i = \mathbf{1}\{A_T^{(i)} < C_i\} \sim B\{1; \mathrm{P}(A_T^{(i)} < C_i)\} \tag{22.12}$$

Thus A_T can be regarded as a latent variable, which drives the default event implicitly replacing the notion of default correlation for the asset correlation. How is the correlation matrix of the latent variables defined? The answer lies in the basic assumption of both models, according to which the asset value dynamics relate to the changes in some common, economic factors. In the typical model parametrization of the latent variables one looks at asset log-returns:

$$r_i = \frac{\tilde{r}_i - \mathsf{E}(\tilde{r}_i)}{\sqrt{\mathsf{Var}(\tilde{r}_i)}} \quad \text{with} \quad \tilde{r}_i = \log(A_T^{(i)}/A_0^{(i)}).$$

Suppose that the standardized log return can be written as:

$$r_i = R_i Y_i + \varepsilon_i. \tag{22.13}$$

Here Y_i represents a weighted sum of many industry and country indices (composite factor). From regression analysis we conclude that R_i^2 defines how much the volatility of r_i can be explained by the volatility of Y_i and therefore it stands for the systematic risk of the obligor i. Respectively ε_i is the firm-specific effect.

The core assumption of *CreditMetrics^{TM}* and KMV models is the multivariate normal (Gaussian) distribution of the latent variables r_i:

$$r_i \sim \mathrm{N}(0, 1)$$
$$Y_i \sim \mathrm{N}(0, 1)$$
$$\varepsilon_i \sim \mathrm{N}(0, 1 - R_i^2)$$

In this case we can rewrite (22.12) as:

$$L_i = \mathbf{1}\{r_i < c_i\} \tag{22.14}$$

where c_i is the threshold corresponding to C_i after replacing A_T for the standardized log returns r_i . Using (22.13) we can rewrite the threshold condition $r_i < c_i$ as $\varepsilon_i < c_i - R_i Y_i$. Because $r_i \sim N(0, 1)$, from $p_i = P(r_i < c_i)$ we obtain

$$c_i = \Phi^{-1}(p_i).$$

After the standardizing of ε_i the threshold condition changes to:

$$\frac{\varepsilon_i}{\sqrt{1 - R_i^2}} < \frac{\Phi^{-1}(p_i) - R_i Y_i}{\sqrt{1 - R_i^2}}. \tag{22.15}$$

On the right-hand side of (22.15) Y_i is the only stochastic element. We therefore obtain (conditional on $Y_i = y$)

$$p_i(y) = \Phi\{\frac{\Phi^{-1}(p_i) - R_i y}{\sqrt{1 - R_i^2}}\}. \tag{22.16}$$

Transforming this into the Bernoulli mixture setting yields

$$P(L_1 = l_1, \ldots, L_m = l_m)$$

$$= \int_{[0,1]^m} \prod_{i=1}^{m} q_i^{l_i} (1 - q_i)^{1 - l_i} dF(q_1, \ldots, q_m).$$

Now we are able to specify the probability distribution function:

$$F(q_1, \ldots, q_m) = \Phi_m(\mu, \Gamma)$$

where $\mu = (p_1^{-1}(q_1), \ldots, p_m^{-1}(q_m))^\top$ and Γ is the asset correlation matrix of the log returns $\tilde{r}_i$.

The described modelling framework belongs to the KMV model. Though being based on the same assumptions, *CreditMetrics*[TM] differs from the KMV mainly in two areas: the asset value is replaced by the equity process and composite factors are redefined. For further information on the model please refer to *CreditMetrics*[TM] Technical Document.

22.4.2   *CreditRisk$^+$ Model*

In contrast to the KMV and *CreditMetrics*[TM] the *CreditRisk$^+$* belongs to the class of Poisson mixture models. Also, instead of common factors, it introduces the notion of sectors and is therefore sometimes called "sector model". A sector plays

a similar role in the model as compared to the factor but unlike the factor which has a certain economic interpretation, the sector may reflect any issues that have a systematic effect on the obligor's economic performance. Each sector is assumed to be Poisson distributed random variable with gamma distributed intensity $\Lambda^{(s)}$, where the variables $\Lambda^{(1)}, \ldots, \Lambda^{(m_S)}$ are assumed to be independent. For the reader's convenience the gamma distribution pdf is stated here:

$$\{\beta^\alpha \Gamma(\alpha)\}^{-1} x^{\alpha-1} \exp(-x/\beta).$$

The mean and variance of a gamma distributed random variable Λ are:

$$\mathsf{E}\,\Lambda = \alpha\beta \tag{22.17}$$

$$\mathsf{Var}(\Lambda) = \alpha\beta^2.$$

Consider a credit portfolio of m loans of m different obligors. In *CreditRisk$^+$* a weight ω_{is} of default intensity of obligor i to the systematic default risk of sector s is given, with $\sum_{s=1}^{S} \omega_{is} = 1$. The basic idea is that the risk of the sector $s \in \{1, \ldots, m_S\}$ is driven by two factors. The first driver is the mean default intensity:

$$\lambda_{(s)} = \mathsf{E}(\Lambda^{(s)}) = \alpha_s \beta_s. \tag{22.18}$$

The second driver is the default intensity's volatility:

$$\sigma_{(s)}^2 = \mathsf{Var}(\Lambda^{(s)}) = \alpha_s \beta_s^2. \tag{22.19}$$

The default risk of obligor i is modelled as well by a mixed Poisson variable L with the default intensity Λ_i with mean value $\mathsf{E}(\Lambda_i) = \lambda_i$. From the probability theory we know that for Poisson distributed variables with small intensity λ_i the following expression is true:

$$p_i = \mathsf{P}(L_i \geq 1) = 1 - e^{-\lambda_i} \approx \lambda_i. \tag{22.20}$$

Thus we can calibrate the default intensity from the obligor's 1-year default probability. The sector parametrization of Λ_i is given:

$$\Lambda_i = \sum_{s=1}^{S} \omega_{is} \lambda_i \frac{\Lambda^{(s)}}{\lambda_{(s)}} \tag{22.21}$$

This means that obligors admit a common source of systematic default risk (i.e. they are correlated), if, and only if, there is at least one sector with a positive weight with respect to the considered obligors. The formula (22.21) is consistent with the assumption that λ_i equals the expected default intensity of obligor i.

According to the formula (22.20), the conditional default intensity of obligor i that arises from the realizations $\theta_1, \ldots, \theta_{m_S}$ of the sector default intensities $\Lambda^{(1)}, \ldots, \Lambda^{(m_S)}$ generates a conditional 1-year default probability:

$$p_i(\theta_1, \ldots, \theta_{m_S}) = P(L_i \geq 1 | \Lambda^{(1)} = \theta_1, \ldots, \Lambda^{m_S} = \theta_{m_S})$$

$$= 1 - \exp(-\lambda_i \sum_{s=1}^{m_S} w_{is} \theta_s / \lambda_{(s)}) \qquad (22.22)$$

To make the idea behind this model more pellucid, we can relate the notion of sector to the notion of factor and the obligor's default intensity Λ_i—to the default probability p_i in the respective factor models. The latter expression could be related then to the formula (22.16) in the Bernoulli framework of the KMV model. As in the previous section, we will proceed now to the problem of the portfolio loss distribution specification. Assuming we know the distribution of defaults in every single sector we can obtain the portfolio's default distribution as the convolution of the sector distributions due to the independence of the sector variables $\Lambda^{(1)}, \ldots, \Lambda^{(m_S)}$. For details on how to find the sectors' default distributions please refer to Bluhm et al. (2010) or directly to *CreditRisk*$^+$ Technical Document.

22.4.3   Other Models

Among other "best-practice" industrial models is the CreditPortfolioView (CPV) generated from a macroeconomic approach. Before moving on we should explain the notion of migration matrix used in the model. Migration matrix contains probabilities for the rated bonds of being downgraded, upgraded or retaining the same grade in a set time period. For every rating class the number of transitions and, respectively, the probabilities of these transitions will add up to the total number of possible ratings within a certain rating system.

CPV is a rating-based portfolio model that incorporates the dependence of default and migration probability on business cycles. The model uses Monte Carlo simulation to generate migration matrices and thus generates macro-scenarios. The losses arising in every possible economic environment are then tabulated and, based on these losses, the portfolio loss distribution is constructed. A significant advantage of this model is in its ability to allow for the fat tails in loss distribution, which can be observed empirically in the periods of deep depression.

Another effort to model fat-tailed loss distributions was undertaken by the so-called credit risk contagion models. The intuitive assumption is that a credit event at one company affects the solvency of related companies directly, thus increasing the concentration of loss events. Contagion models implement the framework of factor models adding contagion as changes to asset values of related companies as a result of defaults. The strength of a contagion is usually measured as a function

of the percentage of sales from each company to each of the other companies. Unfortunately, this approach requires data that is seldom available. For details please refer to Egloff, Leippold and Vanini (2004).

A fundamentally different approach to credit risk modelling is provided by the dynamic intensity models. The theory underlying these models is mathematically complex and beyond the scope of this chapter. The basic idea is that an intensity process $\lambda(t)$ described by a stochastic differential equation exists. This process $\lambda(t)$ drives the default times. By definition default in these models happens if the corresponding default time is less than the planning horizon T. More details can be found in Duffie and Singleton (1997).

22.5 One Factor Models

The multiple factor model was introduced in (22.13). A one factor model simplifies the analysis since there is only one driving factor common to all obligors: $Y \sim$ N(0, 1). In our discussion we concentrate on the KMV-Model. In a one factor set-up we model the (standardised) log returns:

$$r_i = \sqrt{\omega} Y + \sqrt{1 - \omega} Z_i \tag{22.23}$$

with idiosyncratic $Z_i \sim$ N(0, 1). The uniform asset correlation is denoted ω. As before Z_i is assumed to be independent from the factor Y. Given a single factor and identical for all obligors ω, we can rewrite Eq. (22.16) as:

$$p_i(y) = \Phi \left\{ \frac{\Phi^{-1}(p_i) - \sqrt{\omega} y}{\sqrt{1 - \omega}} \right\} \tag{22.24}$$

In order to demonstrate the dependence of $p_i(y)$ on the default probability given default y values, let us fix $\omega = 20\%$ and $y \in \{-3, 0, 3\}$. The variable $y \sim$ N(0, 1) can be interpreted as the state of the economy, $y = -3$ corresponds to a bad state, $y = 0$ means a typical state and $y = 3$ indicates a good state of the economy, see Fig. 22.1. The joint default probability (JDP) is given in the following proposition.

Proposition 22.1 *In a one-factor portfolio model with* $L_i \sim B(1, p_i)$*,* p_i *from (22.24), the JDP of two obligors is:*

$$JDP_{ij} = P(L_i = 1, L_j = 1) = \Phi_2 \left\{ \Phi^{-1}(p_i), \Phi^{-1}(p_j); \omega \right\},$$

where $\Phi_2(\cdot, \cdot; \omega)$ *denotes the bivariate normal cdf with correlation* ω *and marginal variances equal to* 1.
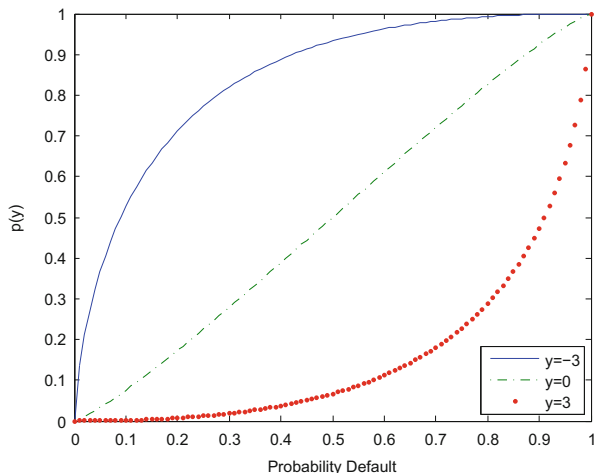
Fig. 22.1 The default probability $p(y)$ as a function of the state of the economy

SFEdefaproba

For a portfolio of m obligors, the portfolio marginal loss relative to the portfolio's total exposure is given by:

$$L^{(m)} = \sum_{i=1}^{m} w_i LGD_i L_i, \quad w_i = \frac{EAD_i}{\sum_{j=1}^{m} EAD_j}. \tag{22.25}$$

Bluhm et al. (2010) illustrate that for $m \to \infty$ the randomness of the portfolio loss $L^{(m)}$ solely depends on the randomness of the factor Y:

$$P[\lim_{m \to \infty} \{L^{(m)} - \mathsf{E}(L^{(m)}|Y)\} = 0] = 1. \tag{22.26}$$

Thus by increasing the number of obligors in the portfolio, the specific risk is completely removed and only the systematic risk arising from the volatility of the common factor remains in the portfolio.

Assuming uniform default probabilities p_i for all obligors i and applying the KMV framework to our analysis we infer:

$$\mathsf{E}(L^{(m)}|Y) = \sum_{i=1}^{m} w_i \, \mathsf{E}(L_i|Y) = \Phi\left\{\frac{\Phi^{-1}(p_i) - \sqrt{\omega}Y}{\sqrt{1-\omega}}\right\} = p(Y),$$

so that from formula (22.26) it follows that for $m \to \infty$

$$L^{(m)} \longrightarrow p(Y) = \Phi\left\{\frac{\Phi^{-1}(p_i) - \sqrt{\omega}Y}{\sqrt{1-\omega}}\right\} \qquad \text{a.s.} \tag{22.27}$$

We have thus established that for sufficiently large portfolios the percentage of defaulted loans given a certain state of economy $Y = y$ is approximately equal to the conditional default probability $p(Y)$. Now we want to derive the cdf of the limit loss variable $p(Y)$ and thus define the loss distribution. We denote here the limit of $L^{(m)}$ by L. For every $0 \leq x \leq 1$ we then have:

$$
\begin{aligned}
\mathrm{P}(L \leq x) = \mathrm{P}\{p(Y) \leq x\} &= \mathrm{P}\left[-Y \leq \frac{1}{\sqrt{\omega}} \left\{\Phi^{-1}(x)\sqrt{1-\omega} - \Phi^{-1}(p)\right\}\right] \\
&= \Phi\left[\frac{1}{\sqrt{\omega}} \left\{\Phi^{-1}(x)\sqrt{1-\omega} - \Phi^{-1}(p)\right\}\right].
\end{aligned}
\tag{22.28}
$$

Thus we can now calculate the corresponding pdf and find such portfolio loss characteristics such as expected and unexpected losses. Obviously, we first need a factor model to define asset correlation ω and some market data to calibrate the default probability p.

22.6 Copulae and Loss Distributions

Even holding asset correlations and default probabilities fixed there are alternative models which yield heavier-tailed loss distributions as compared to those induced by the assumptions of factor and sector models. In this section we describe how to use copulae for constructing such loss distributions. For details on the concept of copulae please refer to Chap. 17.

The copula presents an elegant way of understanding how a multivariate latent variable (e.g. asset returns) distribution determines the portfolio default distribution. From Sklar's Theorem we know that it is possible to extract copulae from known multivariate distribution functions or create new multivariate distributions by joining arbitrary marginal distributions in a known dependence structure, i.e. copula. In factor models we assume latent variables r to have multivariate Gaussian distribution with correlation matrix Γ that leads to the following copula of r:

$$
C_{r,Ga}(u_1, \ldots, u_m) = \Phi_m\left\{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_m)\right\}
\tag{22.29}
$$

$C_{r,Ga}$ is the well-known Gaussian copula, which represents the dependence structure underlying *CreditMetrics*^{TM} and KMV models, $\Phi^{-1}(u_i)$ denotes the u_i quantile of $N(0, 1)$. Replacing the normal dependency with a t-dependency will significantly shift mass into the tails of loss distribution as t-distributions possess fatter tails compared to normal. Fatness of tails is highly sensitive to the chosen degrees of freedom, however the problem of calibrating copulae is beyond the scope of this book. Let $n \geq 3$ and F_n be a t-distribution function with n degrees of freedom. We denote the multivariate t-distribution function as $F_{n,\Gamma}$ where Γ is the correlation matrix. The t-copula is:

$$
C_{n,\Gamma}(u_1, \ldots, u_m) = F_{n,\Gamma}\left\{F_n^{-1}(u_1), \ldots, F_n^{-1}(u_m)\right\}
\tag{22.30}
$$

where $u_1, \ldots, u_m \in [0,1]$. A new multivariate loss distribution can be created by combining the t-copulae $C_{n,\Gamma}$ with different marginal distributions. Below are a few examples of the two dimensional case.

- Bivariate Gaussian copulae with normal marginals. We generate pairs (X_1, X_2) with:

$$X_i = \sqrt{\omega} Y + \sqrt{1-\omega} Z_i$$

  where $Y, Z_1, Z_2 \sim N(0,1)$ i.i.d.
- Bivariate t-copulae with t-distributed margins. Generate (X_1, X_2) with

$$X_i = \sqrt{3}(\sqrt{\omega} Y + \sqrt{1-\omega} Z_i)/\sqrt{W}$$

  where $Y, Z_1, Z_2 \sim N(0,1)$ i.i.d. and $W \sim \chi^2(3)$.
- Bivariate t-copulae with normal margins. Generate (X_1, X_2) with

$$X_i = \Phi^{-1}[F_3 \left\{ \sqrt{3}(\sqrt{\omega} Y + \sqrt{1-\omega} Z_i)/\sqrt{W} \right\}]$$

  where $Y, Z_1, Z_2 \sim N(0,1)$ i.i.d. $W \sim \chi^2(3)$, W independent of Y, Z_1, Z_2, and F_3 denoting the t-distribution function with 3 degrees of freedom.

In order to study the goodness-of-fit of different copulae on the default models we recall (22.23) and rewrite it in the form of a t-distributed log return:

$$r_i = \sqrt{n/W} \sqrt{\omega} Y + \sqrt{n/W} \sqrt{1-\omega} Z_i \sim t(n)$$

for $i = 1, \ldots, m$. The loss function is $L_i = \mathbf{1}\{\tilde{r}_i < F_n^{-1}(p)\}$, and the default probability conditional on Y and W turns into:

$$
\begin{aligned}
p(y, w) &= P(r_i \leq F_n^{-1}(p)|Y = y, W = w) \\
&= P(\sqrt{n/W} \sqrt{\omega} Y + \sqrt{n/W} \sqrt{1-\omega} Z_i \leq F_n^{-1}(p)|Y = y, W = w) \\
&= \Phi\{ \frac{\sqrt{w/n} F_n^{-1}(p) - \sqrt{\omega} y}{\sqrt{1-\omega}} \}
\end{aligned}
\tag{22.31}
$$

Bluhm et al. (2010) simulated the portfolio loss in such a t-copulae model by looking at the distribution of (22.31) with independent $\{Y_i\}_{i=1}^N \sim N(0,1)$, $\{W_i\}_{i=1}^N \sim \chi^2(n)$, with $N = 100{,}000$ simulations and sample size $n = 10, 40, 10{,}000$. They concluded that for large n the t-copulae resemble the Gaussian copulae. For small n though there may be a big difference in the 99 % quantiles.

To conclude this chapter we will show how to estimate asset correlation from historic default frequencies using a one factor model. Our first step is to calibrate default probabilities. Table 22.1 presents Moody's historic corporate bond default frequencies from 1970 to 2011. For each rating class we calculate the mean and the

Table 22.1 Moody's corporate bond historic default frequency 1970–2011. Frequencies different than 0 are bold

Year	Aaa	Aa	A	Baa	Ba	B
1970	0.00	0.00	0.00	**0.27**	**4.12**	**20.78**
1971	0.00	0.00	0.00	0.00	**0.42**	**3.85**
1972	0.00	0.00	0.00	0.00	0.00	**7.14**
1973	0.00	0.00	0.00	**0.46**	0.00	**3.77**
1974	0.00	0.00	0.00	0.00	0.00	**10.00**
1975	0.00	0.00	0.00	0.00	**1.02**	**5.97**
1976	0.00	0.00	0.00	0.00	**1.01**	0.00
1977	0.00	0.00	0.00	**0.28**	**0.52**	**3.28**
1978	0.00	0.00	0.00	0.00	**1.08**	**5.41**
1979	0.00	0.00	0.00	0.00	**0.49**	0.00
1980	0.00	0.00	0.00	0.00	0.00	**4.94**
1981	0.00	0.00	0.00	0.00	0.00	**4.49**
1982	0.00	0.00	**0.26**	**0.31**	**2.72**	**2.41**
1983	0.00	0.00	0.00	0.00	**0.91**	**6.31**
1984	0.00	0.00	0.00	**0.36**	**0.83**	**6.72**
1985	0.00	0.00	0.00	0.00	**1.40**	**8.22**
1986	0.00	0.00	0.00	**1.33**	**2.03**	**11.73**
1987	0.00	0.00	0.00	0.00	**2.71**	**6.23**
1988	0.00	0.00	0.00	0.00	**1.24**	**6.36**
1989	0.00	**0.61**	0.00	**0.59**	**2.98**	**8.95**
1990	0.00	0.00	0.00	0.00	**3.35**	**16.18**
1991	0.00	0.00	0.00	**0.27**	**5.35**	**14.56**
1992	0.00	0.00	0.00	0.00	**0.30**	**9.03**
1993	0.00	0.00	0.00	0.00	**0.56**	**5.71**
1994	0.00	0.00	0.00	0.00	**0.24**	**3.82**
1995	0.00	0.00	0.00	0.00	**0.69**	**4.81**
1996	0.00	0.00	0.00	0.00	0.00	**1.44**
1997	0.00	0.00	0.00	0.00	**0.19**	**2.12**
1998	0.00	0.00	0.00	**0.12**	**0.63**	**4.26**
1999	0.00	0.00	0.00	**0.10**	**1.01**	**5.85**
2000	0.00	0.00	0.00	**0.38**	**0.89**	**5.49**
2001	0.00	0.00	**0.16**	**0.19**	**1.57**	**9.36**
2002	0.00	0.00	**0.16**	**1.21**	**1.54**	**4.97**
2003	0.00	0.00	0.00	0.00	**0.95**	**2.66**
2004	0.00	0.00	0.00	0.00	**0.19**	**0.65**
2005	0.00	0.00	0.00	**0.18**	0.00	**1.04**
2006	0.00	0.00	0.00	0.00	0.00	**0.20**
2007	0.00	0.00	0.00	0.00	**0.20**	**1.21**
2008	0.00	**0.54**	**0.47**	**0.48**	**1.19**	**2.12**
2009	0.00	0.00	**0.18**	**0.87**	**2.46**	**7.41**
2010	0.00	0.00	**0.22**	0.00	0.00	**0.50**
2011	0.00	0.00	0.00	**0.08**	**0.19**	**0.11**

All numbers are percentages values   ▣ SFEdefault

standard error of the historic default frequencies. Then we use simple regression to fit the mean by an exponential function. As a result we obtain fitted default probabilities $\mu_1, \ldots, \mu_6$ for all rating classes, see Table 22.2. Analogously, the historic volatilities are modelled by an exponential function and $\sigma_1, \ldots, \sigma_6$ denotes the regression fit to the historical default volatilities for all rating classes.

The second step includes calculating the asset correlations. We refer to the formula (22.24) from the uniform factor model, in which we replace the true default probability p_i for the fitted mean default rate μ_i. It can be shown that the following expression is true for the considered model:

$$\mathsf{Var}\{p(Y)\} = \Phi_2\{\Phi^{-1}(p), \Phi^{-1}(p); \omega\} - p^2 \qquad (22.32)$$

where we again replace the true unknown variance of default rate for the fitted default volatility σ. Thus, the asset correlation ω is the only unknown parameter in (22.32). The calibrated correlations are shown in the last column of Table 22.2 (Fig. 22.2).

Table 22.2 Calibration results due to exponential function fitting   ⧠ SFEdefault

Rating	Mean (%)	Stand. dev. (%)	μ (%)	σ (%)	ρ (%)
Aaa	0.000	0.000	0.0022	0.019	20
Aa	0.014	0.094	0.010	0.056	27
A	0.034	0.095	0.047	0.155	23
Baa	0.173	0.315	0.219	0.434	18
Ba	1.071	1.227	1.014	1.215	14
B	5.477	4.483	4.684	3.400	10
Mean	1.128	1.035	0.9964	0.8803	21
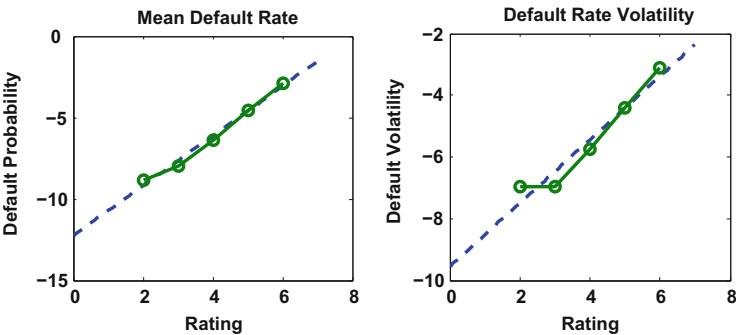
Fig. 22.2 Mean default rate and default rate volatility. The *solid* and *circle-dashed lines* represent the historic default and the regression by exponential function fitting correspondingly. All values are logged data   ⧠ SFEdefault

22.7 Collateralized Debt Obligations

A CDO is a financial instrument that enables securitization of a large portfolio of assets. The portfolio's risk is sliced into tranches of increasing seniority and then sold separately on the secondary market. Investors, according to their risk preferences, provide a risk protection of a particular tranche and receive for that a fixed premium. Premium payments are made by the owner of the assets on a regular basis until the termination of the agreement. The obligation of investors is to cover all losses generated by the portfolio entities within their tranche. Each tranche has specified priority of bearing claims and of receiving periodic payments. Assets in the collateral may be mortgages, auto loans, credit card debt, corporate debt, or CDS.

A CDO transaction has two sides—asset and liability—linked by cash flows (see Fig. 22.3). The asset side refers to the underlying reference portfolio while the liability side consists of securities issued by an issuer, which is often a special purpose vehicle (SPV). An SPV is a company created by an owner of a pool especially for the transaction to insulate investors from the credit risk of the CDO originator. An originating institution, usually a bank, sells assets to an SPV to manage exposure and shrink their balance sheets but it often keeps the administration of the pool.

A CDO tranche is defined by the attachment (also called lower attachment) and detachment (upper attachment) points which are the percentages of the portfolio losses. Table 22.3 presents the tranching of the iTraxx index. It shows that the most subordinated tranche, called the equity tranche, bears the first 3 % losses of the portfolio nominal. It is common that a bank keeps this riskiest piece. If losses constitute 5 % of the collateral notional, the equity investors carry the first 3 % (thus losing all their investment), and the next 2 % is covered by those who invested in the mezzanine junior tranche. The tranches that carry the lowest risk are called senior tranches and get the highest credit ratings. They receive periodic income at first and as the last, are affected by any losses. The senior tranche holders suffer only if the
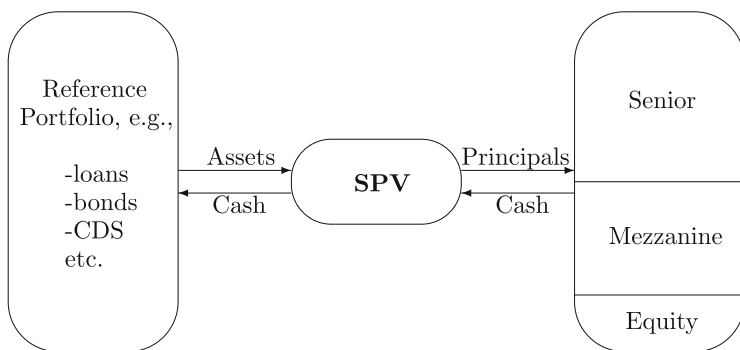
Fig. 22.3 Illustration of a CDO cash flow mechanism

Table 22.3 Example of a CDO tranche structure, iTraxx

Tranche number	Tranche name	Attachment points (%)	
		Lower	Upper
1	Equity	0	3
2	Mezzanine Junior	3	6
3	Mezzanine	6	9
4	Senior	9	12
5	Super Senior	12	22
6	Super Super Senior	22	100

total collateral portfolio loss exceeds 22 % of its notional value. However, the higher the risk a tranche carries, the better return it gives. Mezzanines have lower ratings than senior tranches but give better returns. The unrated equity tranche offers the highest coupons and gets paid at the end.

Each loss in the portfolio reduces the notional on which the payments are based and also reduces the value of the periodic fee. After each default the seller of the protection makes a payment equal to the loss to the protection buyer. When the portfolio losses exceed the detachment point, no notional remains and no payment is made.

iTraxx is a group of international credit derivative indices that are administered by Markit Group Limited, the financial data provider. The iTraxx indices cover credit derivatives markets in Europe, Asia and Australia. iTraxx Europe is one of the most popular indices for which the standard tranches are quoted. The reference portfolio of it consists of 125 equally weighted and most liquidly traded CDS contracts on European companies which represent six different industry sectors: consumer, financial, technology-media-telecommunications (TMT), industrials, energy and auto. The number of swaps in the groups is 30, 25, 20, 20, 20, 10, respectively. There are also indices of sub-portfolios: HiVol, Non-Financials, and Financials. Every 6 months, on 20 March and 20 September, new series of iTraxx Europe are issued and the underlying pool is reconstituted. The standardized tranches of iTraxx are available with maturities of 3, 5, 7 and 10 years.

Consider a CDO with a maturity of T years, J tranches, a total notional M and a pool of d entities. We assume that two parties of the CDO transaction enter into the contract on a trade date t_0 and on this day the protection begins. The first step in pricing a multiname credit derivative is to construct the default term structure of each underlying asset. Let τ_i be a positive random variable representing the time of default of the ith obligor, $i = 1, \ldots, d$. A term structure of a default probability, also called a credit curve, is defined as

$$p_i(t) = P(\tau_i \leq t)$$

and represents the probability that the obligor defaults within the time interval $[0, t]$. In this framework the obligor's default is modelled as the time until the first jump of

a Poisson process with a deterministic or a stochastic intensity. The unconditional default probabilities are related to the intensity function $\lambda_i(t)$ by the following equality

$$p_i(t) = 1 - \mathsf{E}\left[\exp\left\{-\int_0^t \lambda_i(u)du\right\}\right],$$

where the corresponding survival probability term structure is given by

$$\bar{p}_i(t) = 1 - p_i(t) = \mathrm{P}(\tau_i > t).$$

Every obligor is represented by a default indicator

$$I_i(t) = \mathbf{1}(\tau_i \leq t).$$

At time t the proportion of defaulted entities in the portfolio is given by

$$\tilde{L}(t) = \frac{1}{d} \sum_{i=1}^{d} I_i(t), \quad t \in [t_0, T].$$

The portfolio loss at time t is defined as

$$L(t) = \mathrm{LGD}\tilde{L}(t), \tag{22.33}$$

where LGD is a common LGD. Each tranche $j = 1, \ldots, J$ is defined by the attachment l_j and detachment u_j point which are the percentages of the portfolio losses and $l_j < u_j$. The loss L_j at time t is expressed as

$$L_j(t) = \frac{1}{u_j - l_j}\{L^u(t, u_j) - L^u(t, l_j)\},$$

where L^u is specified from (22.33)

$$L^u(t, x) = \min\{L(t), x\} \quad \text{for } x \in [0, 1].$$

The outstanding notional $\Gamma_j(t)$ of the tranche j is written as

$$\Gamma_j(t) = \frac{1}{u_j - l_j}\{\Gamma^u(t, u_j) - \Gamma^u(t, l_j)\},$$

where Γ^u is

$$\Gamma^u(t, x) = x - L^u(t, x) \quad \text{for } x \in [0, 1].$$

At the predefined dates $t = t_1, \ldots, T, t_1 > t_0$, the protection seller and the protection buyer exchange the payments. The fair spread of a CDO tranche is defined by the equivalence of the protection and premium leg. The protection (also called floating) leg DL_j is calculated as the expected value of the discounted stream of payments made upon defaults

$$\mathrm{DL}_j(t_0) = \sum_{t=t_1}^{T} \beta(t_0, t)\, \mathsf{E}\{L_j(t) - L_j(t - \Delta t)\}M, \ j = 1, \ldots, J,$$

where β is a discount factor and Δt is a time between t and the previous payment day. The credit event can happen at a random time point but to get a close form solution we make a slightly simplifying assumption that defaults occur at the end of the payment period. The premium (fixed) leg PL_j is expressed as the expectation of the present value of all premium payments

$$\mathrm{PL}_j(t_0) = \sum_{t=t_1}^{T} \beta(t_0, t) s_j(t_0) \Delta t\, \mathsf{E}\{\Gamma_j(t)\}M, \ j = 2, \ldots, J,$$

where s_j denotes the spread of the tranche j.

The equity tranche is priced differently than the other tranches. It pays an upfront fee once, at the inception of the trade and a fixed coupon of 500 bp during the life of the contract. The upfront payment, denoted by α, is expressed in percent and is quoted on the market. The premium leg of the most subordinated tranche turns into

$$\mathrm{PL}_1(t_0) = 100\alpha(t_0)M + \sum_{t=t_1}^{T} \beta(t_0, t) \cdot 500 \cdot \Delta t\, \mathsf{E}\{\Gamma_1(t)\}M.$$

A tranche spread s_j is calculated once, at t_0 in a way that the marked-to-market value of the tranche is zero, i.e. the value of the premium leg equals the value of the protection leg. Solving $PL_j(t_0) = DL_j(t_0)$ for $s_j(t_0)$ leads to the following solution

$$s_j(t_0) = \frac{\sum_{t=t_1}^{T} \beta(t_0, t)\, \mathsf{E}\{L_j(t) - L_j(t - \Delta t)\}}{\sum_{t=t_1}^{T} \beta(t_0, t)\Delta t\, \mathsf{E}\{\Gamma_j(t)\}}, \ \text{for } j = 2, \ldots, J. \qquad (22.34)$$

For the equity tranche the upfront payment given in percent is equal

$$\alpha(t_0) = 100 \sum_{t=t_0}^{T} (\beta(t, t_0)\, [\mathsf{E}\{L_1(t) - L_1(t - \Delta t)\} - 0.05\Delta t\, \mathsf{E}\{\Gamma_1(t)\}]\, M).$$

Since 2009 the quoting convention of the iTraxx Europe tranches has changed. Now all the tranches (0–3, 3–6, 6–9, 9–12, 12–22, 22–100 %) have a structure of

the equity tranche and trade with an upfront fee and a fixed running spread of 500, 500, 300, 100, 100 and 25 bp, respectively.

In the above formulae all expected values are computed under the risk-neutral measure with respect to a portfolio distribution which describes the dependency structure between the defaults of the reference entities. At the core of the CDO pricing lies a dependency model for portfolio credit risk. The standard market CDO pricing method is the KMV and *CreditMetrics^{TM}* one factor Gaussian model with the portfolio distribution (22.28).

A compound correlation $\rho(l_j, u_j)$, $j = 1, \ldots, J$, is defined as a parameter that prices the tranche j so that it fits the market value. A present value PV_j of a tranche $j = 2, \ldots, J$ is given by

$$\text{PV}_j(t_0) = \sum_{t=t_1}^{T} \beta(t_0, t) M \big[s_j(t_0) \Delta t \, \mathsf{E}_{\rho(l_j, u_j)} \{ \Gamma_j(t) \}$$

$$- \mathsf{E}_{\rho(l_j, u_j)} \{ L_j(t) - L_j(t - \Delta t) \} \big],$$

where the expected value is calculated with respect to the distribution determined by the compound correlation $\rho(l_j, u_j)$.

An implied correlation is a parameter calculated out of a market spread by inverting the pricing model, i.e. we keep the value of other parameters fixed and look for the correlation that matches the quoted tranche spread. The standard Gaussian model uses only one correlation to specify the loss distribution and to price all the tranches. However, the implied correlations are not the same across the tranches, see an example on the left panel of Fig. 22.5. The phenomenon observed is called an implied correlation smile.

Still, the main disadvantage of the compound parameters is that the mezzanine tranches are not monotonic in correlation. In consequence, there could be two parameters that yield the same market spread. Moreover, there is no guarantee that the implied parameter exists. These shortcomings caused the enhanced popularity of base correlations.

The base correlations were introduced by McGinty and Ahluwalia (2004) in the framework of the one factor Gaussian model. The main idea behind the concept of the base correlation is that every tranche can be decomposed into two tranches that have the lower attachment point zero. Being long the mezzanine tranche with the attachment points l_j and u_j can be viewed as being simultaneously long the equity tranche with the upper attachment point u_j and short the equity tranche with the upper attachment point l_j. The base correlations are uniquely determined as the equity tranche is monotone in correlation. Figure 22.4 shows the expected loss of the equity tranche as a decreasing function of the correlation parameter in the one factor Gaussian model. The base correlations are computed using a bootstrapping technique, i.e. we use the base correlation of the first tranche to calculate the second tranche, and so on.
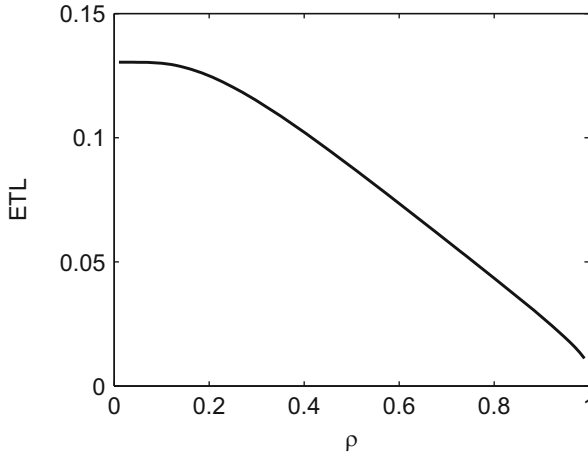
Fig. 22.4 Expected loss of the equity tranche calculated using the one factor Gaussian model with a 1-year default probability computed from the iTraxx index Series 8 with 5 years maturity on 20071022 ▣ SFEETLGaussTr1

The expected losses of successive tranches are calculated recursively

$$
\mathsf{E}\{L_{(3\%,6\%)}\} = \mathsf{e}_{\rho(0,6\%)}\{L_{(0,6\%)}\} - \mathsf{e}_{\rho(0,3\%)}\{L_{(0,3\%)}\},
$$

$$
\mathsf{E}\{L_{(6\%,9\%)}\} = \mathsf{e}_{\rho(0,9\%)}\{L_{(0,9\%)}\} - \mathsf{e}_{\rho(0,6\%)}\{L_{(0,6\%)}\}, \dots
$$

A present value PV_j of a tranche $j = 2, \dots, J$ in this approach is given by

$$
\mathrm{PV}_j(t_0) = \sum_{t=t_1}^{T} \beta(t_0, t) M\left(s_j(t_0)\Delta t \left[\mathsf{E}_{\rho(0,u_j)}\{\Gamma_j^u(t, u_j)\} - \mathsf{E}_{\rho(0,l_j)}\{\Gamma_j^u(t, l_j)\}\right] - \right.
$$

$$
\left. \mathsf{E}_{\rho(0,u_j)}\{L_j^u(t, u_j) - L_j^u(t - \Delta t, u_j)\} - \mathsf{E}_{\rho(0,l_j)}\{L_j^u(t, l_j) - L_j^u(t - \Delta t, l_j)\}\right)
$$

and the tranche spread (22.34) is calculated as

$$
s_j(t_0) = \frac{\sum_{t=t_1}^{T} \beta(t_0, t)\left[\mathsf{E}_{\rho(0,u_j)}\{L_j^u(t, u_j) - L_j^u(t - \Delta t, u_j)\}\right]}{\sum_{t=t_1}^{T} \beta(t_0, t)\Delta t \left[\mathsf{E}_{\rho(0,u_j)}\{\Gamma_j^u(t, u_j)\} - \mathsf{E}_{\rho(0,l_j)}\{\Gamma_j^u(t, l_j)\}\right]} -
$$

$$
\frac{\sum_{t=t_1}^{T} \beta(t_0, t)\left[\mathsf{E}_{\rho(0,l_j)}\{L_j^u(t, l_j) - L_j^u(t - \Delta t, l_j)\}\right]}{\sum_{t=t_1}^{T} \beta(t_0, t)\Delta t \left[\mathsf{E}_{\rho(0,u_j)}\{\Gamma_j^u(t, u_j)\} - \mathsf{E}_{\rho(0,l_j)}\{\Gamma_j^u(t, l_j)\}\right]},
$$

where the expected value $\mathsf{E}_{\rho(0,u_j)}$ is calculated with respect to the portfolio loss distribution determined by the base correlation $\rho(0, u_j)$ of the tranche $[0, u_j]$.

   Although the base correlations overcome some limitations of the compound correlations, they also have drawbacks. It can be shown that even if the true default correlation increases, base correlations might decrease. Moreover, the expected losses for mezzanine tranches can be negative.

   In the Gaussian copula model the base correlations are non-decreasing with respect to the seniority of tranches and the implied correlation smile turns into a correlation skew. Figure 22.5 compares both implied correlation curves computed using the tranches of iTraxx Europe from 22 October 2007 and assuming a constant LGD of 60 % and an interest rate of 3 %.
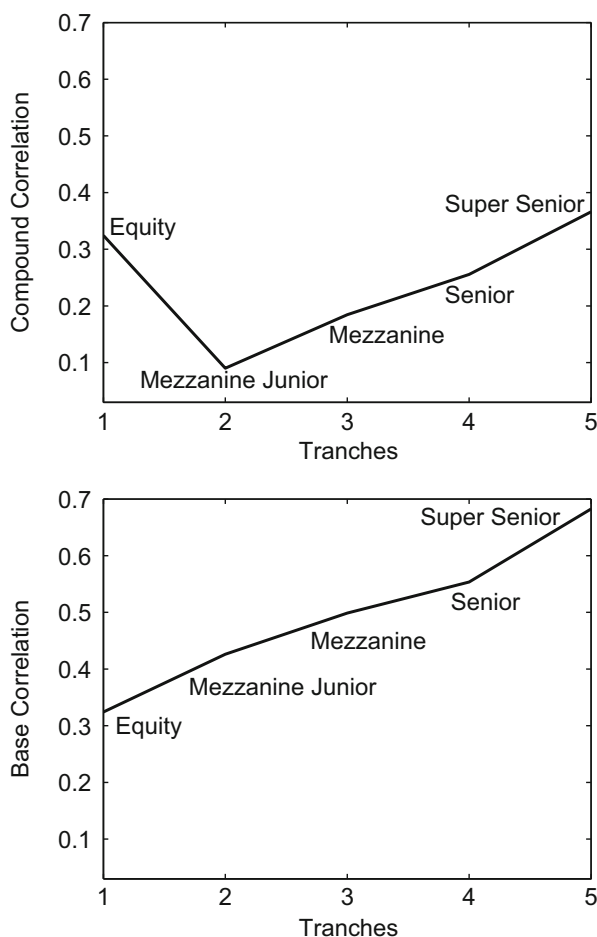
Fig. 22.5 Implied compound (*upper panel*) and base (*lower panel*) correlation smile from the one factor Gaussian model. Data from 20071022, iTraxx Series 8 with 5 years maturity

SFEbaseCorr SFEcompCorr

In addition, this measure allows pricing of the tranches that are not quoted on the market. Since the base correlation depends only on a lower attachment point, we can use it to value off-market tranches by an interpolation approach.

22.8 Exercises

Exercise 22.1 (Expected Loss) *Assume a zero coupon bond repaying full par value 100 with probability 95 % and paying 40 with probability 5 % in 1 year. Calculate the expected loss.*

Exercise 22.2 (Expected Loss) *Consider a bond with the following amortization schedule: the bond pays 50 after half a year (T_1) and 50 after a full year (T_2). In case of default before T_1 the bond pays 40 and in case of default in $[T_1, T_2]$ pays 20. Calculate the expected loss when the probabilities of default in $[0, T_1)$ and $[T_1, T_2]$ are*

(a) 1 and 4 %
(b) 2.5 and 2.5 %
(c) 4 and 1 %

respectively.

Exercise 22.3 (Joint Default) *Consider a simplified portfolio of two zero coupon bonds with the same probability of default (PD), par value 1 and zero recovery. The loss events are correlated with correlation ρ.*

(a) Calculate the loss distribution of the portfolio,
(b) Plot the loss distribution for PD = 20 % and $\rho = 0; 0.2; 0.5; 1$.

Exercise 22.4 (Bernoulli Model) *Consider a simplified Bernoulli model of $m = 100$ homogeneous risks with the same loss probabilities $P_i = P$ coming from the beta distribution. The density of the beta distribution is*

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \mathbf{1}\{x \in (0, 1)\}.$$

Plot the loss distribution of $L = \sum_{i=1}^{m} L_i$ for the following set of parameters

(i) $\alpha = 5, \beta = 25$
(ii) $\alpha = 10, \beta = 25$
(iii) $\alpha = 15, \beta = 25$
(iv) $\alpha = 5, \beta = 45$
(v) $\alpha = 10, \beta = 90$
(vi) $\alpha = 20, \beta = 180$

Exercise 22.5 (Poisson Model) *Consider a simplified Poisson model of $m = 100$ homogeneous risks with same intensities $\Lambda_i = \Lambda$ coming from the gamma distribution. The density of the gamma distribution is*

$$f(x) = \{\Gamma(\alpha)\beta^\alpha\}^{-1} x^{\alpha-1} \exp(-x/\beta).$$

Plot the loss distribution of $L = \sum_{i=1}^m L_i$ for the following set of parameters

  (i) $\alpha = 2$, $\beta = 5$
 (ii) $\alpha = 4$, $\beta = 5$
(iii) $\alpha = 6$, $\beta = 5$
 (iv) $\alpha = 3$, $\beta = 3.33$
  (v) $\alpha = 2$, $\beta = 5$
 (vi) $\alpha = 10$, $\beta = 1$

Exercise 22.6 (Bernoulli vs. Poisson Model) *Consider the Bernoulli model with the same loss probabilities $P_i = P$ and the Poisson model with intensities $\Lambda_i = \Lambda$. Assume that P and Λ have the same mean and variance.*

(a) *Show that the variance of the individual loss in the Poisson model exceeds the variance of the individual loss in the Bernoulli model.*
(b) *Show that the correlation of two losses in the Poisson model is smaller than in the Bernoulli model.*

Exercise 22.7 (Calibrate of Representative Portfolio) *Assume a portfolio of N obligors. Each asset has a notional value EAD_i, probability of default p_i, correlation between default indicators $\mathsf{Corr}(L_i, L_j) = \rho_{i,j}$ for $i, j = 1, \ldots, N$. For simplicity assume no recovery. Analysis of the loss distribution of this portfolio can be simplified by assuming a homogeneous portfolio of D uncorrelated risks with same notional $\widetilde{EAD}$ and probability of default $\tilde{p}$. Calibrate the representative portfolio such that total exposure, expected loss and variance match the original portfolio.*

Exercise 22.8 (Homogeneous Portfolio) *Follow the assumption from Exercise 22.7 for the homogeneous portfolio, i.e. $EAD_i = EAD$, $p_i = p$, $\rho_{i,i} = 1$ for each $i = 1, \ldots, N$ and $\rho_{i,j} = \rho$ for $i \neq j$. Calculate the value of D for $N = 100$, and $\rho = 0, 25, 10\%$.*

Appendix A
Technical Appendix

A.1 Integration Theory

Definition A.1 A *decomposition* $\mathcal{Z}$ of the interval $[a, b]$ is understood to be a set $\mathcal{Z} \stackrel{\text{def}}{=} \{t_0, t_1, \ldots, t_n\}$ of points t_j with $a = t_0 < t_1 < \ldots < t_n = b$. Through this the interval $[a, b]$ is decomposed into n sub-intervals $[t_k, t_{k+1}]$, where $k = 0, 1, 2, \ldots, n-1$. $|\mathcal{Z}| \stackrel{\text{def}}{=} \max_k (t_{k+1} - t_k)$, that is, the length of the largest resulting sub-interval and is referred to as the *refinement* of the decomposition $\mathcal{Z}$.

Definition A.2 For a function $w : [a, b] \longrightarrow \mathbb{R}$ and a decomposition $\mathcal{Z} \stackrel{\text{def}}{=} \{t_0, t_1, \ldots, t_n\}$ one defines the *variation* of w with respect to $\mathcal{Z}$ as:

$$V(\mathcal{Z}) \stackrel{\text{def}}{=} \sum_{k=0}^{n-1} |w(t_{k+1}) - w(t_k)|$$

$V \stackrel{\text{def}}{=} \sup_{\mathcal{Z}} V(\mathcal{Z})$ is called the *total variation* of w on $[a, b]$. If $V < \infty$ holds, then w is of *finite variation* on $[a, b]$.

Theorem A.1 *For a function $w : [a, b] \longrightarrow \mathbb{R}$ it holds that:*

1. w is of finite variation when w is monotone,
2. w is of finite variation when w is Lipschitz continuous,
3. w is bounded when w is of finite variation.

Moreover, sums, differences and products of functions of finite variation are themselves of finite variation.

Definition A.3 Given the functions $f, w : [a, b] \rightarrow \mathbb{R}$ and a decomposition $\mathcal{Z}$, choose for $k = 0, 1, \ldots, n - 1$ partitions $\tau_k \in [t_k, t_{k+1}]$ and form:

$$I(\mathcal{Z}, \boldsymbol{\tau}) \stackrel{\text{def}}{=} \sum_{k=0}^{n-1} f(\tau_k) \cdot \{w(t_{k+1}) - w(t_k)\}$$

If $I(\mathcal{Z}, \boldsymbol{\tau})$ converges for $|\mathcal{Z}| \rightarrow 0$ to a limiting value I, which does not depend on the chosen decomposition $\mathcal{Z}$ nor on the choice of the partitions τ_k, then I is called the **Riemann–Stieltjes integral** of f. One writes:

$$I = \int_a^b f(t) dw(t).$$

For $w(t) = t$ we get the **Riemann integral** as a special case of the Stieltjes integrals.

Theorem A.2 (Characteristics of the Riemann–Stieltjes Integral)

1. *If the corresponding integrals on the right-hand side exist, then the* linearity *characteristics hold:*

$$\int_a^b (\alpha \cdot f + \beta \cdot g) \, dw = \alpha \int_a^b f \, dw + \beta \int_a^b g \, dw \quad (\alpha, \beta \in \mathbb{R})$$

$$\int_a^b f \, d(\alpha \cdot w + \beta \cdot v) = \alpha \int_a^b f \, dw + \beta \int_a^b f \, dv \quad (\alpha, \beta \in \mathbb{R})$$

2. *If the integral $\int_a^b fdw$ and the integrals $\int_a^c fdw$ exist, then for $\int_c^b f \, dw, a < c < b$ it holds that:*

$$\int_a^b fdw = \int_a^c fdw + \int_c^b fdw$$

3. *If f is continuous on $[a, b]$ and w is of finite variation, then $\int_a^b fdw$ exists.*
4. *If f is continuous on $[a, b]$ and w is differentiable with a bounded derivative, then it holds that:*

$$\int_a^b f(t)dw(t) = \int_a^b f(t) \cdot w'(t)dt$$

5. Partial integration: *If $\int_a^b fdg$ or $\int_a^b gdf$ exist, so does the other respective integral and it holds that:*

$$\int_a^b fdg + \int_a^b gdf = f(b)g(b) - f(a)g(a)$$

6. *If w is continuous, it holds that $\int_a^b dw(t) = w(b) - w(a)$*

7. *If f is continuous on $[a,b]$ and w is step-wise constant with discontinuity points $\{c_k, k = 1, \ldots, m\}$, then:*

$$\int_a^b f dw = \sum_{k=1}^m f(c_k) \cdot \left\{ w(c_k^+) - w(c_k^-) \right\}$$

where c_k^+ (c_k^-) is the right (left) continuous limit and $w(c_k^+) - w(c_k^-)$ is the step height of w on $\{c_k\}$.

Theorem A.3 (Radon–Nikodym) *Let λ and μ be positive measures on $(\Omega, \mathcal{F})$ with*

1. *$0 < \mu(\Omega) < \infty$ and $0 < \lambda(\Omega) < \infty$*
2. *λ is absolutely continuous with respect to μ, i.e. from $\mu(A) = 0$ it follows that $\lambda(A) = 0$ for all $A \in \mathcal{F}$ (written: $\lambda \ll \mu$).*

Then a non-negative $\mathcal{F}$-measurable function h exists on Ω, such that:

$$\forall A \in \mathcal{F} : \lambda(A) = \int_A h d\mu;$$

In particular, for all measurable functions f it holds that:

$$\int f d\lambda = \int f \cdot h \, d\mu.$$

Remark A.1 One often uses the abbreviation $\lambda = h \cdot \mu$ in the Radon–Nikodym theorem and refers to h as the *density* of λ with respect to μ. Due to its construction h is also referred to as the *Radon–Nikodym derivative*. In this case one writes $h = \frac{d\lambda}{d\mu}$. One calls λ and μ equivalent measures if $\lambda \ll \mu$ and $\mu \ll \lambda$.

An important tool in stochastic analysis is the transformation of measure, which is illustrated in the following example.

Example A.1 Let $Z_1, \ldots, Z_n$ be independent random variables with standard normal distributions on the measurable space $(\Omega, \mathcal{F}, P)$ and $\mu_1, \ldots, \mu_n \in \mathbb{R}$. Then by

$$Q(d\omega) \stackrel{\text{def}}{=} \xi(\omega) \cdot P(d\omega) \quad \text{with} \quad \xi(\omega) \stackrel{\text{def}}{=} \exp\{\sum_{i=1}^n \mu_i Z_i(\omega) - \frac{1}{2}\mu_i^2\}$$

an equivalent probability measure Q for P is defined. For the distribution of the $Z_1, \ldots, Z_n$ under the new measure Q it holds that:

$$Q(Z_1 \in dz_1, \ldots, Z_n \in dz_n)$$

$$= \exp\{\sum_{i=1}^n (\mu_i z_i - \frac{1}{2}\mu_i^2)\} \cdot P(Z_1 \in dz_1, \ldots, Z_n \in dz_n)$$

$$= \exp\{\sum_{i=1}^{n}(\mu_i z_i - \frac{1}{2}\mu_i^2)\} \cdot (2\pi)^{-\frac{n}{2}} \exp\{-\frac{1}{2}\sum_{i=1}^{n} z_i^2\} dz_1 \ldots dz_n$$

$$= (2\pi)^{-\frac{n}{2}} \exp\{-\frac{1}{2}\sum_{i=1}^{n}(z_i - \mu_i)^2\} dz_1 \ldots dz_n,$$

in other words $Z_1, \ldots, Z_n$ are, with respect to Q, independent and normally distributed with expectations $\mathsf{E}_Q(Z_i) = \mu_i$ and $\mathsf{E}_Q[(Z_i - \mu_i)^2] = 1$. Thus the random variables $\widetilde{Z_i} \overset{\text{def}}{=} Z_i - \mu_i$ are independent random variables with standard normal distributions on the measurable space $(\Omega, \mathcal{F}, Q)$.

Going from P to Q by multiplying with ξ changes the expectations of the normally distributed random variables, but the volatility structure remains notably unaffected.

The following Girsanov theorem generalizes this method for the continuous case, that is, it constructs for a given P-Brownian motion W_t an equivalent measure Q and an appropriately adjusted process W_t^*, so that it represents a Q-Brownian motion. In doing so the ("arbitrarily" given) expectation μ_i is replaced by an ("arbitrarily" given) drift, that is, a stochastic process X_t.

Theorem A.4 (Girsanov) *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space,* W_t *a Brownian motion with respect to* P, $\mathcal{F}_t$ *a filtration in* $\mathcal{F}$ *and* X_t *an adapted stochastic process. Then*

$$\xi_t \overset{\text{def}}{=} \exp(\int_0^t X_u dW_u - \frac{1}{2}\int_0^t X_u^2 du)$$

defines a martingale with respect to P *and* $\mathcal{F}_t$. *The process* W_t^* *defined by*

$$W_t^* \overset{\text{def}}{=} W_t - \int_0^t X_u du$$

is a Wiener process with respect to the filtration $\mathcal{F}_t$ *and*

$$Q \overset{\text{def}}{=} \xi_T \cdot P \tag{A.1}$$

is a P *equivalent probability measure* Q.

The Girsanov theorem thus shows that for a P-Brownian motion W_t an equivalent probability measure Q can be found such that W_t^*, as a Q-Brownian motion at time t, contains the drift X_t. In doing so (A.1) means that: $\int_\Omega \mathbf{1}(\omega \in A) dQ(\omega) = Q(A) \overset{\text{def}}{=} \int_\Omega \mathbf{1}(\omega \in A)\xi_T dP(\omega) = \mathsf{E}_P[\mathbf{1}(\omega \in A)\xi_T]$ for all $A \in \mathcal{F}$.

Remark A.2 With the relationships mentioned above ξ_t is by all means a martingale with respect to P and $\mathcal{F}_t$ when the so-called *Novikov Condition*

$$\mathsf{E}_\mathsf{P}\left[\exp\left(\int_0^t X_u^2 du\right)\right] < \infty \quad \text{for all } t \in [0, T]$$

is met, that is, when X_t does not vary too much.

Another important tool used to derive the Black–Scholes formula by means of martingale theory is the martingale representation theory. It states that every Q-martingale under certain assumptions can be represented by a predetermined Q-martingale by means of a square-integrable process.

Theorem A.5 (Martingale Representation Theorem) *Let M_t be a martingale with respect to the probability measure Q and the filtration $\mathcal{F}_t$, for which the volatility process $\sigma_t \neq 0$ a.s. for all $t \in [0, T]$, where $\sigma_t^2 = \mathsf{E}_\mathsf{Q}[M_t^2|\mathcal{F}_t]$. If N_t is another martingale with respect to Q and $\mathcal{F}_t$, then (uniquely defined) on $\mathcal{F}_t$ an adapted stochastic process H_t exists with $\int_0^T H_t^2 \sigma_t^2 dt < \infty$ with:*

$$N_t = N_0 + \int_0^t H_s dM_s.$$

Example A.2 It is easy to show that the standard Wiener process W_t with respect to the probability measure P is a martingale with respect to P and its corresponding filtration $\mathcal{F}_t$. If X_t is another martingale with respect to P and $\mathcal{F}_t$, then according to the previous theorem a $\mathcal{F}_t$ adapted stochastic process H_t exists, so that

$$X_t = X_0 + \int_0^t H_s dW_s.$$

Remark A.3 Writing the last expression in terms of derivatives:

$$dX_t = H_t dW_t.$$

The example shows once again that a martingale cannot possess a drift.

A.2 Portfolio Strategies

The portfolio of an investor at time t, i.e., the market value of the single equities (contracts) in his portfolio at time t, is dependent on the development of the price $\{S_s; s < t\}$, $S_s = (S_s^1, \ldots, S_s^d)^\top \in \mathbb{R}^d$ up to time t, that is, on the information that is available at that particular time point. Given this, it is obvious that his strategy, i.e., the development of his portfolio's value over time, should be modelled as a

$\mathcal{F}_t$ adapted d-dimensional stochastic process ϕ_t. In doing so $\phi_t^i(\omega)$ represents how much in state ω of the security i is in his portfolio at time t, where negative values indicate a short sell of the corresponding contract.

Definition A.4 Assume the following market model: $\mathcal{M} = (\Omega, \mathcal{F}, P, \mathcal{F}_t, S_t)$. A d-dimensional stochastic process ϕ_t adapted on the filtration $\mathcal{F}_t$ is called a portfolio strategy. The stochastic process $V(\phi_t)$ with $V(\phi_t) \overset{\text{def}}{=} \sum_{i=1}^{d} \phi_t^i S_t^i$ is called the value of the strategy ϕ.

Example A.3 In the Black–Scholes model two financial instruments are traded on the market: a risky security S (stock) and a riskless security B (zero bond). As in Chap. 5, the stock price S_t is assumed to follow a geometric Brownian motion, so that the following stochastic differential equation is satisfied:

$$dS_t = S_t(\mu dt + \sigma dW_t) \qquad\qquad (A.2)$$

The price of the zero bond B_t satisfies the differential equation:

$$dB_t = rB_t dt$$

with a constant r. Without loss of generality it can be assumed that $B_0 = 1$, which leads to $B_t = exp(rt)$.

The corresponding market model is thus $\mathcal{M}_{BS} = (\Omega, \mathcal{F}, P, \mathcal{F}_t, S_t)$, where $S_t \overset{\text{def}}{=} (S_t, B_t)^\top \in \mathbb{R}^2$.

The two-dimensional stochastic process $\phi_t = (a_t, b_t)^\top$ now describes a portfolio strategy in which $a_t(\omega)$ gives the number of stocks and $b_t(\omega)$ gives the number of bonds in the portfolio at time t in state ω. The value of the portfolio at time t is then a random variable

$$V(\phi_t) = a_t S_t + b_t B_t.$$

A particularly important portfolio strategy is that once it is implemented it does not result in any cash flows over time, i.e., when the portfolio is re-balanced no payments are necessary. This means that eventual income (through selling securities, receiving dividends, etc.) is exactly offset by required payments (through buying additional securities, transaction costs, etc.) This is referred to as a self-financing strategy. For an outside observer the change in value of the portfolio only occurs as the price of the participating securities changes.

Definition A.5 Let $\mathcal{M} = (\Omega, \mathcal{F}, P, \mathcal{F}_t, S_t)$ be a market model and ϕ a portfolio strategy with the value $V(\phi_t)$. Then ϕ is called

1. *self-financing*, when $dV(\phi_t) = \sum_{i=1}^{d} \phi_t^i dS_t^i$ holds (P-a.s.),
2. *admissible*, when $V(\phi_t) \geq 0$ holds (P-a.s.).

Below the Black–Scholes model will be considered. The subsequent specification shows that arbitrage is not possible in such a market: There is no admissible self-financing strategy with a starting value of $V(\boldsymbol{\phi}_0) = 0$, whose end value $V(\boldsymbol{\phi}_T)$ is positive with a positive probability.

Lemma A.1 *In the Black–Scholes model* $\mathcal{M}_{BS} = (\Omega, \mathcal{F}, P, \mathcal{F}_t, S_t)$, $S_t = (S_t, B_t)^\top$, *the portfolio strategy* $\phi_t = (a_t, b_t)^\top$ *is exactly self-financing when the discounted process* $\tilde{V}_t$ *with* $\tilde{V}_t = e^{-rt}V_t$ *satisfies the stochastic differential equation*

$$d\tilde{V}_t = a_t d\tilde{S}_t,$$

where $\tilde{S}_t = e^{-rt}S_t$ *describes the discounted stock price.*

The explicit specification of the corresponding strategy can be left out when it is clear from the context and we write $V_t = V(\phi_t)$. With the help of the Girsanov theorem a P equivalent measure Q can be constructed, under which the process of the discounted stock prices is a martingale. Using (A.2) one obtains

$$d\tilde{S}_t = \tilde{S}_t\{(\mu - r)dt + \sigma dW_t\}. \tag{A.3}$$

By setting

$$X_t \stackrel{\text{def}}{=} -\frac{\mu - r}{\sigma}$$

the Novikov condition (see Remark A.2) is obviously fulfilled. Therefore, for Q with

$$\frac{dQ}{dP} = \xi_T = \exp\left(\int_0^T X_u dW_u - \frac{1}{2}\int_0^T X_u^2 du\right)$$

$$= \exp\left\{-\frac{\mu - r}{\sigma}W_T - \frac{1}{2}\left(\frac{\mu - r}{\sigma}\right)^2 T\right\}$$

$W_t^* \stackrel{\text{def}}{=} W_t + \frac{\mu - r}{\sigma}t$ is a Q-Brownian Motion according to the Girsanov theorem. Because of (A.3) and using the definition of W_t^* it holds that

$$d\tilde{S}_t = \tilde{S}_t\sigma dW_t^*. \tag{A.4}$$

According to Itô's lemma this becomes

$$\tilde{S}_t = \tilde{S}_0 \exp\left(\int_0^t \sigma dW_u^* - \frac{1}{2}\int_0^t \sigma^2 du\right)$$

and solves the stochastic differential equation (A.4). Since σ is constant, for all t the Novikov condition holds

$$\mathsf{E}\left[\exp\left(\int_0^t \sigma^2 du\right)\right] < \infty.$$

According to Remark A.2

$$\exp\left(\int_0^t \sigma dW_u^* - \frac{1}{2}\int_0^t \sigma^2 du\right),$$

that is $\tilde{S}_t$, is also a Q-martingale.

Q represents with respect to $\tilde{S}_t$ a P equivalent martingale measure. It can be shown that given this form, it can be uniquely determined.

From the Definition of W_t^* and with the help of (A.2) one obtains

$$dS_t = S_t(rdt + \sigma dW_t^*),$$

i.e., under the measure Q the expected value of the risky securities is equivalent to the certain value of the riskless bonds. Because of this the martingale measure Q is also called the *risk neutral* probability measure and contrary to this P is called the *objective* or *physical* probability measure of the Black–Scholes markets.

As a result of the Q-martingale properties of $\tilde{S}_t$, due to Lemma A.1, the discounted value of a self-financing strategy $\tilde{V}_t$ is itself a Q-martingale. Consequently it holds that: If the starting value of an admissible self-financing strategy is equal to zero, then its value at all later time points t must also be equal to zero. Thus in using an admissible self-financing strategy, there is no riskless profit to be made: The Black–Scholes market is free of arbitrage.

The following theorem represents the most important tool used to value European options with the help of the Black–Scholes model. It secures the existence of an admissible self-financing strategy that duplicates the option, thus the value of which can be calculated using martingale theory.

Theorem A.6 *Assume that the Black–Scholes model $\mathcal{M}_{BS}$ is given. The function X describes the value of a European option at the time to maturity T and is Q-integrable.*

a) *Then an admissible self-financing strategy $(a_t, b_t)^\top$ exists, which duplicates X and whose value V_t for all t is given by*

$$V_t = \mathsf{E}_Q[e^{-r(T-t)}X \mid \mathcal{F}_t]. \tag{A.5}$$

b) *If the value V_t in a) is dependent on t and S_t and is written as a function $V_t = F(t, S_t)$ with a smooth function F, then it holds for the corresponding strategy that*

$$a_t = \frac{\partial F(t, S_t)}{\partial S_t}.$$

Proof 1. One defines V_t by (A.5), where the function defined follows from the Q-integrability of X. Due to

$$\tilde{V}_t = e^{-rt}V_t = \mathsf{E}_Q[e^{-rT}X \mid \mathcal{F}_t]$$

one identifies $\tilde{V}_t$ as Q-martingale. One should notice that $e^{-rT}X$, exactly like X, is only dependent on the state at date T and thus it can be classified as a random variable on $(\Omega, \mathcal{F}_t, Q)$.

$\mathcal{F}_t$ represents, at the same time, the natural filtration for the process W^*, which, as was seen above, is also a Q-martingale. Therefore, according to Theorem A.5 using the martingale representation, a process H_t exists, adapted on $\mathcal{F}_t$ with $\int_0^T H_t^2\sigma^2 dt < \infty$ Q-almost sure, so that for all t it holds that:

$$\tilde{V}_t = \tilde{V}_0 + \int_0^t H_s dW_s^* = V_0 + \int_0^t H_s dW_s^*.$$

Thus one sets:

$$a_t \stackrel{\text{def}}{=} \frac{H_t}{\sigma \cdot \tilde{S}_t} \quad , \quad b_t \stackrel{\text{def}}{=} \tilde{V}_t - a_t \tilde{S}_t.$$

Then after a simple calculation it holds that:

$$a_t S_t + b_t B_t = V_t$$

and $(a_t, b_t)^\top$ is a X duplicating strategy. Furthermore, with (A.4) it holds for all t:

$$a_t d\tilde{S}_t = a_t \tilde{S}_t \sigma dW_t^* = H_t dW_t^* = d\tilde{V}_t,$$

i.e. $(a_t, b_t)^\top$ is according to Lemma A.1 self-financing. Due to the non-negativity of X and the definition of V_t, $(a_t, b_t)^\top$ is also admissible.

2. For $V_t = F(t, S_t)$ it holds using Itô's lemma:

$$\begin{aligned}
d\tilde{V}_t &= d\{e^{-rt}F(t, S_t)\} \\
&= \frac{\partial\{e^{-rt}F(t, S_t)\}}{\partial S_t}dS_t + A(t, S_t)dt \\
&= \frac{\partial F(t, S_t)}{\partial S_t}e^{-rt}S_t(rdt + \sigma dW_t^*) + A(t, S_t)dt \\
&= \frac{\partial F(t, S_t)}{\partial S_t}\tilde{S}_t \sigma dW_t^* + \tilde{A}(t, S_t)dt \\
&= \frac{\partial F(t, S_t)}{\partial S_t}d\tilde{S}_t + \tilde{A}(t, S_t)dt.
\end{aligned}$$

Since not only $\tilde{V}_t$ but also $\tilde{S}_t$ are Q-martingales, the drift term $\tilde{A}(t, S_t)$ must disappear. According to part a) of the theorem the corresponding strategy is self-financing and thus using Lemma A.1 the claim follows.

<div align="right">□</div>

Remark A.4 With the relationships of the preceding theorems, V_t is called the *fair price* for option X at date t, because at this price, according to the previous arguments, there is no arbitrage possible for either the buyer or the seller of the option. Equation (A.5) is called the risk neutral valuation formula, since it gives the fair price of the option as the (conditional) expectation of the (discounted) option value at maturity with respect to the risk neutral measure of the Black–Scholes model.

The result obtained from the last theorem has already been formulated in Chap. 6 as Eq. (6.24).

Corollary A.1 *The relationships of the preceding theorems hold. If the value X of the European option at date T is a function $X = f(S_T)$ dependent on the stock price S_T, then it holds that $V_t = F(t, S_t)$, where F for $x \in [0, \infty[$ and $t \in [0, T]$ is defined by:*

$$F(t, x) = e^{-r(T-t)} \int_{-\infty}^{+\infty} f \left\{ x e^{(r-\frac{\sigma^2}{2})(T-t)+\sigma y \sqrt{T-t}} \right\} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy . \tag{A.6}$$

Proof With respect to Q, S_t contains the drift r and thus it holds that

$$S_t = S_0 \exp\{(r - \frac{\sigma^2}{2})t + \sigma W_t^*\}.$$

Thus S_T can be written in the following form:

$$S_T = S_t(S_T S_t^{-1}) = S_t \exp\{(r - \frac{\sigma^2}{2})(T - t) + \sigma(W_T^* - W_t^*)\}.$$

Since S_t is measurable with respect to $\mathcal{F}_t$ and $W_T^* - W_t^*$ is independent of $\mathcal{F}_t$, one obtains

$$\begin{aligned} V_t &= \mathsf{E}_Q[e^{-r(T-t)} f(S_T) \mid \mathcal{F}_t] \\ &= \mathsf{E}_Q\left[e^{-r(T-t)} f(S_t e^{(r-\frac{\sigma^2}{2})(T-t)+\sigma(W_T^*-W_t^*)}) \mid \mathcal{F}_t \right] \\ &= \mathsf{E}_Q\left[e^{-r(T-t)} f(x e^{(r-\frac{\sigma^2}{2})(T-t)+\sigma(W_T^*-W_t^*)}) \right]_{x=S_t} \end{aligned}$$

From this it can be calculated that $V_t = F(t, S_t)$.

<div align="right">□</div>

Example A.4 Consider a European call $X = \max\{0, S_T - K\}$. Using (A.6) the value at date t is exactly the value given by the Black–Scholes formula in Chap. 6.

$$C(t, S_t) \stackrel{\text{def}}{=} V_t = S_T \Phi(d_1) - K e^{-r(T-t)} \Phi(d_2)$$

with

$$d_1 \stackrel{\text{def}}{=} \frac{\log(\frac{S_T}{K}) + (r + \frac{\sigma^2}{2})(T - t)}{\sigma \sqrt{T - t}}, \quad d_2 \stackrel{\text{def}}{=} \frac{\log(\frac{S_T}{K}) + (r - \frac{\sigma^2}{2})(T - t)}{\sigma \sqrt{T - t}}.$$

Symbols and Notations

Basics

$x \stackrel{\text{def}}{=} \dots$	x is defined as ...
$\overline{\mathbb{R}}$	$\mathbb{R} \cup \{\infty, \infty\}$
X, Y	random variables or vectors
$X_1, X_2, \dots, X_p$	random variables
$X = (X_1, \dots, X_p)^\top$	random vector
A, B	matrices
$A^\top$	transpose of matrix A
Σ	covariance matrix
1_n	vector of ones $(\underbrace{1, \dots, 1}_{n\text{-times}})^\top$
0_n	vector of zeros $(\underbrace{0, \dots, 0}_{n\text{-times}})^\top$
$\boldsymbol{I}(.)$	indicator function, i.e. for a set M is $\boldsymbol{I} = 1$ on M, $\boldsymbol{I} = 0$ otherwise
$[x]$	integer part of x
$(X)^+$	$\lvert X \rvert * \boldsymbol{I}(X > 0)$
$\Rightarrow$	implication
$\Leftrightarrow$	equivalence
$\approx$	approximately equal
iff	if and only if, equivalence
$\alpha_n = \mathcal{O}(\beta_n)$	iff $\frac{\alpha_n}{\beta_n} \longrightarrow$ constant, as $n \longrightarrow \infty$
$\alpha_n = o(\beta_n)$	iff $\frac{\alpha_n}{\beta_n} \longrightarrow 0$, as $n \longrightarrow \infty$

Probability Calculus

P[A] or P(A)	probability of a set A
$X \sim \cdot$	X has distribution $\cdot$
$a.s.$	almost surely, i.e. $X > M$ $a.s.$ *iff* $P(X > M) = 1$
$\xrightarrow{\mathcal{L}}$	convergence in distribution
$\xrightarrow{P}$	convergence in probability
$\{A_n\}_{n=1}^{\infty}, \{B_n\}_{n=1}^{\infty}$	sequences of random variables
$A_n = \mathcal{O}_p(B_n)$	iff $\forall \varepsilon > 0 \; \exists M, \; \exists N$ such that P[$|A_n/B_n| > M$] $< \varepsilon, \; \forall n > N$
$A_n = o_p(B_n)$	iff $\forall \varepsilon > 0 \; : \; \lim_{n \to \infty} $P[$|A_n/B_n| > \varepsilon$] $= 0$
$\mathcal{F}_t$	the information set generated by all information available at time t
W_t	Wiener process
SDE	stochastic differential equation
CLT	Central Limit Theorem

Characteristics of Distribution

$f(x)$	pdf or density of X
$f(x, y)$	joint density of X and Y
$f_X(x), f_Y(y)$	marginal densities of X and Y
$f_{X_1}(x_1), \ldots, f_{X_p}(x_p)$	marginal densities of $X_1, \ldots, X_p$
$\hat{f}_h(x)$	histogram or kernel estimator of $f(x)$
$F(x)$	cdf or distribution function of X
$F(x, y)$	joint distribution function of X and Y
$F_X(x), F_Y(y)$	marginal distribution functions of X and Y
$F_{X_1}(x_1), \ldots, F_{X_p}(x_p)$	marginal distribution functions of $X_1, \ldots, X_p$
$f_{Y|X=x}(y)$	conditional density of Y given $X = x$
$\varphi_X(t)$	characteristic function of X
m_k	kth moment of X
κ_j	cumulants or semi-invariants of X

Moments

$\mathsf{E}\,X, \mathsf{E}\,Y$	mean values of random variables or vectors X and Y
$\mathsf{E}(Y\mid X = x)$	conditional expectation of random variable or vector Y given $X = x$
$\mu_{Y\mid X}$	conditional expectation of Y given X
$\mathsf{Var}(Y\mid X = x)$	conditional variance of Y given $X = x$
$\sigma^2_{Y\mid X}$	conditional variance of Y given X
$\sigma_{XY} = \mathsf{Cov}(X, Y)$	covariance between random variables X and Y
$\sigma_{XX} = \mathsf{Var}(X)$	variance of random variable X
$\rho_{XY} = \dfrac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)\,\mathsf{Var}(Y)}}$	correlation between random variables X and Y
$\Sigma_{XY} = \mathsf{Cov}(X, Y)$	covariance between random vector X and Y, i.e., $\mathsf{Cov}(X, Y) = E(X = EX)(Y - EY)^{\top}$
$\Sigma_{XX} = \mathsf{Var}(X)$	covariance matrix of the random vector X

Samples

x, y	observations of X and Y
$x_1, \ldots, x_n = \{x_i\}_{i=1}^n$	sample of n observations of X
$\mathcal{X} = \{x_{ij}\}_{i=1,\ldots,n;\, j=1,\ldots,p}$	$(n \times p)$ data matrix of observations of $X_1, \ldots, X_p$ or of $X = (X_1, \ldots, X_p)^{\top}$
$x_{(1)}, \ldots, x_{(n)}$	the order statistic of $x_1, \ldots, x_n$

Empirical Moments

$\overline{x} = \dfrac{1}{n}\sum_{i=1}^n x_i$	average of X sampled by $\{x_i\}_{i=1,\ldots,n}$
$s_{XY} = \dfrac{1}{n}\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$	empirical covariance of random variables X and Y sampled by $\{x_i\}_{i=1,\ldots,n}$ and $\{y_i\}_{i=1,\ldots,n}$
$s_{XX} = \dfrac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2$	empirical variance of random variable X sampled by $\{x_i\}_{i=1,\ldots,n}$
$r_{XY} = \dfrac{s_{XY}}{\sqrt{s_{XX}s_{YY}}}$	empirical correlation of X and Y
$\mathcal{S} = \{s_{X_i X_j}\}$	empirical covariance matrix of $X_1, \ldots, X_p$ or of the random vector $X = (X_1, \ldots, X_p)^{\top}$
$\mathcal{R} = \{r_{X_i X_j}\}$	empirical correlation matrix of $X_1, \ldots, X_p$ or of the random vector $X = (X_1, \ldots, X_p)^{\top}$

Distributions

$\varphi(x)$	density of the standard normal distribution
$\Phi(x)$	distribution function of the standard normal distribution
$\mathrm{N}(0, 1)$	standard normal or Gaussian distribution
$\mathrm{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathrm{N}_p(\mu, \Sigma)$	p-dimensional normal distribution with mean μ and covariance matrix Σ
$B(n, p)$	binomial distribution with parameters n and p
$Poiss(\lambda)$	Poisson (Bortkiewicz) distribution with parameter λ
$Exp(\lambda)$	exponential distribution with parameter λ
$lognormal(\mu, \sigma^2)$	lognormal distribution with mean μ and variance σ^2
χ^2_p	χ^2 distribution with p degrees of freedom
$\chi^2_{1-\alpha;p}$	$1 - \alpha$ quantile of the χ^2 distribution with p degrees of freedom
t_n	t-distribution with n degrees of freedom
$t_{1-\alpha/2;n}$	$1 - \alpha/2$ quantile of the t-distribution with n degrees of freedom
$F_{n,m}$	F-distribution with n and m degrees of freedom
$F_{1-\alpha;n,m}$	$1 - \alpha$ quantile of the F-distribution with n and m degrees of freedom

References

Abberger, K. (1997). Quantile smoothing in financial time series. *Statistical Papers, 38*, 125–148.

Ait-Sahalia, Y., & Duarte, J. (2003). Nonparametric option pricing under shape restrictions. *Journal of Econometrics, 116*, 9–47.

Anders, U. (1997). *Statistische neuronale Netze*. München: Vahlen.

Ango Nze, P. (1992). *Critères d'ergodicité de quelques modèles à représentation markovienne*. Technical Report 315, ser. 1. Paris: C. R. Acad. Sci.

Applebaum, D. (2004). *Lévy processes and stochastic calculus*. Cambridge: Cambridge University Press.

Artzner, P., Delbaen, F., Eber, J., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance, 9*(3), 203–228.

Artzner, P., Delbaen, F., Eber, J.-M., & Heath, D. (1997). Thinking coherently. *Risk, 10*, 68–71.

Baillie, R. T. (1996). Long memory and fractional integration in econometrics. *Journal of Econometrics, 73*, 5–59.

Bauwens, L., Laurent, S., & Rombouts, J. (2005). Multivariate garch models: A survey. *Journal of Applied Econometrics, 21*, 79–109.

Baxter, M., & Rennie, A. (1996). *Financial calculus: An introduction to derivative pricing*. Cambridge: Cambridge University Press.

Beisser, J. (2001). *Topics in Finance - A conditional expectation approach to value Asian, Basket and Spread options*. Ph.D. thesis, Johannes Gutenberg University of Mainz, Mainz.

Beran, J. (1994). *Statistics for long-memory processes. Monographs of statistics and applied probability* (Vol. 61). London: Chapman and Hall.

Bertoin, J. (1996). *Lévy processes*. Cambridge: Cambridge University Press.

Black, F. (1975). The pricing of commodity contracts. *Journal of Financial Economics, 3*, 167–179.

Black, F. (1976). Studies in stock price volatility changes. In *Proceedings of the 1976 Meeting of the Business and Economic Statistics Section* (pp. 177–181). Washington, DC: American Statistical Association.

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *Journal of Political Economy, 81*, 637–654.

Bluhm, C., Overbeck, L., & Wagner, C. (2010). *An introduction to credit risk modeling* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Bol, G., Nakhaeizadeh, G., & Vollmer, K.-H. (1996). *Finanzmarktanalyse und -prognose mit innovativen quantitativen Verfahren*. Heidelberg: Physica.

Bollerslev, T., Engle, R., & Wooldridge, J. (1988). A capital asset pricing model with time varying covariances. *Journal of Political Economy, 96*, 116–131.

Bollerslev, T. P. (1986). Generalized autoregressive conditional heteroscedasticity. *Journal of Econometrics, 31*, 307–327.

Bollerslev, T. P. (1990). Modelling the coherence in short-run nominal exchange rates: A multivariate generalized arch model. *Review of Economics and Statistics, 72*, 498–505.

Bollerslev, T. P., & Wooldridge, J. M. (1992). Quasi maximum likelihood estimation of dynamic models with time-varying covariances. *Econometric Reviews, 11*, 143–172.

Bossaerts, P., & Hillion, P. (1993). Test of a general equilibrium stock option pricing model. *Mathematical Finance, 3*, 311–347.

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., & Roncalli, T. (2000). *Copulas for finance. A reading guide and some applications*. Groupe de Recherche Opérationnelle Crédit Lyonnais.

Box, G. E. P., & Jenkins, G. M. (1976). *Time series analysis: Forecasting and control*. San Francisco: Holden-Day.

Brace, A., Gatarek, D., & Musiela, M. (1997). The market model of interest rate dynamics. *Mathematical Finance, 7*, 127–147.

Brigo, D., & Mercurio, F. (2001). *Interest rate models: Theory and practice*. Berlin: Springer.

Briys, E., Bellalah, M., Mai, H., & de Varenne, F. (1998). *Options, futures and exotic derivatives*. Chichester: Wiley.

Brockwell, P., & Davis, R. (1991). *Time series: Theory and methods*. New York: Springer.

Brockwell, P. J., Chadraa, E., & Lindner, A. (2006). Continuous time GARCH processes. *Annals of Applied Probability, 16*, 790–826.

Brzezniak, Z., & Zastawniak, T. (1999). *Basic stochastic processes*. London: Springer.

Buchmann, B., & Müller, G. (2008). *Limit experiments of GARCH*. Technical Report, Monash University and Technische Universität München.

Campbell, J., Lo, A., & MacKinlay, A. (1997). *The econometrics of financial markets*. Princeton: Princeton University Press.

Carroll, R. J., Härdle, W., & Mammen, E. (2002). Estimation in an additive model when the components are linked parametrically. *Econometric Theory, 18*, 886–912.

Chan, K., & Tong, H. (1985). On the use of deterministic Lyapunov functions for the ergodicity of stochastic difference equations. *Advanced Applied Probability, 17*, 666–678.

Chen, R., & Tsay, R. S. (1993a). Functional-coefficient autoregressive models. *Journal of the American Statistical Association, 88*, 298–308.

Chen, R., & Tsay, R. S. (1993b). Nonlinear additive ARX models. *Journal of the American Statistical Association, 88*, 955–967.

Chen, X., & Fan, Y. (2006a). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification. *Journal of Econometrics, 135*, 125–154.

Chen, X., & Fan, Y. (2006b). Estimation of copula-based semiparametric time series models. *Journal of Econometrics, 130*, 307–335.

Chen, X., Fan, Y., & Tsyrennikov, V. (2004). *Efficient estimation of semiparametric multivariate copula models*. Working Paper.

Chung, C. F. (2001). *Estimating the fractionally integrated garch model*. Working Paper, National Taïwan University.

Cizek, P., Härdle, W., & Weron, R. (2011). *Statistical tools in finance and insurance* (2nd ed.). Berlin: Springer.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika, 65*, 141–151.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829–836.

Collomb, G. (1984). Propriétés de convergence presque complète du prédicteur à noyau. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete, 66*, 441–460.

Comte, F., & Lieberman, O. (2003). Asymptotic theory for multivariate GARCH processes. *Journal of Multivariate Analysis, 84*, 61–84.

Copeland, T., & Weston, J. (1992). *Financial theory and corporate policy*. Reading: Addison-Wesley.

Cox, J., Ross, S., & Rubinstein, M. (1979). Option pricing: A simplified approach. *Journal of Financial Economics, 7*, 229–263.

Cox, J., & Rubinstein, M. (1985). *Options markets*. Englewood Cliffs: Prentice-Hall.

Cox, J. C., Ingersoll, Jr., J. E., & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica, 53*, 385–407.

Cox, J. C., & Ross, S. A. (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Economics, 3*, 145–166.

Crack, T. F. (2004). *Basic Black-Scholes: Option pricing and trading*, USA: Timothy Crack.

Czado, C., & Haug, S. (2007). *Quasi maximum likelihood estimation and prediction in the compound poisson ECOGARCH(1,1) model*. Technical Report, Technische Universität München.

Das, S. (1997). *Risk management and financial derivatives*. New York: McGraw-Hill.

Davidson, J. (2004). Moment and memory properties of linear conditional heteroscedasticity models, and a new model. *Journal of Business and Economic Statistics, 22*, 16–19.

Delbaen, F., & Schachermayer, W. (1994). A general version of the fundamental theorem of asset pricing. *Mathematische Annalen, 300*, 463–520.

Deutsch, H., & Eller, R. (1999). *Derivatives and internal models*. Houndmill: Macmillan Press.

Devroye, L. (1986). *Non-uniform random variate generation*. New York: Springer.

Dewynne, J., Howison, S., & Wilmott, P. (1993). *Mathematical models and computation*. Oxford: Oxford University Press.

Dickey, D., & Fuller, W. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association, 74*, 427–431.

Diebold, F. X., & Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics, 105*, 131–159.

Diebolt, J., & Guégan, D. (1990). *Probabilistic properties of the general nonlinear autoregressive process of order one*. Technical Report 128, Université de Paris VI, Paris.

Ding, Z., Granger, C. W. J., & Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance, 1*, 83–106.

Doukhan, P., & Ghindès, M. (1980). Estimation dans le processus $x_{n+1} = f(x_n) + \varepsilon_{n+1}$. *Les Comptes Rendus de l'Académie des sciences, 297*, 61–64.

Doukhan, P., & Ghindès, M. (1981). Processus autorégressifs non-linéaires. *Les Comptes Rendus de l'Académie des sciences, 290*, 921–923.

Duan, J.-C. (1995). The GARCH option pricing model. *Mathematical Finance, 5*, 13–32.

Duan, J.-C. (1997). Augmented GARCH(p,q) process and its diffusion limit. *Journal of Econometrics, 79*, 97–127.

Duffie, D. (1996). *Dynamic asset pricing theory* (2nd ed.). Princeton: Princeton University Press.

Duffie, D., & Singleton, K. J. (1997). An econometric model of the term structure of interest-rate swap yields. *Journal of Finance, 52*, 1287–1321.

Durrett, R. (1991). *Probability: Theory and examples*. Belmont, CA: Wadsworth and Brooks/Cole.

Egloff, D., Leippold, M., & Vanini, P. (2004). *A simple model of credit contagion*. EFA Maastricht Meetings Paper.

Elton, E., Gruber, M., Brown, S., & Goztmann, W. (2002). *Modern portfolio theory and investment analysis*. Hoboken: Wiley.

Embrechts, P. (2002). *Selfsimilar processes*. Princeton: Princeton University Press.

Embrechts, P., Frey, R., & McNeil, A. (2005). *Quantitative risk management: Concepts, techniques and tools*. Princeton: Princeton University Press.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). *Modelling extremal events*. Berlin: Springer.

Embrechts, P., McNeil, A., & Straumann, D. (1999). Correlation: Pitfalls and alternatives. *Risk, 12*, 69–71.

Embrechts, P., Resnick, S., & Samorodnitsky, G. (1998). Living on the edge. *Risk Magazine, 11*(1), 96–100.

Engle, R., & Kroner, F. (1995). Multivariate simultaneous generalized arch. *Econometric Theory, 11*, 122–150.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of U.K. inflation. *Econometrica, 50*, 987–1008.

Engle, R. F. (2002). Dynamic conditional correlation - A simple class of multivariate garch models. *Journal of Business and Economic Statistics, 20*, 339–350.

Engle, R. F., & Bollerslev, T. P. (1986). Modelling the persistence of conditional variances. *Econometric Reviews, 5*, 1–50, 81–87.

Engle, R. F., & Gonzalez-Rivera, G. (1991). Semiparametric ARCH models. *Journal of Business and Economic Statistics, 9*, 345–360.

Engle, R. F., Lilien, D. M., & Robins, R. P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica, 55*, 391–407.

Engle, R. F., & Ng, V. K. (1993). Measuring and testing the impact of news on volatility. *Journal of Finance, 48*, 1749–1778.

Fahrmeir, L., & Tutz, G. (1994). *Multivariate statistical modelling based on generalized linear models.* Heidelberg: Springer.

Fama, E. F. (1965). The behavior of stock market prices. *Journal of Business, 38*, 34–105.

Fan, J., & Gijbels, I. (1996). *Local polynomial modeling and its application – Theory and methodologies*. New York: Chapman and Hall.

Fan, J., & Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika, 85*, 645–660.

Fan, J., & Yao, Q. (2003). *Nonlinear time series: Nonparametric and parametric methods*. New York: Springer.

Fasen, V., Klüppelberg, C., & Lindner, A. (2005). Extremal behavior of stochastic volatility models. In A. N. Shiryaev, M. R. Grossinho, P. E. Oliviera, & M. L. Esquível (Eds.), *Stochastic finance*. New York: Springer.

Fengler, M., Härdle, W., & Schmidt, P. (2002). An empirical analysis of the common factors governing implied volatility movements measured by the German vdax. *Financial Markets and Portfolio Management, 16*, 16–29.

Fengler, M., Härdle, W., & Villa, C. (2003). The dynamics of implied volatility: A common principal component approach. *Review of Derivative Research, 6*(3), 179–202.

Fisher, R., & Tippett, L. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society, 24*, 180–190.

Föllmer, H., & Schweizer, M. (1991). Hedging of contingent claims under incomplete information. In M. H. A. Davis & R. J. Elliot (Eds.), *Applied stochastic analysis* (pp. 389–414). London: Gordon and Breach.

Föllmer, H., & Sondermann, D. (1991). Hedging of non-redundant contingent claims. In W. Hildenbrand & A. Mas-Colell (Eds.), *Contributions to mathematical economics* (pp. 205–223). Amsterdam: North Holland.

Forsythe, R., Palfrey, T., & Plott, C. (1982). Asset valuation in an experimental market. *Econometrica, 50*, 537–567.

Frank, M. J. (1979). On the simultaneous associativity of $f(x, y)$ and $x + y - f(x, y)$. *Aequationes Mathematicae, 19*, 194–226.

Franke, J. (1999). Nonlinear and nonparametric methods for analyzing financial time series. In P. Kall & H.-J. Luethi (Eds.), *Operation research proceedings 98*. Heidelberg: Springer.

Franke, J. (2000). Portfolio management and market risk quantification using neural networks. In *Statistics and finance: An interface*. London: Imperial College Press.

Franke, J., & Diagne, M. (2006). Estimating market risk with neural networks. *Statistics and Decisions, 24*, 233–253.

Franke, J., Härdle, W., & Kreiss, J. (2003). Nonparametric estimation in a stochastic volatility model. In *Recent advances and trends in nonparametric statistics* (pp. 303–314). Amsterdam: Elsevier.

Franke, J., Härdle, W., & Stahl, G. (Eds.) (2000). *Measuring risk in complex stochastic systems*. Lecture notes in statistics (Vol. 147). New York: Springer.

Franke, J., Kreiss, J., & Mammen, E. (2002). Bootstrap of kernel smoothing in nonlinear time series. *Bernoulli, 8*(1), 1–37.

Franke, J., Kreiss, J., Mammen, E., & Neumann, M. (2003). Properties of the nonparametric autoregressive bootstrap. *Journal of Time Series Analysis, 23*, 555–585.

Franke, J., & Mwita, P. (2003). *Nonparametric estimates for conditional quantiles of time series*. Technical Report, Department of Mathematics, University of Kaiserslautern.

Franke, J., & Neumann, M. (2000). Bootstrapping neural networks. *Neural Computation, 12*, 1929–1949.

Friedmann, R. (1992). *Zur Güte von Einheitswurzeltests*. Technical Report, unveröffentlichtes Manuskript.

Genest, C., & Rivest, L.-P. (1989). A characterization of Gumbel family of extreme value distributions. *Statistics and Probability Letters, 8*, 207–211.

Geweke, J., & Porter-Hudak, S. (1983). The estimation and application of long-memory time series models. *Journal of Time Series Analysis, 4*, 221–238.

Giacomini, E., & Härdle, W. (2005). Value-at-risk calculations with time varying copulae. In *Proceedings 55th International Statistical Institute, Sydney 2005*.

Gikhman, l. I., & Skorokhod, A. V. (1974). *The theory of stochastic processes* I. New York: Springer.

Gilli, M., & Këllezi, E. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics, 27*(1), 1–23.

Giraitis, L., Kokoszka, P., Leipus, R., & Teyssière, G. (2003). Rescaled variance and related tests for long memory in volatility and levels. *Journal of Econometrics, 112*, 265–294.

Glasserman, P. (2004). *Monte Carlo methods in financial engineering*. New York: Springer.

Glosten, L., Jagannathan, R., & Runkle, D. (1993). Relationship between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance, 48*, 1779–1801.

Gouriéroux, C. (1997). *ARCH models and financial applications*. New York: Springer.

Gouriéroux, C., & Jasiak, J. (2002). Nonlinear autocorrelograms: An application to inter-trade durations. *Journal of Time Series Analysis, 23*, 127–154.

Gouriéroux, C., & Monfort, A. (1992). Qualitative threshold ARCH models. *Journal of Econometrics, 52*, 159–199.

Gouriéroux, C., & Monfort, A. (1996). *Time series models*. Paris: Economica.

Gouriéroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica, 52*, 681–700.

Grama, I., & Spokoiny, V. (2003). *Tail index estimation by local exponential modeling*. Technical Report, WIAS, Berlin.

Graumert, U., & Stahl, G. (2001). *Interne Modelle. Handwörterbuch des Bank- und Finanzwesens*. Stuttgart: Schäffer-Poeschel.

Gregory, A. (1989). A nonparametric test for autoregressive conditional heteroscedasticity: A markov chain approach. *Journal of Business and Economic Statistics, 7*, 107–115.

Gumbel, E. J. (1960). Distributions des valeurs extrêmes en plusieurs dimensions. *Publications de l'Institut de statistique de l'Université de Paris, 9*, 171–173.

Hafner, C. (1998). Estimating high frequency foreign exchange rate volatility with nonparametric ARCH models. *Journal of Statistical Planning and Inference, 68*, 247–269.

Hafner, C., & Herwartz, H. (1998). Time-varying market price of risk in CAPM, approaches, empirical evidence and implications. *Finance, 19*, 93–112.

Hafner, C., & Herwartz, H. (2000). Testing linear autoregressive dynamics under heteroskedasticity. *Econometrics Journal, 3*, 177–197.

Hafner, C., & Herwartz, M. (2003). *Analytical quasi maximum likelihood inference in bekk garch models*, Econometric Institute Report 21.

Hamilton, J. D. (1994). *Time series analysis*. Princeton: Princeton University Press.

Härdle, W. (1990). *Applied nonparametric regression*. Cambridge: Cambridge University Press.

Härdle, W., & Hafner, C. (2000). Discrete time option pricing with flexible volatility estimation. *Finance and Stochastics, 4*, 189–207.

Härdle, W., Lütkepohl, H., & Chen, R. (1997). Nonparametric time series analysis. *International Statistical Review, 12*, 153–172.

Härdle, W., Müller, M., Sperlich, S., & Werwatz, A. (2004). *Non- and semiparametric modelling*. Heidelberg: Springer.

Härdle, W., & Simar, L. (2012). *Applied multivariate statistical analysis* (3rd ed.). Heidelberg: Springer.

Härdle, W., & Tsybakov, A. (1997). Local polynomial estimation of the volatility function. *Journal of Econometrics, 81*, 223–242.

Härdle, W., Tsybakov, A., & Yang, L. (1996). Nonparametric vector autoregression. *Journal of Statistical Planning and Inference, 68*, 221–245.

Harrison, M., & Kreps, D. (1979). Martingales and arbitrage in multiperiod securities markets. *Journal of Economic Theory, 20*, 381–408.

Harrison, M., & Pliska, S. (1981). Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes Applications, 11*, 215–260.

Hassler, U. (1994). Einheitswurzeltests – ein Überblick. *Allgemeines Statistisches Archiv, 78*, 207–228.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models. Monographs on statistics and applied probability* (Vol. 43). London: Chapman and Hall.

Hastings, C. (1955). *Approximations for digital computers*. Princeton: Princeton University Press.

Haug, S., & Czado, C. (2007). An exponential continuous time GARCH process. *Journal of Applied Probability, 44*, 960–976.

Haug, S., Klüppelberg, C., Lindner, A., & Zapp, M. (2007). Method of moments estimation in the COGARCH(1,1) model. *Econometrics Journal, 10*, 320–341.

Haug, S., & Stelzer, R. (2010). Multivariate ECOGARCH processes. *Econometric Theory, 27*, 344–371

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. Upper Saddle River: Prentice-Hall.

He, C., & Teräsvirta, T. (1999). Properties of moments of a family of GARCH processes. *Journal of Econometrics, 92*, 173–192.

Heath, D., Jarrow, R., & Morton, A. (1992). Bond pricing and the term structure of interest rates: A new methodology for contingent claims valuation. *Econometrica, 60*, 77–105.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics, 3*, 1163–1174.

Ho, S., & Lee, T. (1986). Term structure movements and pricing interest rate contingent claims. *Journal of Finance, 41*, 1011–1029.

Hoeffding, W. (1940). Masstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin, 5*(3), 179–233.

Hoeffding, W. (1941). Masstabinvariante Korrelationsmasse für diskontinuierliche Verteilungen. *Archiv für die mathematische Wirtschafts- und Sozialforschung, 7*, 49–70.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks, 2*, 359–366.

Hull, J., & White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance, 42*, 281–300.

Hull, J., & White, A. (1990). Pricing interest rate derivatives. *The Review of Financial Studies, 3*, 573–592.

Hull, J., & White, A. (1994). Numerical procedures for implementing term structure models II: Two factor models. *Journal of Derivatives, 2*, 37–47.

Hull, J. C. (2006). *Options, futures and other derivatives*. Upper Saddle River: Prentice Hall.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers, 116*, 770–799.

Hurst, H. E. (1957). A suggested statistical model for some time series that occur in nature. *Nature, 180*, 494.

Hurvich, C., Deo, R., & Brodsky, J. (1998). The mean square error of geweke and porter-hudak's estimator of the memory parameter of a long-memory time series. *Journal of Time Series Analysis, 19*, 19–46.

Ingersoll, Jr., J. E. (1987). *Theory of financial decision making*. Totowa: Rowman & Littlefield.

Jacod, J., & Protter, P. (2000). *Probability essentials*. Heidelberg: Springer.

Jamshidian, F. (1997). Libor and swap market models and measures. *Finance and Stochastics, 1*, 293–330.

Jarchow, H.-J., & Rühmann, P. (1994). *Monetäre Außenwirtschaft I. Monetäre Außenwirtschaftstheorie*. Göttingen: Vandenhoeck & Ruprecht.

Jarrow, R. (1992). *Finance theory* (2nd ed.). Englewood Cliffs: Prentice-Hall.

Jaschke, S., & Küchler, U. (1999). *Coherent risk measures*. Discussion paper, No. 64, SFB 373, Humboldt-Universität zu Berlin.

Jeantheau, T. (1998). Strong consistency of estimators for multivariate arch models. *Econometric Theory, 14*, 70–86.

Jensen, B., & Nielsen, J. (1996). Pricing by no arbitrage. In D. Cox, D. Hinkley, & O. Barndorff-Nielsen (Eds.), *Time series models in econometrics, finance and other fields*. London: Chapman and Hall.

Joe, H. (1996). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In L. Rüschendorf, B. Schweizer, & M. Taylor (Eds.), *Distribution with fixed marginals and related topics. IMS lecture notes – Monograph series*. Hayward: Institute of Mathematical Statistics.

Joe, H. (1997). *Multivariate models and dependence concepts*. London: Chapman and Hall.

Joe, H., & Xu, J. J. (1996). *The estimation method of inference functions for margins for multivariate models*. Technical Report 166, Department of Statistics, University of British Columbia.

Jorion, P. (2000). *Value at risk: The new benchmark for managing financial risk*. New York: McGraw-Hill.

Ju, E. (1992). Pricing asian and basket options via Taylor expansion. *Journal of Computational Finance, 5*, 79–103.

Kallsen, J., & Vesenmayer, B. (2009). COGARCH as a continuous-time limit of GARCH(1,1). *Stochastic Processes and their Applications, 119*, 74–98.

Karatzas, I., & Shreve, S. (1999). *Brownian motion and stochastic calculus*. Heidelberg: Springer.

Katkovnik, V. (1979). Linear and nonlinear methods for nonparametric regression analysis (in Russian). In *Avtomatika i Telemehanika* (pp. 35–46). New York: Book on Demand Ltd.

Katkovnik, V. (1985). *Nonparametric identification and data smoothing*. Moscow: Nauka.

Kim, C. S., & Phillips, P. C. B. (2006). *Log periodogram regression in the nonstationary case*. Cowles Foundation Discussion Paper, Yale University.

Klüppelberg, C., Lindner, A., & Maller, R. A. (2004). A continuous-time GARCH process driven by a lévy process: Stationarity and second-order behaviour. *Journal of Applied Probability, 41*, 601–622.

Klüppelberg, C., Lindner, A., & Maller, R. A. (2006). Continuous time volatility modelling: COGARCH versus ornstein–uhlenbeck models. In Y. Kabanov, R. Lipster, & J. Stoyanov (Eds.), *From stochastic calculus to mathematical finance*. Berlin: Springer.

Knuth, D. (1997). *The art of computer programming*. Boston: Addison-Wesley.

Korn, R. (1999). *Optimal portfolios: Stochastic models for optimal investment and risk management in continuous time*. Singapore: World Scientific.

Korn, R., & Korn, E. (1999). *Optionsbewertung und portfolio-optimierung*. Braunschweig: Vieweg.

Korn, R., & Korn, E. (2001). *Option pricing and portfolio optimization*. Providence: American Mathematical Society.

Krekel, M. de Kock, J. K. R., & Man, T. (2007). An analysis of pricing methods for basket options. *Wilmott, 3*, 82–89.

Krengel, U. (2000). *Einführung in die Wahrscheinlichkeitstheorie und Statistik*. Braunschweig: Vieweg.

Kwiatkowski, D., Phillips, P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics, 54*, 159–178.

Leadbetter, M., Lindgren, G., & Rootzen, H. (1983). *Extremes and related properties of random sequences and processes. Springer series in statistics*. New York: Springer.

Lee, S., & Hansen, B. (1994). Asymptotic theory for the GARCH(1,1) quasi-maximum likelihood estimator. *Econometric Theory, 10*, 29–52.

Lehrbass, F. (2000). A simple approach to country risk. In J. Franke, W. Härdle, & G. Stahl (Eds.), *Measuring risk in complex stochastic systems* (pp. 33–65). New York: Springer.

Leland, H. (1980). Who should buy portfolio insurance. *Journal of Finance, 35*, 581–594.

Li, X. D. (1998). *Constructing a credit curve*. Credit Risk. A RISK Special Report.

Lintner, J. (1965). Security prices, risk and maximal gains from diversification. *Journal of Finance, 20*, 587–615.

Lobato, I., & Robinson, P. (1998). A nonparametric test for i(0). *Review of Economic Studies, 3*, 475–495.

Longin, F. (2003). From value at risk to stress testing: The extreme value approach. *Journal of Banking & Finance, 24*(7), 1097–1130.

Longstaff, F., & Schwartz, E. (1992). Interest rate volatility and the term structure: A two-factor general equilibrium model. *Journal of Finance, 47*, 1259–1282.

Lubrano, M. (1998). *Smooth transition GARCH models: A Bayesian perspective*. Technical Report, CORE, Louvain-la-Neuve.

Lumsdaine, R. (1996). Consistency and asymptotic normality of the quasi maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models. *Econometrica, 64*, 575–596.

Maercker, G. (1997). *Statistical inference in conditional heteroskedastic autoregressive models*. Aachen: Shaker.

Maller, R. A., Müller, G., & Szimayer, A. (2008). GARCH modelling in continuous time for irregularly spaced time series data. *Bernoulli, 14*, 519–542.

Mandelbrot, B. (1963). The variation of certain speculative prices. *Journal of Business, 36*, 394–419.

Mandelbrot, B. (1977). *Fractals: Form, chance and dimension*. San Francisco, CA: W. H. Freeman

Mandelbrot, k., & Van Ness, J. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review, 10*, 422–437.

McAllister, P. H., & Mingo, J. J. (1996). Bank capital requirements for securitized loan portfolios. *Journal of Banking and Finance, 20*, 1381–1405.

McGinty, L., & Ahluwalia, R. (2004). *A model for base correlation calculation*. Technical Report, JP Morgan.

McKeague, I., & Zhang, M. (1994). Identification of nonlinear time series from first order cumulative characteristics. *Annals of Statistics, 22*, 495–514.

McNeil, A. (1999). Extreme value theory for risk managers. In *Internal modelling and CAD II* (pp. 93–113). London: RISK Books.

McNeil, A., & Frey, R. (2000). Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance, 7*, 271–300.

Melino, A., & Turnbull, S. M. (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics, 45*, 239–265.

Merton, R. (1973). Theory of rational option pricing. *Bell Journal of Economics and Management Science, 4*, 141–183.

Mikosch, T. (1998). *Elementary stochastic calculus with finance in view*. Singapore: World Scientific.

Mikosch, T., & Starica, C. (2004). Non-stationarities in financial time series, the long range dependence and the igarch effects. *Review of Economics and Statistics, 86*(1), 389–390.

Mills, T. (1993). *The econometric modelling of financial time series*. Cambridge: Cambridge University Press.

Miltersen, K., Sandmann, K., & Sondermann, D. (1997). Closed form solutions for term structure derivatives with log-normal interest rates. *Journal of Finance, 52*, 409–430.

Mokkadem, A. (1987). Sur un modèle autorégressif nonlinéaire. ergodicité et ergodicité géometrique. *Journal of Time Series Analysis, 8*, 195–204.

Morillas, P. M. (2005). A method to obtain new copulas from a given one. *Metrika, 61*, 169–184.

Mossin, J. (1966). Equilibrium in a capital asset market. *Econometrica, 34*, 768–783.

Müller, G. (2010). Mcmc estimation of the COGARCH(1,1) model. *Journal of Financial Econometrics, 1*, 1–30.

Müller, M. (2000). Generalized partial linear models. In W. Härdle, Z. Hlavka, & S. Klinke (Eds.), *XploRe application guide*. Heidelberg: Springer.

Müller, M., & Rönz, B. (2000). Credit scoring using semiparametric methods. In J. Franke, W. Härdle, & G. Stahl (Eds.), *Measuring risk in complex stochastic systems*. Heidelberg: Springer.

Müller, T., & Nietzer, H. (1993). *Das große Buch der technischen Indikatoren*. Rosenheim: TM Börsenverlag.

Murata, N., Yoskizawa, S., & Amari, S. (1994). Network information criterion - Determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks, 5*, 865–872.

Mwita, P. (2003). *Semiparametric estimation of conditional quantiles for time series with applications in finance*. Ph.D. thesis, University of Kaiserslautern, Kaiserslautern.

Neftci, S. (1996). *An introduction to the mathematics of financial derivatives*. San Diego: Academic Press.

Nelsen, R. B. (2006). *An introduction to copulas*. New York: Springer.

Nelson, D. B. (1990a). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory, 6*, 318–334.

Nelson, D. B. (1990b). Arch models as diffusion approximations. *Journal of Econometrics, 45*, 7–38.

Nelson, D. B. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica, 59*, 347–370.

Newey, W., & West, K. D. (1987). A simple positive semi-definite, heteroskedastic and autocorrelation consistent covariance matrix. *Econometrica, 55*(3), 703–708.

Okhrin, O., Okhrin, Y., & Schmid, W. (2013). On the structure and estimation of hierarchical Archimedean copulas. *Journal of Econometrics, 173*(2), 21–53.

Okhrin, O., Okhrin, Y., & Schmid, W. (2013). Properties of the hierarchical Archimedean copulas. *Statistics and Risk Modeling, 30*(1), 189–204.

Overbeck, L. (2000). Allocation of economic capital in loan portfolios. In J. Franke, W. Härdle, & G. Stahl (Eds.), *Measuring risk in complex stochastic systems* (pp. 1–15). New York: Springer.

Overbeck, L., & Rydén, T. (1997). Estimation in the Cox-Ingersoll-Ross model. *Econometric Theory, 13*, 430–461.

Phillips, P. C. B., & Shimotsu, K. (2004). Local whittle estimation in nonstationary and unit root cases. *Annals of Statistics, 32*, 656–692.

Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics, 3*, 119–131.

Pitman, J. (1997). *Probability*. Heidelberg: Springer.

Polzehl, J., & Spokoiny, V. (2006). Local likelihood modeling by adaptive weights smoothing. *Probability Theory and Related Fields, 135*, 335–362.

Press, W., Teukolsky, S., & Vetterling, W. (1992). *Numerical recipes in Fortran 77: The art of scientific computing*. Reading: Addison-Wesley.

Protter, P. E. (2005). *Stochastic integration and differential equations*. Berlin: Springer.

Rabemananjara, R., & Zakoian, J. M. (1993). Threshold ARCH models and asymmetries in volatility. *Journal of Applied Econometrics, 8*, 31–49.

Rebonato, R. (2002). *Modern pricing of interest rate derivatives*. Princeton: Princeton University Press.

Rebonato, R. (2004). *Volatility and correlation*. Chichester: Wiley.

Redelberger, T. (1994). *Grundlagen und Konstruktion des VDAX-Volatilitätsindex der Deutsche Börse AG*, Deutsche Börse AG.

Refenes, A.-P. (1995a). *Neural networks for pattern recognition.* Oxford: Clarendon Press.

Refenes, A.-P. (1995b). *Neural networks in the capital market*. New York: Wiley.

Rehkugler, H., & Zimmermann, H. G. (1994). *Neuronale Netze in der Ökonomie*. München: Vahlen.

Reiss, R., & Thomas, M. (1997). *Statistical analysis of extreme values*. Basel: Birkhäuser.

Reiss, R., & Thomas, M. (2000). Extreme value analysis. In *XploRe learning guide* (pp. 353–374). Heidelberg: Springer.

Renault, E., & Touzi, N. (1996). Option hedging and implied volatilities in a stochastic volatility model. *Mathematical Finance, 6*, 277–302.

Ripley, B. (1996). *Pattern recognition and neural networks*. Cambridge: Cambridge University Press.

RiskMetrics. (1996, 17 Dec). Technical document (4th ed.). http://www.jpmorgan.com/RiskManagement/RiskMetrics/RiskMetrics.html.

Robinson, P. (1983). Non-parametric estimation for time series models. *Journal of Time Series Analysis, 4*, 185–208.

Robinson, P. (1984). Robust nonparametric autoregression. In J. Franke, W. Härdle, & D. Martin (Eds.), *Robust and nonlinear time series analysis*. Heidelberg: Springer.

Robinson, P. M. (1995a). Gaussian semiparametric estimation of long-range dependence. *Annals of Statistics, 23*, 1630–1661.

Robinson, P. M., & Henry, M. (1999). Long and short memory conditional heteroscedasticity in estimating the memory parameter in levels. *Economic Theory, 15*, 299–336.

Ross, S. (1994). *A first course in probability*. New York: Macmillan.

Ross, S., Westerfield, R., & Jaffe, J. (2005). *Corporate finance*. New York: McGraw-Hill.

Rüeger, S., & Ossen, A. (1997). The metric structure of weightspace. *Neural Processing Letters, 5*, 63–72.

Samaskij, A. (1984). *Theorie der Differenzenverfahren*. Leipzig: Akademische Verlagsgesellschaft Geest und Portig K.-G.

Samorodnitsky, G., & Taqqu, M. (1994). *Stable and non-Gaussian random processes*. New York: Chapman and Hall.

Sato, K. (1999). *Lévy processes and infinitely divisible distributions*. Cambridge: Cambridge University Press.

Savu, C., & Trede, M. (2010). Hierarchies of archimedean copulas. *Quantitative Finance, 10*, 295–304.

Schlittgen, R., & Streitberg, B. (1995). *Zeitreihenanalyse*. München: Oldenbourg.

Schönbucher, P. (1999). A market model for stochastic implied volatility. *Philosophical Transactions of the Royal Society A, 357*, 2071–2092.

Severini, T., & Staniswallis, J. (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association, 89*, 501–511.

Severini, T., & Wong, W. (1992). Generalized profile likelihood and conditionally parametric models. *Annals of Statistics, 20*, 1768–1802.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance, 19*, 425–442.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility models. In D. R. Cox, D. V. Hinkley, & O. E. Barndorff-Nielsen (Eds.), *Time series models in econometrics, finance and other fields*. London: Chapman and Hall.

Shimotsu, K. (2006). *Simple (but effective) tests of long memory versus structural breaks*. Working Paper, Queen's Economics Department, Queen's University, Ontario-Canada.

Skaug, C., & Naess, A. (2007). Fast and accurate pricing of discretely monitored barrier options by numerical path integration. *Computational Economics, 30*, 143–151.

Skiadopoulos, G., Hodges, S., & Clewlow, L. (1998). *The dynamics of implied volatility surfaces*. Working Paper, Financial Options Research Centre, Warwick Business School University of Warwick, Coventry.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publica- tions de l'institut de statistique de l'Universit e de Paris, 8*, 229–231.

Sowell, F. (1992). Maximum likelihood estimation of stationary univariate fractional integrated time series models. *Journal of Econometrics, 53*, 165–188.

Starica, C. (2003). *Is garch (1,1) as good a model as the accolades of the nobel prize would imply?* Working Paper.

Stelzer, R. (2009). First jump approximation of a lévy-driven sde and an application to multivariate ECOGARCH processes. *Stochastic Processes and their Applications, 119*, 1932–1951.

Stelzer, R. (2010). Multivariate COGARCH(1,1) processes. *Bernoulli, 16*, 80–115.

Stone, C. (1977). Consistent nonparametric regression. *Annals of Statistics, 5*, 595–645.

Studer, G. (1995). *Value at risk and maximum loss optimization*. Discussion Paper, ETHZ, RiskLab: Technical Report.

Susanne Kruse, U. N. (2005). On the pricing of forward starting options in heston's model on stochastic volatility. *Finance Stochastic, 9*, 233–250.

Szimayer, A., & Maller, R. A. (2007). Finite approximation schemes for lévy processes, and their application to optimal stopping problems. *Stochastic Processes and their Applications, 117*, 1422–1447.

Taleb, N. (1997). *Dynamic hedging: Managing vanilla and exotic options*. New York: Wiley.

Taleb, N. (2001). *Fooled by randomness: The hidden role of chance in the markets and in life*. New York: TEXERE LLC.

Taylor, S. J. (1986). *Modelling financial time series*. Chichester: Wiley.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association, 89*, 208–218.

Teräsvirta, T., Lin, C.-F., & Granger, C. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis, 14*, 209–220.

Tong, H. (1983). *Threshold models in nonlinear time series analysis. Lecture notes in statistics* (Vol. 21). Heidelberg: Springer.

Tsay, R. S. (2002). *Analysis of financial time series*. New York: Wiley.

Tse, Y. K. (1998). The conditional heteroscedasticity of the yen-dollar exchange rate. *Journal of Applied Econometrics, 5*, 49–55.

Tse, Y. K., & Tsui, A. (2002). A multivariate garch model with time-varying correlations. *Journal of Business and Economic Statistics, 20*, 351–362.

Tsevas, G., & Panaretos, J. (1998). Extreme value theory and its applications to financial risk management. In *Proceeding of the 4th Hellenic European Conference on Computer Mathematics and its Applications* (pp. 509–516).

Tsybakov, A. (1986). Robust reconstruction of functions by the local–approximation method. *Problems of Information Transmission, 22*, 133–146.

Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of Financial Economics, 5*, 177–188.

Velasco, C. (1999). Non-stationary log-periodogram regression. *Journal of Econometrics, 91*, 325–371.

Vieu, P. (1995). *Order choice in nonlinear autoregressive models*. Discussion Paper, Laboratoire de Statistique et Probabilités, Université Toulouse.

von Weizsäcker, H., & Winkler, G. (1990). *Stochastic integrals*. Braunschweig: Vieweg.

Wand, M., & Jones, M. (1995). *Kernel smoothing*. London: Chapman and Hall.

Wang, Y. (2002). Asymptotic nonequivalence of garch models and diffusions. *Annals of Statistics, 30*, 754–783.

Weiss, A. A. (1986). Asymptotic theory for ARCH models: Estimation and testing. *Econometric Theory, 2*, 107–131.

Welcker, J. (1994). *Technische Aktienanalyse*. Zürich: Verlag Moderne Industrie.

Welcker, J., Kloy, J., & Schindler, K. (1992). *Professionelles Optionsgeschäft*. Zürich: Verlag Moderne Industrie.

Whelan, N. (2004). Sampling from Archimedean copulas. *Quantitative Finance, 4*, 339–352.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*, 1–25.

White, H. (1989a). An additional hidden unit test for neglected nonlinearities in multilayer feedforward networks. In *Proceedings of the International Joint Conference on Neural Networks, Zürich*, Washington, DC.

White, H. (1989b). Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Association, 84*, 1008–1013.

White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks can learn arbitrary mappings. *Neural Networks, 3*, 535–550.

Wiggins, J. (1987). Option values under stochastic volatility: Theory and empirical estimates. *Journal of Financial Economics, 19*, 351–372.

Williams, D. (1991). *Probability with martingales*. Cambridge: Cambridge University Press.

Wilmott, P., Howison, S., & Dewynne, J. (1995). *The mathematics of financial derivatives: A student introduction*. Cambridge: Cambridge University Press.

Yan, J. (1999). *Martingale approach to option pricing - A brief review with examples*. Beijing: Institute of Applied Mathematics, Academia Sinica.

Yang, L., Härdle, W., & Nielsen, J. (1999). Nonparametric autoregression with multiplicative volatility and additive mean. *Journal of Time Series Analysis, 20*(5), 579–604.

Zakoian, J. (1994). Threshold heteroskedastic functions. *Journal of Economic Dynamics and Control, 18*, 931–955.

Zakoian, J. M. (1991). *Threshold heteroskedastic models*. Technical Report, INSEE.

Index