Bayesian Analysis of Financial Time Series

Andreas Berg

A thesis submitted in partial fulfilment of the
requirements for the degree of Doctor of Philosophy in Statistics,
The University of Auckland, 2005
Abstract

In the last decades new ideas in the analysis of time series which are based on the assumption that trading is performed continuously in time have been developed. This idea has its strength in its flexibility and can overcome many theoretical and practical difficulties experienced with discrete time analysis. Due to Itô's Lemma continuous time models are mathematically tractable, however in practice, statistical analysis is generally difficult since observations are available only at discrete times. Modelling continuous time models using stochastic differential equations including Brownian motions is a natural approach owing to the diffusion type motions of the models involved. I will introduce the most widely used processes for analysing financial models based on continuous time series and review common discretization methods, as well as approximation and estimation techniques.

I will then concentrate on stochastic volatility models, which can also be derived from continuous time models and are nowadays commonly applied to analyse stock market data or exchange rates. Additionally, stochastic volatility models offer an alternative to the Black-Scholes model as a basis for pricing options. Latest advances in estimation methods for this demanding model will be reviewed. I will focus on several ideas on performing a Bayesian analysis for
parameter estimation, one of them is based on automatic differentiation and the extended Kalman filter. This technique is compared with other Bayesian estimation procedures in terms of efficiency with simulated and real data sets.

The need for reliable model checking criteria within the class of stochastic volatility models is revealed to be a complicated issue. Many of the classical or Bayesian standard approaches can not be adapted to this task. Therefore, a new idea on how to perform Bayesian model checking is taken up and further developed for the needs of comparing stochastic volatility models. The resulting deviance information criterion is compared with the more popular but far more computationally demanding Bayes factors and the easy to derive, but less accurate harmonic means.

Furthermore I will give an overview of interest spot-rate models, which have their origin in continuous time. After introducing the latest literature on interest rate specific models, a new estimation technique based on latent observations for discretely observed diffusions using Bayesian methods is reviewed and a prior sensitivity analysis has been performed to assess estimation bias in the Vasicek model. Several case studies and simulations are carried out and results are reported at the end of each chapter.
Acknowledgements

I would like to express my gratitude towards my supervisors Dr. Renate Meyer and Dr. Jun Yu for their sensible guidance, patience, motivation, support and their constant availability. I would also like to thank Dr. Rachel Fewster for her honesty and for giving me valuable advice and encouragement in understanding the background of the PhD process. I am grateful to Prof. Angelika van der Linde and Dr. William M. (Bill) Bolstad for refereeing my thesis.

Many thanks to my former colleagues Bo Cai and Rusheu Shuo for endless discussions, practical on-the spot suggestions, positive criticisms and for sharing their thoughts with me.

I would further like to thank Gerhard and Gertrud Berg for their trust and parental support and of course my wife Ayako Sato for her love and understanding.

Financial and material support by the Royal Society of New Zealand Marsden Fund, the DAAD and the Department of Statistic at the University of Auckland is gratefully acknowledged.
# Contents

## 1 Continuous time series

1.1 Continuous time models .................................. 2  
1.2 Discrete approximation methods .......................... 7  
1.3 Continuous time approximation and estimation methods .... 11  
1.4 Summary ................................................. 24  

## 2 Stochastic volatility models  

2.1 Introduction ........................................... 27  
2.2 Models ................................................. 28  
2.3 Classical estimation methods ............................ 45  
2.4 Bayesian estimation ..................................... 52  

2.4.1 Bayesian computation using automatic differentiation and  
the extended Kalman filter .................................. 61  
2.4.2 Implementation and Comparison of EKF-AD ............. 75  
2.5 Summary ................................................. 85
CONTENTS

3 DIC and other model checking criteria 88
   3.1 Introduction ............................................. 89
   3.2 The Deviance Information Criterion ......................... 95
   3.3 Marginal Likelihood and Harmonic mean ..................... 100
   3.4 Case studies ............................................... 108
      3.4.1 A simulation study ................................... 108
      3.4.2 An empirical study ................................... 120
   3.5 Summary .................................................. 129

4 Estimation of interest rate diffusions 132
   4.1 Models .................................................... 133
   4.2 Estimation methods ........................................ 139
   4.3 Bayesian estimation ....................................... 143
      4.3.1 Bayesian estimation with augmented data ............... 146
      4.3.2 Estimation bias and prior sensitivity .................. 150
   4.4 Case study ............................................... 159
   4.5 Summary .................................................. 175
List of Figures

1.1 Realization of an arithmetic Brownian motion ............... 3
1.2 Realization of a geometric Brownian motion ............... 4
1.3 Realization of a mean-reverting process .................. 6

2.1 Trace plots and kernel density estimates of the marginal posterior distribution of $\beta, \phi, \tau$, obtained by EKF-AD for the basic SV model .......................................................... 80

2.2 Trace plots and kernel density estimates of the marginal posterior distribution of $\beta, \phi, \phi_2, \tau, \tau_2$, obtained by EKF-AD for the 2 independent AR(1) SV model .................................................. 86

3.1 Time Series Plots for Simulated Data, S&P100, VIX, Logarithm of Absolute Value of S&P100 Returns .................. 110
List of Tables

2.1 Comparison of Bayesian estimates, IACTs and CPU times ... 78
2.2 Bayesian estimates, IACTs and CPU time (in seconds) using EKF-AD for the 2 independent AR(1) model ................. 85
3.1 Parameter Estimates for Simulated Data ..................... 115
3.2 Chib's Marginal Likelihood, Harmonic Mean, DIC for Simulated Data .......................................................... 116
3.3 Deviance and Harmonic Mean (HM) Summaries for Simulated Data ................................................................. 119
3.4 Parameter Estimates for S&P 100 Data ....................... 123
3.5 Chib's Marginal Likelihood, Harmonic Mean, DIC for S&P 100 Data ................................................................. 125
3.6 Deviance and Harmonic Mean (HM) Summaries for S&P100 Data 127
3.7 Sensitivity of DIC and Chib's Marginal Likelihood to the Prior 129
<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>Short-term interest rate models as in CKLS (1992)</td>
<td>135</td>
</tr>
<tr>
<td>4.2</td>
<td>Parameter estimates for the Vasicek Model (Euler approximation, 0 auxiliary data points)</td>
<td>169</td>
</tr>
<tr>
<td>4.3</td>
<td>Parameter estimates for the Vasicek Model (Euler approximation, 1 auxiliary data point)</td>
<td>170</td>
</tr>
<tr>
<td>4.4</td>
<td>Parameter estimates for the Vasicek Model (Euler approximation, 3 auxiliary data points)</td>
<td>171</td>
</tr>
<tr>
<td>4.5</td>
<td>Parameter estimates for the Vasicek Model (Euler approximation, 5 auxiliary data points)</td>
<td>172</td>
</tr>
</tbody>
</table>
Chapter 1

Continuous time series
1.1 Continuous time models

Modern model description of financial time series is often established in continuous time with the use of stochastic differential equations (SDE). Continuous time models are able to give additional insights in practical and theoretical senses compared to discrete time models (for major issues in empirical finance, which reveal a strong preference of continuous time modelling over discrete time modelling see Cootner, 1964). This representation via SDEs needs more advanced mathematical tools to deal with the additional flexibility.

In finance, SDEs generally encompass a, possibly nonlinear, drift and a diffusion term with uncertainty described through Brownian motion. A basic dynamic model can be described through a SDE known as the \textit{Itô stochastic differential equation} as follows:

\[ dy(t) = a\{y(t), \theta\} dt + b\{y(t), \theta\}dW(t), \ t \in (0, T], \]

(1.1)

where \(a(.)\) and \(b(.)\) are the drift, interpretable as instantaneous mean, and diffusion, interpretable as instantaneous variance. Both terms depend on the underlying time series \(y(t)\) and a latent parameter \(\theta\). \(W(t)\) is a Brownian motion with \(dW(t)\) the increment of such.
1.1 Continuous time models

Frequently occurring continuous stochastic processes used in finance emerge naturally from this general description, such as (Shimko, 1992):

*The Arithmetic Brownian Motion* with

\[ dy(t) = a \, dt + b \, dW(t), \ t \in (0, T], \]  

(1.2)

\(a\) and \(b\) constants. This is a process which is widely used for modelling linear growth with increasing uncertainty (see also Figure 1.1).

Figure 1.1: Realization of an arithmetic Brownian motion \(X\) with drift 0.02 and standard deviation 0.2
1.1 Continuous time models

The Geometric Brownian Motion

\[ dy(t) = a y(t) \, dt + b y(t) \, dW(t), \quad t \in (0, T], \]  

which is an appropriate model description when exponential growth and volatility proportional to the level of the modelled variable (see also Figure 1.2) appears.

Figure 1.2: Realization of a geometric Brownian motion $X$ with drift 0.02 and standard deviation 0.2
1.1 Continuous time models

The Mean reverting (or Ornstein-Uhlenbeck process, if $\gamma = 0$)

\[ dy(t) = \kappa(\mu - y(t)) \, dt + b \, y(t) \gamma \, dW(t), \quad t \in (0, T], \]  

with speed of reversion parameter $\kappa$, long run mean $\mu$ and $\gamma$ as a sensitivity parameter for the volatility of the process. This model is suitable for variables that tend to reveal a long-term mean, but can also account for short-term shocks. The Ornstein-Uhlenbeck process has the property that the further the process departs from the long-run mean, the stronger it tends to return to the long-run mean.

A range of financial time series seem to display such a behavior. Amongst them are interest rates or inflation rates. The mean reverting process will be the starting point for discussion and analysis in Chapter 2 of this thesis. A possible realization of an Ornstein-Uhlenbeck process is displayed in Figure 1.3.

Stochastic differential equations have analytical solutions, so that the transition densities are available in closed form and direct maximum-likelihood estimation is feasible, only in the rarest cases. Wong (1964) gives an overview of these exceptions. Kloeden and Platen (1992) give an excellent introduction of stochastic differential equations, as well as methods and conditions for their
1.1 Continuous time models

Figure 1.3: Realization of an Ornstein-Uhlenbeck process $X$ with mean $\mu=2$, $\kappa=0.1$ and standard deviation 0.2

I will investigate interest rate, exchange rate and stock market index time series models which are derived from the continuous time and stochastic differential equation setup in Chapters 2 and 3.

A starting point for further research in advanced time series modelling which embeds the above mentioned processes in an 2- or 3-stage hierarchical context is given with the development of dynamic term structure models. These
1.2 Discrete approximation methods

provide additional flexibility when modelling bond and option prices, see for example Duffie and Kan (1996), Dai and Singleton (2000) and Duffee and Stanton (2004).

1.2 Discrete approximation methods

When stating financial models in continuous time, the underlying assumption is of course that trading occurs in continuous time, which is an abstract image. "However [as noted by Merton, 1992, p. 58] if the length of time between revisions is very short (or indeterminately small), then the continuous-trading solution will be a reasonable approximation to the discrete-trading solution."

In some cases, continuous time data in form of "tick-by-tick" data is available, as may be experienced when observing electronic stock markets. Nevertheless the unprejudiced use of this kind of data is questionable. As Shephard (1996) points out, institutional factors like patterns of volatility during a day, due to for example the bid-ask spread or opening and closing of various markets around the world, prevent continuous time models to be appropriate over shorter time periods. If, however, nonetheless utilized, above mentioned nuisance features need to be included into the modelling process.
Other institutional constraints determine that in some cases only weekly or monthly data are made available. In this situation, characteristic dynamics of the continuous system evolve too fast to be captured by the sampling rate. It is then nearly impossible to reconcile discrete parameter estimates and corresponding continuous time estimates, a problem that will even amplify when further implications like option pricing strategies depend on these estimates. The immediate implication is the presence of an enormous discretization bias which will inevitably result in insufficient descriptions of internal dynamics and erratic parameter estimation.

Nevertheless time discretization techniques are often employed in continuous time analysis and several of them will be summarized here:

Continuous time processes often don’t possess analytical solutions. Popular ways to circumvent this problem can be followed by using discretizing approximation techniques. The most popular and widely feasible methods are based on time discretizing the continuous time model. This can be done by introducing a finite number of time points \( \tau_0, \tau_1, \ldots, \tau_N \) replacing the time interval \( (0, T] \) in continuous time. For simplicity one often assumes equidistant time points, but this assumption can easily be relaxed. Of course, the accuracy
of time discretization relies on the frequency the underlying data is sampled. Depending on the problem specification and the data in mind, the sample frequency for which the specific data is available can vary enormously, which may necessitate refraining from the use of time discretization in some cases.

The most widely applied time discretization methods are embedded in the class of strong Taylor approximations (Kloeden and Platen, 1992). The simplest form of strong Taylor approximations is the Euler-Maruyama or often simply Euler approximation. For the Itô process (1.1), this discretization method yields:

\[ y(t + 1) = y(t) + a\{y(t), \theta\} \Delta(t) + b\{y(t), \theta\} \Delta W(t), \ t \in (0, T], \]

with \( y(0) = x(0), \Delta(t) = \tau_{t+1} - \tau_t \) and \( \Delta W(t) = W(\tau_{t+1}) - W(\tau_t) \). The increment of a Brownian motion \( \Delta W(t) \) is defined as a Gaussian random variable with mean \( E(\Delta W(t)) = 0 \) and variance \( E((\Delta W(t))^2) = \Delta(t) \).

If the next term of the stochastic Taylor formula is also included, one obtains:

\[
y(t + 1) = y(t) + a\{y(t), \theta\} \Delta(t) + b\{y(t), \theta\} \Delta W(t) + \frac{1}{2} b\{y(t), \theta\} b'\{y(t), \theta\} \Delta W(t) \times ((\Delta W(t))^2 - \Delta(t)), \ t \in (0, T],
\]

(1.6)
1.2 Discrete approximation methods

with

\[
\frac{1}{2} \{(\Delta W(t))^2 - \Delta(t)\} = \int^r_{r_t} \int^{s_2}_{r_t} dW(s_1)dW(s_2). \tag{1.7}
\]

This resulting approximation is known as Milstein scheme.

Additional time approximation schemes for this class can easily be derived by simply further expanding the stochastic Taylor formula.

For other closely related approximation schemes, as for instance weak Taylor approximations, extrapolation methods or derivative-free versions involving Runge-Kutta methods, see Butcher (1987) and Kloeden and Platen (1992).

The Euler discretization method has the property that resulting transition densities belong to the class of Gaussian densities and it is therefore a simple task to derive (log-) likelihood functions. This property is shared with discretization methods developed by Ozaki (1985) and Shoji and Ozaki (1998) who discretize the transition density between observations directly. They assume local linearity for the drift function and constant diffusion in the underlying SDE and approximate the drift term by using the differential of the drift function. Similarly Nowman (1997) approximates the nonlinear diffusion
term, as illustrated for interest rates in Chapter 4.2 of this thesis. Another approach which results in normal distributed transition densities is developed by Yu and Phillips (2001) using the Dambis, Dubins-Schwarz theorem which allows to transfer the underlying SDE into a Brownian motion after a suitable time change.

Alternatively one can examine the exact likelihood function of the continuous time stochastic process and apply discretization methods as demonstrated for instance by Kutoyants (1984) and Yoshida (1992).

1.3 Continuous time approximation and estimation methods

Based on discrete observations, the standard estimation technique for stochastic differential equations is the maximum likelihood method.

If transition densities are known, it is straightforward to derive the (log-) likelihood function and carry out maximum likelihood estimation. Of course, the accuracy of estimation depends strongly on the number of observations, which
must be large, if sensible estimation results are expected.

A second, very important estimation technique is the method of moments. There is a vast literature about how to choose the moment conditions and how many conditions should be chosen. In general, this estimation method is asymptotically less efficient than maximum likelihood, but has a far greater range of applicability.

However, since time series data are usually of finite length and are only available in finite number, asymptotic properties don’t need to have any relevance any more. Taking this point of view, Shoji and Ozaki (1997) investigate finite sample properties of maximum likelihood and generalized method of moments methods when different discretization schemes have been applied. They also take into account the behavior of different discretization methods regarding different lengths of time intervals. These finite sample performance features have been investigated via Monte Carlo experiments.

Similarly, Durham and Gallant (2002) provide an extensive Monte Carlo study on the performance of simulation-based maximum likelihood approaches for the case of a square-root process (see also Chapter 4.1). Since these methods suffer from a high degree of computational effort, they aim to improve both the existing approximation techniques and to speed up convergence of the involved Monte Carlo integration. Considering the augmented data ap-
1.3 Continuous time approximation and estimation methods

approach (as introduced at the end of this chapter), a range of discretization and importance sampling techniques, different sizes of sample paths and different numbers of latent auxiliary data points are examined in various combinations to find highly efficient approximations.

When, as is usually the case, the transition densities are unknown one can rely on approximations of the log-likelihood function or score functions. Liptser and Shiryayev (1977) provide under some conditions the log-likelihood function for \( \theta \) in equation (1.1) based on continuous time observations, assuming \( b\{y(t), \theta \} \) is known. This resulting log-likelihood function can then be approximated based on discrete observations. When using instead the corresponding continuous time score function (under regularity conditions), Hutton and Nelson (1986) show that in this case the restrictions on \( b\{y(t), \theta \} \) can be relaxed.

For an overview of estimation techniques for continuous time data it is referred to Jiang and Knight (1997) and references therein.

Pedersen (1995) remarked the strong bias in discrete time estimates and developed a method to improve the approximation of the continuous time log-likelihood function. Pedersen's idea is to derive a sequence of approximations \( \{l_n^{(N)}(\theta)\}_{N=1}^{\infty} \) with the property that for \( N \to \infty \), \( l_n^{(N)}(\theta) \) improves and approaches the correct log-likelihood function. He accomplished this by deriving
1.3 Continuous time approximation and estimation methods

A series of approximating transition densities based on splitting observation intervals in smaller subintervals, which converge as $N \rightarrow \infty$ to the unknown transition density. Using his method, equidistant time points are not necessary any more and the previously mentioned restrictions on $b\{y(t), \theta\}$ can be relaxed as well. But natural assumptions like the SDE in equation (1.1) having a, at least weak, solution, which requires the coefficients of the SDE to be specified (but not necessarily the Wiener process), and the solutions to be unique in law remain. His results are more of theoretical nature. For practical purposes however, transition densities need to be simulated.

Another way to calculate transition densities numerically has been suggested by Lo (1988). Lo shows that under weak assumptions the exact likelihood function of the underlying stochastic process can be described as a product of conditional densities. These single transition densities can be characterized by functional partial differential equations known as Kolmogorov forward and backward partial differential equations or Fokker–Planck equations. Its solutions, if existent, yield the desired likelihood function. To carry out the numerical solution procedure for each transition density, however, can be quite cumbersome. Maximum likelihood estimation can then be applied. Lo also presents a counterexample showing that maximum likelihood estimation based on a discretized version of the SDE can yield inconsistent estimators.
1.3 Continuous time approximation and estimation methods

Previously mentioned approximations of (log-)likelihoods or transition densities result finally in obtaining a functional form which allows the implementation of maximum likelihood estimation techniques. Alternatively Hansen and Scheinkman (1995) apply the generalized method of moments (GMM, see for example Hansen, 1982) to match moment condition derived under certain conditions with the help of infinitesimal generators. Similar to the well known score generators, the infinitesimal generator has known connections with the coefficients of the SDE and instead of using a likelihood function, different sets of functions, like polynomials of order $n$ have been suggested to specify drift and diffusion terms in a sufficient way. The infinitesimal generator consists of derivatives of these functions and the above mentioned connections to the parameters of the SDE. It can be used to construct sets of moment conditions which then need to be reconciled with sample counterparts. Since many candidate generators can generate the same moment condition, identifiability rules have been established. Other approaches based on moment conditions have been developed, having their roots within the generalized method of moments, which have been gradually improved with the help of efficient simulation methods.

In Duffie and Singleton’s (1993) simulated method of moments (SMM), sample paths are simulated for fixed parameter values and and arbitrarily moments
1.3 Continuous time approximation and estimation methods

can be calculated. Then, sample moments are matched with simulated moments. According to the matching results, based on minimizing the difference between the simulated moments and the sample moments, parameters are adjusted. In an iterative procedure path simulation is repeatedly carried out with the newly adjusted parameter sets. Since the entire sample paths need to be simulated for adjusted parameter estimates, this method can be computationally expensive.

At approximately the same time an approach known as indirect inference has been developed which includes as a special case the simulated method of moments. Gourieroux, Monfort and Renault (1993), and independently Smith (1993) use a discrete approximate likelihood or quasi-likelihood function. They calculate estimates of the approximate (quasi-) likelihood function for an auxiliary parameter based on their analytical form. In a second step the estimation method is repeated for simulated values which have been drawn from the continuous time model and accommodated to the parameter of interest. Calibration of the parameter of interest is carried out by maximizing quadratic forms, which include metrics between the estimated auxiliary parameters from both steps, with respect to the parameters of interest. Clearly, the efficiency of this method relies on how well the auxiliary (quasi-) likelihood function has been chosen.

The efficient method of moments (EMM) of Gallant and Tauchen (1996) repre-
1.3 Continuous time approximation and estimation methods

...sents a further development in method of moments estimation. They combine several aspects of moment condition generation from score functions based on approximative models and simulation. This method will be described in more detail for the special problems occurring in sections 2.3 and 4.2.

Other closed form approximations of transition densities are developed and implemented by Aït-Sahalia (1999, 2002). The author transforms under certain assumptions the underlying differential equation by changing measures into a "standardized" differential equation with unit diffusion. The resulting diffusion process is now based on a state variable which possesses a standardized transition density which is "close" to a standard normal density. Hence an expansion scheme based on Hermite polynomials can be applied to approximate this transition density. Finally, an accurate approximation which converges uniformly to the unknown original transition density can be obtained by back-transforming the standardized transition density via the Jacobian formula. A further refinement which proves to be more applicable in practice is given by Bakshi and Ju (2003) by using a different standardization of the differential equation.

Alternative estimation procedures are made available through considerable advances in semi- or nonparametric estimation. The advantage of parameter-
free approaches is that model misspecification can be avoided since functional forms are not arbitrarily chosen. Early work in this field has been established by Florens-Zmirou (1993) who estimated the diffusion term using discrete time data, leaving the drift term unspecified. Aït-Sahalia (1996a) presents a semi-parametric way for estimating continuous time models. In his method of "density-matching" he aims to estimate the drift and volatility functions together. For identifiability reasons, he restricts the form of the drift function by choosing it from a parametric linear class, which is adopted in most popular models in the interest rate model literature. The economically more relevant volatility function can then be estimated without any restrictions. The densities of the process can be non-parametrically estimated from the data, for instance via kernel density estimation methods and drift and diffusion terms are constructed under certain assumptions by using Kolmogorovs forward and backward partial differential equations, so that they match these densities. Following this procedure, once the drift parameter vector is identified via the expected value of the transitional distribution and estimated by *ordinary least squares* (OLS), Aït-Sahalia shows that the diffusion function only depends on the marginal density of the process.

In another similar semi-parametric approach, Conley, Hansen, Luttmer and Scheinkman (1997) leave the drift function flexible and restrict the volatility function to a power function of the current state. Instead of Kolmogorovs
1.3 Continuous time approximation and estimation methods

Forward and backward partial differential equations they use the concept of infinitesimal generators as an identification scheme. Resulting moment conditions from appropriate test functions (they show that the scores of the marginal distribution of the data will be efficient candidates) are matched to moments obtained from nonparametric marginal density estimation of the underlying process.

The approach of Conley et al. (1997) makes also use of the technique of subordination. They claim that the observations $y_j$ are actually outcomes of a process $x_{r_j}$ with independence between the process $x_t$ and the sample process $r_j$ assumed. Whereas $y_j$ are observables in actual calendar time, the underlying process $x_{r_j}$ describes a more relevant setting of "economic time" and introduces the possibility of modelling unknown time gaps between observations. Therefore it can account for systematically missing data points. This additionally induced uncertainty can also be interpreted as a particular form of stochastic volatility. However, Conley et al. (1997) leave the unobserved sample process for their investigation unspecified. Parameterizations of the drift function are localized regarding $x$ by multiplying linear combinations of the drift functions by an appropriate kernel density with compact support.

Stanton (1997) develops a nonparametric procedure and overcomes the identification problems of Aït-Sahalia (1996a). His approach avoids parametric
specifications for either drift or diffusion terms. He constructs a family of approximations to the true drift and diffusion, assuming the time difference between two observations to be small. Like time discretization procedures, the performance deteriorates when the sampling frequency decreases. However, he points out that this effect seems to be negligible for data sampled monthly, or even more frequently. Based on Chan, Karolyi, Longstaff and Sanders' (1992) naive approximation of conditional moments, the limiting moments of the underlying conditional transition density of the SDE can be represented in the form of a Taylor series expansion of higher order. Stanton (1997) makes use of the infinitesimal generator and performs estimation of conditional moments based on Kernel functions. First order approximations of drift and diffusion correspond then to Chan et al.'s (1992) conditional moments.

Jiang and Knight (1997) and Bandi and Phillips (2003) use the related concept of defining a local time process. Once the diffusion term is identified through the properties of the local time process, it can be discretely approximated by weighted averages of convoluted (smoothing) kernel functions (see also Florens-Zmirou (1993) introducing this approach using naive indicator functions instead of kernels, which proves inferior because of its discontinuities). Whereas Bandi and Phillips (2003) extend this approach to non-stationary processes and repeat the diffusion term estimation procedure again for the estimation of the drift function, Jiang and Knight (1997) use (similar to Aït-Sahalia's,
1.3 Continuous time approximation and estimation methods

1996a semi-parametric approach for the diffusion term) the relationship of the drift term to the marginal distribution of the process for identification and estimation. Bandi and Phillips (2003) show that consistency for the estimates can be achieved when data is sampled at a high frequency and long time span, but practical difficulties arise, when appropriate kernels and bandwidths have to be chosen.

Recent developments in continuous time model estimation in particular for affine models are based on the characteristic function (CF) or Fourier transform. Often, the CF is available in analytic form, whereas the (conditional) density function is not. Because the CF contains the same amount of information as the empirical distribution function, its use will be superior to any approximation to the real likelihood. The CF can also, like the conditional density function be obtained by solving the Kolmogorov forward and backward equations, but using the CF proves often more tractable. Several approaches have been developed based on the conditional CF (Singleton, 2001 and Chacko and Viceira, 2003) or the unconditional CF (Jiang and Knight, 2002). Important differences in these approaches lie in efficiency and computational effort. Whereas Singleton (2001) conditions on the entire sample path, managing very efficient results, Chacko and Viceira (2003) condition only on the previous period which gains computational speed.
1.3 Continuous time approximation and estimation methods

The methodology behind the characteristic function method lies in matching the CF and its empirical counterpart by investigating various (weighted) differences between those two and minimizing them. In the discrete time version of this method weights can be chosen as jumps at certain points. Choosing the weights continuously corresponds to the continuous time version (Jiang and Knight, 2002 and Knight and Yu, 2002).

Alternatively, instead of employing the direct CF method, the CF can also be used to develop moment conditions. Consequently, Chacko and Viceira (2003) also name their approach spectral GMM. Similarly to the continuous weights in the direct CF method, in an extension of Chacko and Viceira (2003), Carrasco, Chernov, Florens and Ghysels (2002) introduce their continuum of moments approach.

Another alternative procedure is based on the Fourier inversion formula of the CF, which for this purpose must be made available explicitly in analytic form, for carrying out maximum likelihood estimation as in Bates (1996). Here the CF back-transformed to a density function. Singleton (2001) develops a full maximum likelihood conditional CF estimator and a limited-information maximum likelihood CF estimator by not using the information included in joint conditional distributions. Whereas the former estimator results in overwhelming computational effort when the dimension of the underlying process is greater than 2, the latter estimator proves more computationally tractable.
in this case.

Like the (in the following reviewed) Bayesian methods using MCMC, this approach has been used to estimate latent variable models by using characteristic functions (Jiang and Knight, 2002) directly or numerically (or by simulation) integrating out the latent variables.

As previously described, Conley et al. (1997) allow for the specification of the observed discrete time process by a subordination, which introduces additional (latent) data points. Whereas Conley et al. (1997) had more specifically holiday or weekend effects in mind when introducing their concept of subordination, Jones (1998), Eraker (2001) and Elerian, Chib and Shephard (2001) have used an augmented latent parameter approach for mimicking missing values between observation to minimize the effect of discretization. Hereby they use the property of the Euler discretization method that for an infinite small time gap, convergence to the true diffusion process will be achieved. It makes sense to regard this approach as a reformulation from estimation of a continuous time model problem into a missing value problem for a Gaussian model. This method derives its theoretical foundations from the previously mentioned work in Pedersen (1995).

The authors choose a Bayesian approach (although also a maximum likelihood approach would be possible) which does not rely on large sample approxi-
1.4 Summary

Financial time series often originate in continuous time. In this chapter, important examples of popular continuous time models which can be described by Itô stochastic differential equations have been introduced. Since only a few
resulting stochastic processes in continuous time possess strong solutions in closed-form, one often has to rely on discretizations and approximations to obtain an analytically tractable representation.

When discretizing a continuous time process, the Euler-Maruyama approach is by far the most common method. However, several other proposals of how to time discretize such a process exist and a few of them have been shortly presented.

Through the limited tractability of continuous time processes, estimation becomes an extremely difficult task. Therefore, recent advances in approximation and estimation techniques have been reviewed. Some of those methods are mainly of theoretical nature, others are more appealing to the practitioner. Most of the methods rely extensively on simulation of moments, transition functions etc.

In the following chapters, some of these approximation methods will reappear and the model-related applications to certain specifications introduced there will be investigated in greater detail.

Since the preferred methodology pursued in this thesis is Bayesian, it will be especially focused on a Bayesian approach using data augmentation and MCMC methods. This method will be discussed in greater detail and applied to interest rate data in Chapter 4.
Chapter 2

Stochastic volatility models
2.1 Introduction

The stochastic volatility (SV) model class represents the discrete time analogue to stochastic volatility models in continuous time, where the log-volatility is modelled as a diffusion process. In both the theoretic finance literature on option pricing and the empirical finance literature, the SV model (Taylor, 1982; Hull and White, 1987; Heston, 1993) has received much attention in recent years. It has become a powerful alternative to the ARCH and GARCH models introduced by Engle (1982) and Bollerslev (1986).

The popularity of SV models can be explained by greater flexibility due to the explicitly modelled non-observed conditional variance. Its logarithm is modelled in continuous time as a diffusion process and displays in discrete time an autoregressive structure. In contrast, ARCH and GARCH models incorporate volatilities exactly determined by lagged disturbances (Hafner, 1998).

SV models were initially introduced by Clark (1973), Taylor (1982) and Tauchen and Pitts (1983) to describe the evolution of asset returns which typically exhibit changing variances over time. SV models can be cast into the framework of nonlinear state-space models with state-dependent variances, which often proves convenient as a setup for its analyses.
2.2 Models

Ghysels, Harvey and Renault (1996) and Shephard (1996) give excellent reviews of the model and Harvey, Ruiz and Shephard (1994) present a detailed comparison between GARCH and SV models.

This and the following chapter includes large parts of recently published work based on Meyer, Fournier and Berg (2003) and Berg, Meyer and Yu (2004).

2.2 Models

The logarithmic continuous time stochastic volatility model is the most popular and most widespread stochastic volatility model. It can be described in integral form (see Johannes and Polson, 2002) as:

\[ y_t = \int_{t-1}^{t} \mu_1^* \, ds + \int_{t-1}^{t} \sqrt{V_s} \, dW_s^* \]

\[ \log(V_t) = \int_{t-1}^{t} \mu_2^* + \phi^*(\log(V_s)) \, ds + \int_{t-1}^{t} \tau \, dW_s \]

with \( y_t \) the returns of the underlying asset at time \( t \), given volatility \( V \). \( \mu_1^* \) is the drift term specifying the observed returns, \( W_s^* \) and \( W_s \) are Wiener processes, which might be assumed correlated or uncorrelated and \( \mu_2^*, \phi^* \) and \( \tau \)
are parameters which drive the volatility evolution process.

Several variations of the logarithmic and affine SV model can be found in the option pricing literature.

The setup of the stochastic volatility model in (2.1) contains only one volatility factor. Current research however, focuses also on more flexible models containing, for instance a two factor diffusion specification to describe return dynamics with the drift function also depending on the volatility (Andersen, Benzoni and Lund, 2002) or having introduced a second independent factor for the diffusion term (Gallant and Tauchen, 2001). Chernov, Gallant, Ghyssels and Tauchen (2003) even consider up to 4 factor models when comparing extensively affine and logarithmic volatility models with various amounts of factors.

On the other hand, in the recent literature promising advances have been made by departing from the implicated Gaussianity of Brownian motions. Consequently more flexible stochastic processes have been suggested. For instance, Barndorff-Nielsen and Shephard (2001) consider a non-Gaussian class of Ornstein-Uhlenbeck processes, replacing the Brownian motion by a more general Lévy process and construct from there continuous time SV models.
2.2 Models

A general book-length overview of Lévy processes, which is a stochastic process with independent and stationary increments, and therefore contains a Brownian motion as a special case is available in Sato (1999), whereas the Lévy process methodology in finance and their applications is presented in Schoutens (2003) and for an option pricing theoretical review, it is referred to Boyarchenko and Levendorskii (2002).

The representation through equations (2.1) is equivalent to describing the model in differential form as:

\[ dy^*(t) = \mu_1^* dt + \sqrt{V(t)} dW^*(t) \]

\[ d\log(V(t)) = \mu_2^* + \phi^*(\log V(t)) d(t) + \tau dW(t) \]  

(2.2)

Applying an Euler approximation with \( \Delta = 1 \) yields:

\[ y_t^* - y_{t-1}^* = \mu_1^* + \sqrt{V_t} u_t, \quad t = 1, \ldots, n \]

\[ \log(V_t) - \log(V_{t-1}) = \mu_2^* + \phi^*(\log(V_{t-1})) + v_t, \quad t = 1, \ldots, n \]  

(2.3)

Here, \( y^*(t) \) are asset prices and \( u_t \) and \( v_t \) are standard Gaussian distributed
2.2 Models

error terms. The application of the Euler discretization scheme for logarithmic continuous time SV models is far from being unproblematic. For sufficient regularity conditions which guarantee that logarithmic continuous time SV models possess well defined asymptotic properties see Chernov et al. (2003, Appendix A).

For convenience one refers to this model often in state space form:

Setting \( y_t = y_t^* - y^*_{t-1}, \) \( logV_t = h_t, \) \( \mu^*_t = 0 \) and reparameterizing the model by defining \( \mu = \frac{\nu^2}{1-\phi} \) and \( \phi = \phi^* + 1 \) gives

\[
y_t | h_t = \exp(h_t/2)u_t, \quad t = 1, \ldots, n \tag{2.4}
\]

the observation equation of the state-space model, which describes the distribution of the data given unknown states, aka the log-volatilities \( h_t, \)

and

\[
h_t | h_{t-1}, \mu, \phi = \mu + \phi(h_{t-1} - \mu) + v_t, \quad t = 1, \ldots, n \tag{2.5}
\]

which forms the state equation and models the day-to-day variation of the volatilities as a Markov process. This model will be referred to as basic SV model or MODEL 1 in empirical applications which will follow in Chapter 3.4.
For the basic SV model the errors $u_t$ and $v_t$ are assumed to be uncorrelated Gaussian sequences with $u_t \sim N(0,1)$ and $v_t \sim N(0,\tau^2)$. Initialization of the unknown first state $h_0$ is established by assuming a Gaussian distribution $h_0 \sim N(\mu,\tau^2)$.

Several extensions of this basic SV model towards SV models with higher complexity will be considered here. An example of such an extension is the inclusion of a level effect in the observation equation, namely

$$y_t|h_t, x_t, \gamma = x_t^\gamma \exp(h_t/2)u_t, \quad t = 1, \ldots, n,$$

(2.6)

where $x_t$ denotes a time-varying covariate. This model serves as an example how one factor short-term interest rate models (to be introduced in Chapter 4.1) can be extended to multi factor models. The level-effect parameter $\gamma$ plays an important role in analyzing interest rate data (for details refer for example to Chan, Karolyi, Longstaff and Sanders, 1992 and Brenner, Harjes and Kroner, 1996). In other applications, for instance stock market data, it is common to set this parameter equal to 0.

Variations of SV models which will be further explored in Chapter 3.4 in-
2.2 Models

MODEL 2: An additional non-zero mean $\alpha$ is added in the observation equation:

$$
y_t | h_t, \alpha = \alpha + \exp(h_t/2)u_t, \quad u_t \sim N(0, 1),
$$

$$
h_t | h_{t-1}, \mu, \phi, \tau^2 = \mu + \phi \cdot (h_{t-1} - \mu) + v_t, \quad v_t \sim N(0, \tau^2),
$$

MODEL 3: An AR(2)-process for the state equation:

$$
y_t | h_t = \exp(h_t/2)u_t, \quad u_t \sim N(0, 1),
$$

$$
h_t | h_{t-1}, \mu, \phi, \psi, \tau^2 = \mu + \phi \cdot (h_{t-1} - \mu) + \psi \cdot (h_{t-2} - \mu) + v_t, \quad v_t \sim N(0, \tau^2),
$$

AR(p) processes can often be simpler expressed by introducing the lag operator $L$ with $L^p(x_t) = x_{t-p}$. This formulation of lagged variables leads to a generalization of SV models, which enables the researcher to consider even non-integer values for $p$. When defining the state equation as

$$
(1 - L)^d(h_t) = \mu + v_t, \quad v_t \sim N(0, \tau^2),
$$
then the hereby defined SV model is known as *long-memory stochastic volatility model* with the log-volatility at time $t$ following a fractionally integrated process with long-memory parameter $p$. This kind of SV models have been considered in discrete time by, among others, Breidt, Crato and de Lima (1998) and Jensen (2001). A continuous time derivation of long-memory processes based on fractional Brownian motions has been analyzed by Comte and Renault (1996) and continuous time SV models based on long memory processes are derived by Comte and Renault (1998). An alternative derivation based on superpositions of Ornstein-Uhlenbeck type processes illustrated in continuous time, see Barndorff-Nielsen (2000).

**Model 4:** Two independent AR(1) processes as in Harvey et al. (1994), Shephard (1996), Gallant and Tauchen (2001) and Chernov et al. (2003):

$$
y_t \mid h_t, \mu = \exp(\mu/2 + h_t^{(1)}/2 + h_t^{(2)}/2)u_t, \quad u_t \sim N(0,1)
$$

$$
h_t^{(1)} \mid h_{t-1}^{(1)}, \phi, \tau^2 = \phi h_{t-1}^{(1)} + v_t^{(1)}, \quad v_t^{(1)} \sim N(0, \tau_1^2),
$$

$$
h_t^{(2)} \mid h_{t-1}^{(2)}, \phi_2, \tau_2^2 = \phi_2 h_{t-1}^{(2)} + v_t^{(2)}, \quad v_t^{(2)} \sim N(0, \tau_2^2),
$$

(2.10)
Here, one process ideally captures long term effects of the underlying time series, whereas the remaining process should be able to describe the speed of mean-reversion. This strategy provides additional flexibility to describe the dynamics of the financial series. One could even think of the possible consideration of models with more than two (volatility) factors. In this case, special care need to be taken for identification purposes.

**MODEL 5:** The basic SV **MODEL 1** including a leverage or asymmetric effect implemented by allowing for correlation between $u_t$ and $v_{t+1}$, i.e.

\[
\begin{pmatrix}
  u_t \\
  v_{t+1}
\end{pmatrix} \overset{iid}{\sim} N\left\{ \begin{pmatrix}
  0 \\
  0
\end{pmatrix}, \begin{pmatrix}
  1 & \rho \tau \\
  \rho \tau & 1
\end{pmatrix} \right\}.
\]

This effect is often observed in financial time series, e.g. in time series of exchange rates and, even stronger, in stock market data. It reveals the market behavior, first discovered by Black (1976), which

"occurs when an unexpected drop in price (bad news) increases predictable volatility more than an unexpected increase in price (good news) of similar magnitude" (Engle and Ng, 1993, page 1752).
A continuous time derivation (Barndorff-Nielsen and Shephard, 2001) can be
gathered by adding an Ornstein-Uhlenbeck-type process component, driven by
a Lévy process, to the continuous time SV process introduced in 2.1.

For a summary of possible economic explanations of leverage effects, see Blair,
Poon and Taylor (2002).

Asai and McAleer (2004a) consider this asymmetric effect together with a
threshold effect (threshold autoregressive models have been developed by Tong,
1993; a threshold stochastic volatility model has been analyzed by So, Li and
Lam, 2002) based on an indicator function which is first analyzed by Glosten,

MODEL 6: The SV+Jumps model includes a jump component and lagged
observations in the observation equation:

\[ y_t \mid y_{t-1}, h_t, s_t, q_t, \beta, = \beta y_{t-1} + s_t q_t + \exp(h_t/2)u_t, \quad u_t \overset{iid}{\sim} N(0, 1) \]

\[ h_t \mid h_{t-1}, \mu, \phi, \tau^2, = \mu + \phi(h_{t-1} - \mu) + v_t, \quad v_t \overset{iid}{\sim} N(0, \tau^2), \]

(2.12)
2.2 Models

where $q_t$ follows a Bernoulli distribution which takes the value of one with probability $\kappa$ and zero with probability $1 - \kappa$, and $\log(1 + s_t) \sim N(-\delta^2/2, \delta^2)$.

**MODEL 7**: This model includes a jump component in the observation equation but without taking the lagged observations into consideration:

$$
y_t | h_t, s_t, q_t, = s_t q_t + \exp(h_t/2)u_t, \quad u_t \overset{iid}{\sim} N(0,1)
$$

(2.13)

$$
h_t | h_{t-1}, \mu, \phi, \tau^2, = \mu + \phi(h_{t-1} - \mu) + v_t, \quad v_t \overset{iid}{\sim} N(0, \tau^2),
$$

Including a jump component accommodates for infrequently observed large jumps in the underlying asset and together with the stochastic volatility process presents an even more flexible tool to accurately describe empirically observed features of asset prices.

"Economically, jumps in stock returns are easily rationalized: The discrete arrival of new information induces an instantaneous revision of stock prices. Adding a jump component should improve the fit to the observed time series of returns since the jumps may help accommodate outliers as well as asymmetry in the return distribution" (Andersen et al., 2002, page 1244).
2.2 Models

In a classical analysis identification of the model without introducing restrictions is not possible if the number of parameters exceeds the number of observables as is the case with all of the stochastic volatility models introduced here. These restrictions of the parameter space are naturally be imposed in a Bayesian analysis by defining proper, informative priors. Although identification can always be established within a Bayesian analysis, sampling from the posterior distribution in the case of overparametrization causes generally efficiency problems due to the associated high cross- and autocorrelations. This becomes apparent when estimating jump components.

Analyzing stock market data, jumps are not very frequently observed (on average less than 4 times a year). Due to the rare occurrence of these events, estimation problems regarding jump parameters arise. Therefore, jump components cannot be estimated with similar precision as the remaining model parameters and this uncertainty amplifies in small samples.

In many jump-diffusion SV models, returns are driven by a simple, pure Poisson jump process and independent of the stochastic volatility component (Chib, Nardari, Shephard, 2002, and Chernov et al., 2003 in a univariate and Chib, Nardari, Shephard, 2004 in a multivariate setting). This way of modelling jumps has been earlier introduced to the option pricing literature by
Very recently, more complex formulations of jumps appeared in the SV model literature, with jump intensities dependent on the underlying volatility (Andersen et al., 2002 for linear dependence and Johannes, Kumar and Polson, 1999, who let jump intensities depend on past jump times and absolute returns).

A very informative and very recent paper concerning the role of jumps in stochastic volatility models has been written by Eraker, Johannes and Polson (2003). They find strong evidence, that a jump component in the returns does not prove sufficient to describe underlying stock market or option prices time series efficiently. They note that an additional jump incorporated in the volatility process can further improve the ability of the stochastic volatility model to capture the crucial features of equity index returns. They analyze a jump in return and jump in volatility model, as previously introduced to the option pricing literature by Duffie, Pan and Singleton, (2000). Eraker et al. (2003) consider both the case that volatility and return jumps arrive independently of each other and, alternatively, that these jumps arrive at the same time and their jump sizes are correlated.
2.2 Models

A great deal of generalization regarding jump formulation in SV models is presented in continuous time by Chernov, Gallant, Ghysels and Tauchen (1999). They develop a generic Lévy process characterization of jumps encompassing most SV with jump models described previously. Chernov et al. (1999) alternatively model jump intensities which depend on the jump size of previous ones and analyze several specifications of jump size distributions.

**Model 8:** Gaussian observation errors are substituted by independent central Student-t distributions with $\nu$ degrees of freedom:

$$
\begin{align*}
    y_t \mid h_t &= \exp(h_t/2) u_t, \ u_t \iid t_{\nu} \\
    h_t \mid h_{t-1}, \mu, \phi, \tau^2 &= \mu + \phi(h_{t-1} - \mu) + v_t, \ v_t \iid N(0, \tau^2).
\end{align*}
$$

This model is introduced to capture extreme observations and stands as an example for the concept of using fat-tailed distributions, compared to the initial Gaussian assumption for the return error distribution. This more outlier-resistant version, compared with the basic SV model, has been discussed for example in Shephard and Pitt (1997) and Jacquier, Polson, Rossi (2002).

Chib et al. (2002) also analyze SV models with the departure from the Gaussian error distribution to the Student-\(t\) distribution and additionally including
2.2 Models

jumps. They give some evidence that jumps in many cases turn out to be tail realizations of the $t$-distributed SV model, whereas the combination of jumps and $t$-distributed errors does not offer a further improvement. This observation justifies the suggestion of fat-tailed distributed error models as an alternative to jump models.

In continuous time the $t$-distributed SV model can be characterized by Student $t$-based Lévy processes.

**Model 9:** This model is very similar to the SVX model introduced in Hol and Koopman (2002) which includes implied volatility as expressed by an additional covariate $x_t$:

\[
y_t \mid h_t = \exp(h_t/2)u_t, \quad u_t \overset{i.i.d.}{\sim} N(0, 1)
\]

\[
h_t \mid h_{t-1}, x_t, \mu, \phi, \tau^2, \lambda = \mu + \phi \cdot (h_{t-1} - \mu) + \lambda(x_t - \bar{x}) + v_t,
\]

\[
v_t \overset{i.i.d.}{\sim} N(0, \tau^2).
\]

The implied volatility is used in this model as an alternative source for predicting volatility and is based on calculations of option price models. The specification of the variance equation is motivated from the empirical result that implied volatilities contain useful information in forecasting future volatil-
2.2 Models

ities (see for example Blair, Poon and Taylor, 2001b). In a continuous time context Jones (2003b) uses implied volatility to estimate a constant elasticity of variance extension of the square-root model.

This collection of 9 varieties of SV models illustrates just a small selection of possibilities and is far from being exhaustive. Other SV models allow for lagged observations in linear form in the state equation (SVL/ARCH model). Additional to those models, Fridman and Harris (1998) also successfully introduce an SV model with kurtotic distribution for the state equation error term, whereas Barndorff-Nielsen (1997) suggested a normal inverse Gaussian distribution to describe volatility dynamics.

Yu and Yang (2002) presented SV models with volatility specifications based on Box-Cox transformations, which allowed for a nonlinear description of volatility dynamics and Steel (1998) proposed skewed exponential power distributions experiment as well with different error distribution specifications.

Alternative features implemented in SV models encompass trading day effects (Asai and McAleer, 2004b).

So, Lam, Li (1998) construct a Markov switching stochastic volatility (MSSV)
model, as such that parameters can switch between different states according to a hidden Markov process (for an introduction see, for example, Hamilton, 1989) based on Hamilton and Susmel’s (1994) proposal of a Markov switching ARCH (SWARCH) model.

Additionally, various combinations of previously mentioned modelling concepts have been introduced to the literature. For instance, Liu (2000) who interestingly introduces the concept of regime switching (through the observation equation) together with the long-memory assumption for the log-volatilities, by modelling them in form of a fractional integrated process; or Garcia, Luger and Renault (2001) who combine in their SV model specification the regime switching idea with non-constant, multiple leverage effects. Gallant and Tauchen (1997) estimate a range of continuous time and discrete time models including, for instance, $t$-distributed errors, long memory, splines and asymmetries.

Andersen (1994) attempts to directly translate GARCH and EGARCH models into SV models. His resulting models consist of monotone polynomial transformations of so-called stochastic autoregressive volatility (SARV) processes and allow for a flexible way of modelling volatility. Similarly, Koopman and Hol Uspensky (2002) generalize the ARCH-M model and introduce the stochastic volatility in mean (SVM) model. Due to their model representation, they
provide an interesting analysis of the relationship between mean and variance of stock returns. With their model, they investigate effects of the volatility feedback theory. Other studies regarding the SVM models include Fridman and Harris (1998) and Watanabe (1999).

Meddahi (2001) assumes in his eigenfunction approach, that the variance is a linear combination of eigenfunctions of the conditional expectation operator. To better accommodate for fat tails he introduces Hermite SV models to replace log-normal SV models and Laguerre SV models to replace square-root SV models. This approach may provide a further alternative to SV + jump or $t$-distributed SV models.

Furthermore, various multivariate SV models have also been analyzed. It turns out, that many univariate approaches can easily be generalized to the higher dimensional case. Examples are Harvey et al. (1994), Jacquier, Polson and Rossi (1995) and Danielsson (1998). One of the most general multivariate setting with a large amount of parameters to be estimated has been proposed recently by Chib et al. (2004).

Several SV models have been developed to accommodate specific economic or financial implications. They sometimes serve as an extension of already
2.3 Classical estimation methods

well established one factor models. The SV model based on CEV processes and used for analyzing the short rate will be introduced in Chapter 4.1. Additionally, SV models with gaining popularity in the option pricing literature have been developed.

2.3 Classical estimation methods

The demanding structure of a SV model with the likelihood not readily available, has been an interesting research area for many statisticians and econometricians to work on developing new and refined estimation methods. Here, several of the most important ideas from the literature about parameter estimation in the SV model setting will be summarized. A further alternative to the methods in this section will be presented in detail in subsection 2.4.1.

Obviously the "gold standard" estimation method by maximum likelihood is not directly applicable because of the absence of an analytical closed-form description of the marginal likelihood function. The crucial point is that volatility is unobservable, therefore the marginal likelihood can only be written, for example as an n-dimensional integral
2.3 Classical estimation methods

\[ \int f(y \mid h) f(h \mid \theta) dh, \]  

(2.16)

with \( y \) the vector of observations \((y_1, ..., y_n)\), \( h \) the vector of log-volatilities \((h_1, ..., h_n)\) and \( \theta \) the underlying vector of parameters.

The analytical treatment of this integral is barely feasible because of its high-dimensionality. An immediately appealing estimation approach for non-normal disturbances is known as the quasi maximum likelihood (QML) method, which is a fast indirect inference method adjusted to the complex specifications of the SV model.

The idea here is to ignore the true underlying distribution of the errors and use instead the Gaussian distribution as an approximation, so that standard Gaussian filter can be applied as has been demonstrated for the SV model by Harvey, Ruiz and Shephard (1994). They have cautiously pointed out that the accuracy of the normality approximation used in the filtering approach highly depends on where the true parameters lie in the parameter space. Indeed, the approximation will become more inaccurate as the volatility of the volatility decreases.

Related to this quasi maximum likelihood technique, but computationally more demanding, are the estimation techniques developed by Frühwirth-Schnatter
2.3 Classical estimation methods

(1994) and applied by Ball and Torous (1999). Instead of assuming a Gaussian density for the errors, the prior states at each filtering step are approximated by a Gaussian distribution with expectation and standard deviation matching those of the true distribution of the prior states, and then updated using Gaussian Kalman filtering.

Fridman and Harris (1998) and Watanabe (1999) proposed an estimation procedure based on numerical maximum likelihood and accomplished by recursive numerical integration and applying a non-Gaussian state space modeling and filtering procedure. This recursive numerical integration method relies on a carefully chosen finite point grid, selected accordingly to the consideration of integration bounds and total number of points, making this method comparatively expensive computationally.

Taking the logarithm of squared errors, Sandmann and Koopman (1998) developed a Monte Carlo maximum likelihood method by exploiting the linear structure inherent to the transformation into the logarithm of squared errors. This has been accomplished with a decomposition into a Gaussian part, analyzed with a Kalman filter and a remainder function with a simulation-based evaluation of its expectation. Their ideas are based on the importance sampling approach of Shephard and Pitt (1997) and Durbin and Koopman (1997).
Problems with Sandmann and Koopman's (1998) method arise only when the unconditional variance of the log-volatility is relatively small, and results can then be heavily biased.

Another Monte Carlo maximum likelihood estimation approach based on the untransformed SV model is applied by Danielsson (1994) and introduced in Danielsson and Richard (1993). With this simulated maximum likelihood (SML) method, Danielsson (1994) is simulating the latent volatility conditional on available information, using their accelerated Gaussian importance sampling technique. Then, Monte Carlo methods are used to indirectly integrate out latent variables and derive an unbiased estimate of the likelihood function.

A fast technique for ML estimation in SV models will be introduced in Chapter 2.4.1 which is applicable in a classical and Bayesian context. It is an efficient alternative to the quasi-ML approach proposed by Harvey et al. (1994), to the approximate ML approach by Fridman and Harris (1998) that uses recursive numerical integration to calculate the marginal likelihood, to the Monte Carlo ML approaches of Sandmann and Koopman (1998), and to the improved ML approach proposed by Durbin and Koopman (2000) that employs Monte Carlo simulation using Gaussian importance sampling and antithetic variables.
Another popular estimation method for the SV model setting is the method of moments. This method is based on the idea of matching empirical with theoretical moments and overcomes the integration problems which occur when direct likelihood approaches are taken up.

Melino and Turnbull (1990) developed and applied the generalized method of moments (GMM) for the stochastic volatility model, and Andersen and Sørensen (1996) further improved this method by choosing different weighting matrices and reducing the influence of non-stationary parameter estimates. A problem, inherent in all method of moments applications is, that one has to face the decision problem of which moments to select for the analysis. The extensive Monte Carlo study of Andersen and Sørensen (1996) provides guidance on GMM estimation of SV models.

Through the developments of ever improved simulation methods the GMM idea has again been taken up. Simulated method of moments (SMM, Duffie and Singleton, 1993) and the method of indirect inference (Gourieroux et al., 1993) emerged. As noted by Danielsson (1994), however, while for SML estimation only latent variables need to be simulated and the form of the observation equation will be retained, by using SMM the entire variable set has to be simulated to provide moment estimates.
The efficient method of moments (EMM) has been proven as one of the most popular varieties (Gallant and Tauchen, 1996). For the EMM one aims to find an auxiliary model, which serves as a good description of the true data generating process. Its score function is then used to derive expectations from which the "efficient" moment conditions can be generated. For the purpose of choosing an appropriate auxiliary model, the semi-nonparametric (SNP) density proposed by Gallant and Tauchen (1989) has often been proven sufficiently successful to approximate the score function. For the SV context, analytical tractable ARCH or GARCH likelihoods offer seemingly appropriate choices for an approximate likelihood or quasi likelihood function. Auxiliary model parameters are usually estimated by quasi maximum likelihood, relevant tuning parameters which determine the SNP density can be identified by statistical model selection criteria, for example BIC (Schwarz, 1978). Matching parameters in the auxiliary model instead of the score vector has been proposed by Gourieroux et al. (1993), which suffers under the drawback, that the auxiliary model has to be simulated repeatedly rather than once for the EMM. After deriving the score function, with the parameter vector of the auxiliary model evaluated as its quasi maximum likelihood estimate, simulation methods have to be applied to calculate its expected value.

Depending on the accuracy of the auxiliary model, needed for the implementa-
2.3 Classical estimation methods

tion of the EMM technique, to the data distribution EMM can be as efficient as maximum likelihood.

For further asymptotic properties of EMM see, for example, Gallant and Tauchen (1996) and Gallant and Long (1997), a Monte Carlo study of the performance of EMM regarding different choices of parametric and nonparametric models for the score function has been presented by Andersen, Chung and Sørensen (1999).

Many discrete time SV models have been estimated and most continuous time SV models have been analyzed with the EMM technique. Despite often reported as not very efficient, especially for small sample situations, its strength is the flexibility to be applicable to a very broad range of model specifications. A large amount of publications can be named, where EMM has been successfully applied amongst them are Andersen et al. (1999), Liu (2000), Gallant and Tauchen (2001), Chernov et al. (2003), Andersen et al. (2002).

Whereas the likelihood function in the SV setting is not directly available, the joint and conditional characteristic functions are. The empirical characteristic function can therefore be used to carry out parameter estimation as has been done for the discrete time model by Yu (1998) and Knight, Satchell and
2.4 Bayesian estimation


Some very special specifications of the SV model demand tailored estimation procedures. For instance, the analysis of the long-memory SV model as introduced in Breidt et al. (1998) has been accomplished via a spectral likelihood estimator, which is based on applying QML estimation in the frequency domain and by Jensen (2001) using wavelet techniques. Further improvements in estimating SV models can be achieved by using case specific algorithms for Bayesian likelihood estimation (Andersen et al., 1999), which will be introduced in the next subsection.

2.4 Bayesian estimation

Estimation procedures introduced in this chapter so far have been developed from a frequentist or classical point of view. Here, a different philosophy, the Bayesian point of view, together with different methods and theories, will be explored. There is a long history of controversy and disagreement between
these two positions (for an introduction of the different philosophies and principles, with important examples of contradictions in interpretation see Carlin and Louis, 2000, Chapter 1 and references therein). Fortunately many researchers now often see that the two philosophies can be used to complement each other.

Both, classical and Bayesian concepts have been applied for the estimation process in SV models. By choosing non-informative priors one cold try to construct a situation which is close to the classical case. However, for many parameters, there exists form an economics point of view a strong a-priori sense of knowledge or assumptions about "reasonable" parameter support. Using this knowledge or assumptions, as for example a fairly strong persistence belief in volatility, one obtains generally improved estimation results by implementing this additional information into a Bayesian analysis. Additionally, the Bayesian paradigm takes into account this induced parameter uncertainty.

Reconsider the basic SV model from Chapter 2.2 which specifies the conditional distribution of the observations given unknown states, i.e., the underlying latent volatilities, $h_t$, in the observation equation

$$ y_t | h_t = \exp(h_t/2) u_t, \quad u_t \sim \text{iid } N(0, 1), \quad t = 1, \ldots, n. $$

(2.17)
The unknown states are assumed to follow a Markovian transition over time given by the state equations

\[ h_{t+1}|h_t, \mu, \phi, \tau^2 = \mu + \phi(h_t - \mu) + v_t, \quad v_t \overset{iid}{\sim} N(0, \tau^2), \quad t = 1, \ldots, n - 1, \] 

(2.18)

with \( h_1 \sim N(\mu, \frac{\tau^2}{1-\phi^2}) \). Here, the state transition from \( h_{t-1} \) to \( h_t \), \( f(h_{t-1}, \theta) = \mu + \phi(h_{t-1} - \mu) \), is linear but can easily be generalized to a nonlinear function \( f \). The formulae given below are expressed in general terms for nonlinear state transition functions \( f \) and not necessarily Gaussian observations and state distributions.

The focus is on estimating the unknown parameters \( \theta = (\mu, \phi, \tau^2) \) given the observations \( y_t, t = 1, \ldots, n \). A fully Bayesian approach specifies the joint distribution of all \( n \) observations \((y = (y_1, \ldots, y_n))\) and \( p = 3 \) parameters \( (\theta = (\mu, \phi, \tau^2)) \).

The change of focus from that in Chapter 3 has been performed for comparability reasons. The integration sampler (see Chapter 2.4.2) as in Kim et al. (1998) is also based on integrating out volatilities to speed up sampling procedures for the parameter vector \( \theta \). It will be shown in Chapter 3.3 how
volatilities can be sampled from particle filter procedures.

The joint probability density function (PDF) \( p(y, \theta) \) can be factorized into the product of the PDF of parameters, \( p(\theta) \), referred to as the prior PDF, and the conditional PDF of the observations given the parameters, \( p(y \mid \theta) \), referred to as the likelihood, i.e., \( p(y, \theta) = p(y \mid \theta)p(\theta) \).

After observing the data, the prior knowledge about the parameters, as quantified through the prior PDF of \( \theta \), is updated to the posterior PDF, \( p(\theta \mid y) \), via Bayesian theorem

\[
p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)}, \tag{2.19}
\]

where \( p(y) = \int p(y \mid \theta)p(\theta)d\theta \) is the marginal PDF of \( y \).

Due to the conditioning on unobserved states in a state-space model, the likelihood \( p(y \mid \theta) \) is not available in closed form but requires \( n \)-dimensional integration over the state vector \( h = (h_1, \ldots, h_n) \) as

\[
p(y \mid \theta) = \int p(y, h \mid \theta)dh = \int p(y \mid h, \theta)p(h \mid \theta)dh. \tag{2.20}
\]
Taking the temporal structure of the observations into account, we can factorize the likelihood by successive conditioning into

$$p(y \mid \theta) = p(y_1 \mid \theta) \prod_{t=2}^{n} p(y_t \mid y_{t-1}, \theta),$$

(2.21)

where \(y_{t-1} = (y_1, \ldots, y_{t-1})\) collects all the observable information obtained until time \(t-1\). Thus, the \(n\)-dimensional integration in (2.20) can be reduced to \(n\) successive one-dimensional integrations, starting with

$$p(y_1 \mid \theta) = \int p(y_1 \mid h_1, \theta)p(h_1 \mid \theta)dh_1$$

(2.22)

and, subsequently, for \(t = 2, \ldots, n\)

$$p(y_t \mid y_{t-1}, \theta) = \int p(y_t \mid h_t, \theta)p(h_t \mid y_{t-1}, \theta)dh_t.$$  

(2.23)

This also implies that the data can be processed in a single sweep, updating knowledge about states as more information is received.

Classical and Bayesian techniques for analyzing state space models have recently been reviewed by Durbin and Koopman (2000) with focus on Gaussian
2.4 Bayesian estimation

importance densities which approximate the models with unspecified densities around the conditional mode of the states given observations. They avoid MCMC methods. More often, however, a Bayesian approach to parameter estimation in the SV model, which is based on a sound statistical paradigm, relies on MCMC techniques (Chen, Shao and Ibrahim, 2000).

With the development of efficient simulation techniques for posterior computation, modern Bayesian inference no longer ends with posterior mode estimation, rather it has become possible to draw a sample from the posterior PDF and use this sample to summarize the posterior distribution. Various techniques are feasible to obtain a sample from the posterior (2.31), e.g. sampling/importance re-sampling, adaptive versions of importance sampling as in Givens and Raftery (1996) and MCMC algorithms such as Gibbs sampling or the MH algorithm (Smith and Roberts, 1993).

Very recently, a large amount of Bayesian approaches based on MCMC techniques have dominated the SV literature. The idea behind MCMC sampling is to construct a Markov chain whose limiting distribution is the target posterior density of interest. The Markov chain is then iterated a large number of times and the sampled draws beyond a burn-in period are treated as variates from the target distribution. For a general discussion see for example Chib and

Carlin, Polson and Stoffer (1992) showed how to implement the Gibbs sampler for nonlinear non-normal state-space models, based on the idea of data augmentation (Tanner and Wong, 1987) and since then, various MCMC techniques have been suggested that are particularly tailored to SV models.

Jacquier, Polson and Rossi (1994) introduced Bayesian MCMC methods for the SV model. They combine the Gibbs sampling technique with Metropolis-Hastings updating schemes and compare their method with with several non-likelihood based classical estimation methods. Their single-update Gibbs sampler is used in Kim et al. (1998) and Meyer and Yu (2000) to generate a sample from the joint posterior distribution of unknown parameters and unknown states. The single state Metropolis-Hastings step updating scheme within Gibbs sampling has been further refined by Geweke and Tanizaki (2001) and Stroud, Polson and Müller (2004).

Alternative examples of its implementation for non-standard SV models are given in So et al. (1998) and Steel (1998). The Lévy process based models of Barndorff-Nielsen and Shephard (2001) for the case that volatility structure
have been modelled by a marked point process and analyzed by Roberts, Papaspiliopoulos and Dellaportas (2004) using Bayesian MCMC methods.

However, due to dependencies between consecutive states the Markov chain to traverse the state space in only very tiny steps and thus mix inefficiently. Therefore, convergence of the Markov chain to the equilibrium distribution is slow, so a large number of iterations is required to achieve stationarity, and the estimation procedure becomes very time-consuming.

This led researchers to focus on the development of more efficient MCMC techniques using *block-updating schemes* as in Carter and Kohn (1994), Frühwirth-Schnatter (1994) and De Jong and Shephard (1995). Further improvements have been achieved by Shephard and Pitt (1997) and the following correction of Watanabe and Omori (2004), as well as the introduction of *auxiliary discrete mixture models* that approximate the given model as in Kim, Shephard and Chib (1998) for the special mixing case of fixed weights, and Stroud, Müller and Polson (2003) for more flexible state-dependent weights.

Kim et al. (1998) linearize the state space model induced by the SV model and approximate the resulting observation equation error distribution by a mixture of 7 normal distributions and improve MCMC algorithms by sampling the variance and autoregressive parameters in one block and the entire volatility in
another block as in Jacquier et al. (1994).

An extension regarding the generalization of the mixture is given by Mahieu and Schotman (1998) who fix the number of mixture components and include mean and variance mixture components into the parameter vector. They perform a fully Bayesian analysis with this setup, which proves to be more flexible than assuming an initial normal distribution of the errors.

The auxiliary offset mixture representation in Kim et al. (1998) was found to improve the correlation behaviour of the simulations, but only to a small extent. A bigger gain was obtained by employing the Gaussian structure of the mixture representation and integrating out the log-volatilities using an augmented version of the Kalman filter. Further developments of this integration sampler technique have been provided for example by Chib et al. (2002). Very recent approaches in this area have been demonstrated by Roberts et al. (2004) by applying the methodology of centred (Gelfand, Sahu and Carlin, 1995) and non-centred (Papaspiliopoulos, Roberts and Sköld, 2003) parameterizations. A general overview of the dependence between convergence of MCMC algorithms and estimation is given in Roberts (2003).

Jones (1998) estimates a continuous time SV model using the missing data technique as to be introduced in Chapter 4.3.1. His approach relies computa-
2.4 Bayesian estimation

tionally heavily on the MCMC samplers developed by Jacquier et al. (1994, 1995) for the discrete time SV model. In his investigation of S&P 500 returns, he also focuses on elimination of measurement and rounding errors by including these features into modelling and sampling procedures.

Various efficiency evaluations and comparisons between different estimation methods have been carried out. Generally, simpler procedures as GMM or QML perform nearly equally efficient, whereas MCMC and simulated methods (SMM, EMM, SML) dominate the former procedures in terms of more accurately estimating the SV model, unfortunately they come with the price of being computationally more intensive. Jacquier et al. (1994) note, for example, that the SML method (Danielsson and Richard, 1993) could theoretically be used to perform Bayesian analysis, but suffers from computational infeasibility.

2.4.1 Bayesian computation using automatic differentiation and the extended Kalman filter

In this section, an alternative approach to the offset mixture representation of Kim et al. (1998) will be developed. It is also based on the idea of gaining ef-
ficiency by first integrating out the unknown states. This reduces the problem of sampling vectors in a high \((n + p)\)-dimensional space to that of sampling in a low \((p)\)-dimensional space.

If the dependence on states were *linear*, this integration could be performed using the Kalman filter. Due to the nonlinearity and the state-dependent variances in the SV model, however, this is not feasible.

Thus an approach that combines the extended Kalman filter (EKF) (Harvey, 1989) with the Laplace approximation (Laplace, 1986) is suggested. The EKF has been developed for non-linear non-Gaussian state-space models whereas the Laplace approximation has a long tradition in Bayesian computation as an asymptotic approximation to the posterior distribution (Gamerman, 1997).

Here, the term 'extended Kalman filter' is used in a broad sense, referring to extensions of the classical Kalman filter where first- and second-order Taylor series expansions with numerical approximations to posterior modes and curvatures result in normal predictive and filtering densities (Fahrmeir and Tutz, 1994).

The proposed technique is not restricted to Gaussian errors but can also be applied to robustify models by allowing for outlying observations through heavy-tailed error distributions. It is also general enough to allow nonlinear state transitions.
Once an analytic approximation to the likelihood function has been obtained via the EKF, one could use a standard numerical optimization routine such as the Newton-Raphson algorithm to find the ML estimate, if one pursues a frequentist approach, or to locate the posterior mode for Bayesian inference. However, the performance of numerical maximization techniques depends crucially on information about derivatives. It can be significantly enhanced by specifying the exact gradient rather than using finite differences which are subject to truncation and subtractive cancellation errors. Hand coding, though, is tedious and error prone. Also, packages such as Mathematica for symbolic differentiation would need to be used with great care to deal with loops and branching and to avoid exponential growth in the expressions produced.

Here automatic differentiation (AD) (Griewank and Corliss, 1991) is suggested to generate first-order (and possibly second-order) derivatives of the log-posterior density. AD performs computational derivation, i.e. given a computer code for a real-valued function of say $p$ parameters, $f(\theta_1, \ldots, \theta_p)$, it returns code for evaluating its gradient $\nabla f = (\partial f / \partial \theta_1, \ldots, \partial f / \partial \theta_p)$.

AD algorithms basically efficiently provide the researcher with numerical derivatives of any order for the required scalar valued differentiable function, calculated to machine precision. By decomposing the original function of interest
into simple components, systematic application of common differentiation rules for each of the components is carried out contemporaneously.

For instance, the function \( f(k) = k(\ln k)^3 \) might be decomposed into
\[
\theta_1 = k, \quad \theta_2 = \ln \theta_1, \quad \theta_3 = \theta_2^3, \quad \theta_4 = \theta_1 \theta_3.
\]

Having obtained derivatives for each currently derived decomposition step, one arrives at the last step at the gradient for the original function:
\[
\nabla \theta_1 = 1,
\]
\[
\nabla \theta_2 = \frac{1}{\theta_1} \nabla \theta_1 = \frac{1}{k},
\]
\[
\nabla \theta_3 = 3 \theta_2^2 \nabla \theta_2 = 3(\ln k)^{\frac{3}{2}} k,
\]
\[
\nabla \theta_4 = \theta_1 \nabla \theta_3 + \theta_3 \nabla \theta_1 = 3k(\ln k)^{\frac{3}{2}} + (\theta_2)^3 = 3(\ln k)^{\frac{3}{2}} + (\ln k)^3.
\]

This algorithm is known as being forward directed. However, this algorithm is generally less efficient as an also developed backwards directed algorithm. In this procedure, starting from the last decomposition element, a sequence of partial derivatives can be evaluated, leading as a by-product to the required gradient.

In the example above, one would obtain:
\[
\frac{\partial \theta_4}{\partial \theta_3} = \theta_1,
\]
2.4 Bayesian estimation

\[ \frac{\partial \theta_2}{\partial \theta_1} \frac{\partial \theta_4}{\partial \theta_3} = \theta_1 \theta_2^2, \]

\[ \frac{\partial \theta_3}{\partial \theta_1} = \theta_3 + \frac{\partial \theta_4}{\partial \theta_2} \frac{\partial \theta_1}{\partial \theta_3} = \theta_3 + \theta_1 \theta_2^2 \frac{1}{\theta_1} = (\ln k)^3 + 3(\ln k)^2. \]

Modern AD based optimisation software uses a combination of both algorithms which reduces the internal memory required.

For every parameter involved at every stage during the function coding process, gradients will be automatically calculated by AD. This proves very advantageous, especially for more demanding models and parameter spaces of higher dimension. Here, approximation errors due to inefficient numerical derivative calculation will likely sum up, having a huge effect on the desired accuracy. In contrast, implementation of symbolic derivatives without an automated mechanism into a computer program can be very time-consuming and limiting with growing memory problems.

An introduction to AD can also be found in Rall and Corliss (1996) and Skau (2002), details are given in Griewank (2000).

The webpage of the Argonne National Laboratory Computational Differentiation Project (http://www-unix.mcs.anl.gov/autodiff/index.html) provides an overview of available software tools in FORTRAN and C for AD.

Its main applications have been for function maximization in disciplines such as Engineering, Computer Science and Chemistry.
2.4 Bayesian estimation

Only very few statistical applications exist, where AD has been employed to find the maximum of the likelihood function (Hovland, Bischof, Spiegelman and Casella, 1997 and Skaug, 2002, in the context of random effects models). Here, the use of AD to find the posterior mode and to design an efficient Metropolis-Hastings (MH) algorithm for Bayesian posterior computation is suggested. It is well known that the mixing behaviour of MH algorithms depends upon a judicious choice of the proposal density which should be close to the target density. A multivariate Gaussian proposal centred at the current state and with covariance matrix equal to the inverse of the Hessian matrix of the log-posterior evaluated at the posterior mode is constructed by using the quasi-Newton algorithm combined with AD to generate exact first-order derivatives of the log-posterior density. To this end, the software AD Model Builder (Fournier, 2000) (http://otter-rsch.com/admodel.htm), a C++ software package that integrates AD and a quasi-Newton algorithm for function minimization is employed. This yields an extremely flexible, effective and user-friendly MCMC technique for SV models referred to as EKF-AD in what follows.

It will now be demonstrated how to approximate the likelihood in (2.21) by successively applying the Laplace approximation (Gamerman, 1997) to the integrals in (2.22) and (2.23).
For instance, in the light of just the first observation $y_1$, the prior $p(h_1 \mid \theta)$ of the unknown state $h_1$ can be updated to the filtering PDF via Bayes theorem

$$p(h_1 \mid y_1, \theta) = \frac{p(y_1 \mid h_1, \theta)p(h_1 \mid \theta)}{p(y_1 \mid \theta)},$$

(2.24)

where the denominator is just the first factor in the likelihood decomposition in (2.21), given in (2.22). If both likelihood and prior, $p(y_1 \mid h_1, \theta)$ and $p(h_1 \mid \theta)$ respectively, were Gaussian and linear in $h_1$, the posterior would again be Gaussian and the normalization constant in the denominator easy to determine.

However, $p(y_1 \mid h_1, \theta)$ is not linear in $h_1$. Nevertheless, the denominator can be approximated by the so-called Laplace approximation, an asymptotic approximation of the posterior distribution that dates back to the work of Laplace in the 18th century (Laplace, 1986). For the integral in (2.22) this yields

$$p(y_1 \mid \theta) \approx \sqrt{2\pi}e^{-\psi_1(y_1, \hat{h}_1, \theta)}|D^2\psi_1(y_1, \hat{h}_1, \theta)|^{-1/2},$$

(2.25)
where

\[ \psi_1(y_1, h_1, \theta) = - \log[p(y_1 | h_1, \theta)p(h_1 | \theta)] \]
\[ = \frac{1}{2} \log(2\pi e^{h_1}) + \frac{1}{2} e^{-h_1} y_1^2 + \frac{1}{2} \log\left( \frac{2\pi r^2}{1 - \phi^2} \right) + \frac{1}{2\tau^2} (h_1 - \mu)^2. \]

\[ D^2\psi_1(y_1, h_1, \theta) \]

denotes the second-order derivative of the function \( \psi_1(y_1, h_1, \theta) \)
with respect to \( h_1 \), \( \hat{h}_1 = \text{arg min}_{h_1} \psi_1(y_1, h_1, \theta) \), and \( \hat{\sigma}^2 := |D^2\psi_1(y_1, \hat{h}_1, \theta)|^{-1} \).

This is easily seen by a second-order Taylor series expansion of \( \psi_1(y_1, h_1, \theta) \) at \( \hat{h}_1 \).

We now learn about a state at time \( t \), successively for \( t = 2, \ldots, n \), given contemporaneously available information. This is done repeatedly in a two-stage procedure by on-line extended Kalman filtering. In the first stage of the EKF, after observing \( y_{t-1} \) but before observing \( y_t \), the predictive PDF of \( h_t | y_{t-1}, \theta \) is approximated by a Normal PDF \( \tilde{p}(h_t | y_{t-1}, \theta) \) with mean and variance given by

\[ \beta_t = f(h_{t-1}, \theta) = \mu + \phi(h_{t-1} - \mu) \]  
\[ (2.26) \]

and
respectively, using a first-order Taylor series expansion of \( f(h_{t-1}, \theta) \) at the approximate mean \( \hat{h}_{t-1} \) of \( h_{t-1} \mid y_{t-1}, \theta \). Here, \( f'(h, \theta) \) denotes the first derivative of \( f(h, \theta) \) with respect to \( h \). In the second stage, after observing \( y_t \), the filtering PDF \( p(h_t \mid y_t, \theta) \) is updated via Bayes theorem to

\[
p(h_t \mid y_t, \theta) \propto p(y_t \mid h_t, \theta)p(h_t \mid y_{t-1}, \theta) \approx p(y_t \mid h_t, \theta)p(h_t \mid y_{t-1}, \theta). \tag{2.28}
\]

Using the Laplace approximation then yields an approximation to the \( t \)th likelihood contribution in (2.23):

\[
p(y_t \mid y_{t-1}, \theta) = \int p(y_t \mid h_t, \theta)p(h_t \mid y_{t-1}, \theta)dh_t \\
\approx \int p(y_t \mid h_t, \theta)p(h_t \mid y_{t-1}, \theta)dh_t \\
= \sqrt{2\pi} e^{-\psi(y_t, \hat{h}_t, \theta)}|D^2\psi_t(y_t, \hat{h}_t, \theta)|^{-1/2}, \tag{2.29}
\]
2.4 Bayesian estimation

where

\[
\psi_t(y_t, h_t, \theta) = - \log[p(y_t \mid h_t, \theta)p(h_t \mid y_{t-1}, \theta)] \\
= \frac{1}{2} \log(2\pi e^{h_t^2}) + \frac{1}{2} e^{-h_t} y_t^2 + \frac{1}{2} \log(2\pi \gamma_t^2) + \frac{1}{2} \gamma_t^2 (h_t - \beta_t)^2,
\]

\(\hat{h}_t = \arg \min_h \psi_t(y_t, h_t, \theta)\), and \(\hat{\sigma}_t^2 = |D^2 \psi_t(y_t, \hat{h}_t, \theta)|^{-1}\). The Newton-Raphson algorithm has been used to derive \(\hat{h}_t\). Already very few iterations usually suffice to get close enough to the minimum.

Completion of this sequential two-stage procedure yields a closed-form approximative expression for the likelihood of (2.21) that no longer depends on the latent states \(h\). More precisely, this likelihood is given by

\[
\tilde{p}(y \mid \theta) = \exp \left\{ -\frac{1}{2} \sum_{t=1}^{n} \log(2\pi e^{\hat{h}_t^2}) - \sum_{t=1}^{n} \frac{1}{2} e^{h_t} y_t^2 - \frac{1}{2} \sum_{t=1}^{n} \log(2\pi \gamma_t^2) \\
- \sum_{t=1}^{n} \frac{1}{2} \gamma_t^2 (\hat{h}_t - \beta_t)^2 - \frac{1}{2} \sum_{t=1}^{n} \log(\frac{1}{\hat{\sigma}_t^2}) \right\},
\]

(2.30)

with \(\beta_1 = \mu\) and \(\gamma_1^2 = \frac{\sigma^2}{1-\phi^2}\). This approximation is exact for linear Gaussian state-space models. From (2.19) the posterior PDF can be obtained up to a normalization constant:

\[
\tilde{p}(\theta \mid y) \propto p(\theta)\tilde{p}(y \mid \theta).
\]

(2.31)
Due to marginalization over the latent volatilities, EKF-AD can provide inference only for the model parameters but not the state variables.

However, it is possible to approximate the conditional likelihood function \( p(y \mid \theta, h) \) by making use of the 'particle filter' method (Kitagawa, 1996) or by more efficient techniques such as the auxiliary particle filter (Pitt and Shephard, 1999a). In the context of SV models, this particular algorithm has been implemented in Kim et al. (1998), Pitt and Shephard (1999b) and Chib et al. (2002). Furthermore, due to iterated Laplace approximations, the inference is based on an approximation \( \tilde{p}(y \mid \theta) \) (compare with equation 2.30) to the exact marginal likelihood, which is exact only for linear Gaussian state-space models. Therefore, posterior inference is not exact, which is reflected in the discrepancy between estimates of the parameter \( \beta \) in an example in Chapter 2.4.2.

Tierney and Kadane (1986) provide guidelines towards the justified use of the Laplace approximation. Lewis and Raftery (1997) note that the Laplace approximation often works well, even if some functions are not entirely defined according Tierney and Kadane's (1986) guidelines. Lewis and Raftery (1997) and references therein present non-Gaussian examples for which the Laplace approximation generates very accurate results.
Once the (approximate) posterior PDF (2.31) has been obtained, traditional Bayesian posterior inference proceeds by using the Bayesian central limit theorem, i.e., finding the multivariate normal approximation $N(\mu, \Sigma)$ by calculating the posterior mode $\mu$ and its approximate variance, the latter by evaluating the Hessian matrix of minus the log-posterior at the posterior mode:

$$\phi(\theta) = -\log[p(\theta)p(y | \theta)],$$

$$\mu = \arg\min_{\theta} \phi(\theta),$$

$$\Sigma = D^{2}\phi(\mu)^{-1}.$$ 

Thus, one has to find the posterior mode, or alternatively minimize $\phi(\theta)$. To this end, the quasi-Newton algorithm has been employed and AD (Griewank and Corliss, 1991) has been used to calculate the first-order partial derivatives of $\phi(\theta)$ as implemented in the software AD Model Builder (Fournier, 2000). This can be done to the same degree of accuracy as the function evaluation itself. AD Model Builder uses AD implemented in a C++ class library which combines an array language with the reverse mode of AD supplemented with precompiled adjoint code for the derivatives of common array and matrix operations. This results in an extremely fast and accurate function minimization
2.4 Bayesian estimation

Here, the MH algorithm, developed by Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (1953) and generalized by Hastings (1970) is used.

The MH algorithm is top of the list of the great algorithms of 20th century scientific computing (Dongarra and Sullivan, 2000). It is a MCMC method which means that it generates a Markov chain whose equilibrium distribution is just the target posterior distribution. The MH algorithm shares the concept of a generating PDF with the well-known simulation technique of rejection sampling (Gamerman, 1997).

However, the candidate generating PDF $q(\theta \mid \theta_c)$, $\int q(\theta \mid \theta_c) d\theta = 1$, can now depend on the current state $\theta_c$ of the sampling process. A new candidate $\theta^*$ is accepted with a certain acceptance probability $\alpha(\theta^* \mid \theta_c)$ also depending on the current state $\theta_c$ and chosen such that the transition probability $p(\theta_c, \theta^*) = q(\theta \mid \theta_c)\alpha(\theta^* \mid \theta_c)$ satisfies detailed balance. This is met by setting

$$\alpha(\theta^* \mid \theta_c) = \min \left\{ \hat{p}(\theta^* \mid y)q(\theta_c \mid \theta^*), 1 \right\}. $$

The outcomes from the MH algorithm can be regarded as a sample from the
invariant PDF only after a certain 'burn-in' period.

The efficiency of the MH algorithm depends crucially on the choice of the proposal PDF. Similar to rejection sampling, the efficiency can be improved by choosing a proposal that is 'close' to the posterior PDF. Having already obtained the Laplace approximation to the posterior PDF via AD, it is obvious to make use of this approximation to \( \tilde{p}(\theta \mid y) \) to determine a good proposal PDF. Thus, a multivariate normal PDF with mean equal to the current state \( \theta^{(i)} \) and covariance matrix \( \Sigma \) equal to the inverse of the Hessian matrix of minus the log-posterior evaluated at the posterior mode is employed. The covariance matrix is dynamically scaled until a reasonable acceptance rate in the MH algorithm is observed.

But one should note that in order to obtain a sample from the posterior, it is not absolutely necessary to employ AD. Once the approximate PDF (2.31) has been obtained by EKF, any common SIR or MCMC technique can be used.
2.4 Bayesian estimation

2.4.2 Implementation and Comparison of EKF-AD

The EKF-AD technique described in section 2.4.1 has been implemented here for the analysis of SV models using the software package AD Model Builder which integrates AD with a quasi-Newton function minimization algorithm. First, the basic SV model specified in equations (2.17) and (2.18) has been fitted to a time series of daily pound/dollar exchange rates. This dataset has been previously analyzed by Harvey et al. (1994) and more recently by Shephard and Pitt (1997), Kim et al. (1998) and Meyer and Yu (2000) using Gibbs sampling and by Durbin and Koopman (2000) using a maximum likelihood as well as a Bayesian approach via importance sampling.

The data consist of a time series of daily pound/dollar exchange rates \( \{x_t\} \) from 01/10/81 to 28/06/85. The series of interest are the daily mean-corrected returns, \( \{y_t\} \), given the transformation \( y_t = \log x_t - \log x_{t-1} - \frac{1}{n} \sum_{t=1}^{n} (\log x_t - \log x_{t-1}), t = 1, \ldots, n. \)

Prior independence of the parameters \( \mu, \phi \) and \( \tau^2 \) has been assumed, and the same priors as in Kim et al. (1998) have been used. A slightly informative prior for \( \mu, \mu \sim N(0,10) \) has been employed. By setting \( \phi = 2\phi^* - 1 \) and specifying a Beta\((\alpha, \beta)\) prior for \( \phi^* \) with \( \alpha = 20 \) and \( \beta = 1.5 \), a prior mean for
\( \phi \) of 0.86 is obtained. A conjugate inverse-gamma prior is chosen for \( \tau^2 \), i.e.,
\[
\tau^2 \sim IG(2.5, 0.025)
\]
with a prior mean of 0.0167 and prior standard deviation of 0.0236.

Using the EKF followed by AD and quasi-Newton algorithm, posterior modes and asymptotic standard deviations (in brackets) for \( \phi \), \( \tau \) and \( \beta \) of 0.9830 (0.00909), 0.1445 (0.03072) and 0.7932 (0.12643) are obtained, respectively.

Then, the MH algorithm as described previously (EKF-AD) has been employed to obtain samples from the posterior distribution. But also an alternative random walk MH algorithm (EKF-RW) has been implemented, that does not require AD. In EKF-RW, after a pilot run of 10,000 iterations using a multivariate normal proposal with mean equal to the current state and a diagonal covariance matrix with rough variance estimates for each parameter obtained from experimental runs, 100,000 iterations have been performed using a multivariate normal proposal with the empirical covariance matrix from the pilot run as proposal covariance. The covariance matrix has been scaled using the optimal scaling constant \( 2.4/\sqrt{d} \) with \( d = 3 \) of Gelman, Roberts and Gilks (1995).

For comparison purposes, additionally the single-update Gibbs sampler as well
as the integration sampler as implemented in the software package SVPack
have been used, a freeware dynamic link library for the Ox programming lan-
guage (Doornik, 1996) discussed by Kim et al. (1998) and results for the
transformed parameters $\beta$, $\phi$ and $\tau$ have been stated. For each MCMC algo-
rithm, 110,000 iterations have been performed, the first 10,000 iterations as
burn-in period discarded and the output has been thinned by taking every
10th sample. This resulted in a final chain length of 10,000. Table 2.1 gives
an overview of the parameter estimates for $\phi$, $\tau$ and $\beta$ obtained by the four
different samplers and compares the CPU times (all computations were per-
formed on a Pentium III PC) and the IACT.

The IACT (Sokal, 1996), also referred to as the 'inefficiency factor' by Kim
et al. (1998) gives a measure of the efficiency of the MCMC sampler. As the
estimate of the posterior mean of a parameter $x$ is the average of $N$ correlated
samples from a Markov chain, its variance is a factor of IACT larger than
the variance of the sample mean based on the same number of independent
samples, i.e.,

$$\text{var}(\bar{x}_{MC}) = IACT \cdot \frac{\text{var}(x)}{N}.$$ 

Hence, IACT is the number of correlated samples with the same variance-
2.4 Bayesian estimation

reducing power as one independent sample. The estimate of IACT is obtained by multiplying the square of the MC standard error by the number of MCMC iterations $N$ and dividing by the square of the standard deviation of the parameter. The MC standard errors in Table 2.1 are calculated by Geweke’s (1992) method.

The discrepancy in the estimates of the parameter $\beta$ between EKD-AD/RW

Table 2.1: Comparison of Bayesian estimates, IACTs and CPU times (in seconds) using the single-update Gibbs sampler, EKF-AD, EKF-RW, and the integration sampler of SVPack as in Kim et al. (1998)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
<th>MC SE</th>
<th>IACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single-update Gibbs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.9775</td>
<td>0.0106</td>
<td>0.00031</td>
<td>86</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.1598</td>
<td>0.0304</td>
<td>0.00126</td>
<td>172</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.6496</td>
<td>0.0990</td>
<td>0.00126</td>
<td>16</td>
</tr>
<tr>
<td>Time (s)</td>
<td>249</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EKF-AD</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.9797</td>
<td>0.0105</td>
<td>0.00018</td>
<td>28</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.1596</td>
<td>0.0365</td>
<td>0.00068</td>
<td>35</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.8127</td>
<td>0.1414</td>
<td>0.00198</td>
<td>19</td>
</tr>
<tr>
<td>Time (s)</td>
<td>818</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EKF-RW</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.9761</td>
<td>0.0112</td>
<td>0.00016</td>
<td>20</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.1756</td>
<td>0.0352</td>
<td>0.00048</td>
<td>19</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.7995</td>
<td>0.1315</td>
<td>0.00168</td>
<td>16</td>
</tr>
<tr>
<td>Time (s)</td>
<td>832</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Integration sampler</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.9775</td>
<td>0.0105</td>
<td>0.00012</td>
<td>12</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.1575</td>
<td>0.0313</td>
<td>0.00040</td>
<td>16</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.6517</td>
<td>0.0999</td>
<td>0.00105</td>
<td>11</td>
</tr>
<tr>
<td>Time (s)</td>
<td>2311</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and the single-update Gibbs sampler/integration sampler is due to the approx-
2.4 Bayesian estimation

imation error incurred by iterated Laplace approximations.

However, the approximation affects only the constant scaling factor $\mu$, respectively $\beta$, but not the estimates of the persistence of the volatility $\phi$ nor the volatility of log-volatilities $\tau$.

The IACTs of the EKF-AD (and EKF-RW) show a substantial decrease over those of the single-update Gibbs sampler, thus yielding a much improved MCMC technique. Its mixing properties come close to those of the integration sampler with a third of its computation time. However, one should note that the implementation of the integration sampler in SVPack did not use optimised code but rather off the shelf SsfPack algorithms (see Koopman, Shephard and Doornik, 1999) to carry out likelihood evaluations and simulation smoothing. Thus, a speed-up of the integration sampler by about a factor of 3 may well be feasible.

The computation times and IACTs of EKF-AD and EKF-RW are comparable. One should note, though, that the specification of the proposal covariance matrix in the pilot run of EKF-RW required several preliminary experimental runs to get reasonable variance estimates of each parameter.

EKF-AD, on the other hand, yields the posterior modes and asymptotic standard errors of parameter estimates. Moreover, its implementation of the MH algorithm makes efficient use of the availability of the posterior mode and the
2.4 Bayesian estimation

Figure 2.1: Trace plots and kernel density estimates of the marginal posterior distribution of $\beta, \phi, \tau$, obtained by EKF-AD for the basic SV model.

Figure 2.1 shows trace and kernel density plots of the marginal posterior distribution of the model parameters obtained by EKF-AD. These can be compared to Figures 2 and 5 of Kim et al. (1998) for the corresponding trace and kernel density plots of the single-update Gibbs sampler and the integration sampler, respectively.
A menu-driven collection of SPLUS functions, CODA (Best, Cowles and Vines, 1995), is available for analysing the samples obtained from MCMC. CODA can be downloaded from http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml. Besides trace plots and convergence diagnostics based on Cowles and Carlin (1996), CODA calculates statistical summaries of the posterior distributions and kernel density estimates.

Extensive convergence diagnostics based on Cowles and Carlin (1996) were calculated for this chain using the CODA software of Best et al. (1995), a menu-driven collection of SPLUS functions. All parameters passed the Heidelberger and Welch stationarity and halfwidth test. Geweke’s Z-scores for $\phi$, $\tau$, and $\beta$ were reasonably low (-0.99, 1.54, and -0.56, respectively).

It should be pointed out that EKF-AD in conjunction with AD can be readily used to obtain ML estimates and their asymptotic standard deviations. This is easily done by choosing a flat prior, i.e., setting $p(\theta) \propto 1$, and using AD and a quasi-Newton algorithm to find the value of $\theta$ where $\phi(\theta) = -\log \tilde{p}(y \mid \theta)$ is minimized.

The ML-estimates thus obtained for $\phi$, $\tau$, and $\beta$ and their standard deviations (in brackets) are 0.9692 (0.01395), 0.1935 (0.04185) and 0.7029 (0.08607), re-
The same pound/dollar exchange rate data set as fitted for the previous analysis by the basic SV model will be considered here, now applying a 2 independent AR(1) model, as introduced in Chapter 2.2, MODEL 4:

\[
\begin{align*}
    y_t | h_{1,t}, h_{2,t} &= \exp(\mu/2 + h_{1,t}/2 + h_{2,t}/2)u_t, & u_t \overset{iid}{\sim} N(0,1) \\
    h_{1,t+1} | h_{1,t}, \phi, \tau_1^2 &= \phi h_{1,t} + v_{1,t}, & v_{1,t} \overset{iid}{\sim} N(0,\tau_1^2), \\
    h_{2,t+1} | h_{2,t}, \phi_2, \tau_2^2 &= \phi_2 h_{2,t} + v_{2,t}, & v_{2,t} \overset{iid}{\sim} N(0,\tau_2^2).
\end{align*}
\] (2.32)

In the case studies to be carried out in Chapters 3.4.1 and 3.4.2, convergence problems were encountered for the 2 independent AR(1) model when using the all-purpose Gibbs-sampler based Bayesian software package BUGS. Consequently this model was removed from the analysis.

Here, the model will be implemented using the EKF-AD technique as described in Chapter 2.4.1. This can easily be done by now defining the states \( h_t \) as a vector with two components \( h_{1,t} \) and \( h_{2,t} \). Due to the independence of the AR(1) processes, derivations can be carried out analogous to the formulas in Chapter 2.4.1. Whenever more convenient, one can use the independence property and
replace a function \( f(h_t) = f(h_{1,t}, h_{2,t}) \) with the product \( f(h_{1,t}) \times f(h_{2,t}) \). This yields for instance for (2.29):

\[
p(y_t \mid y_{t-1}, \theta) = \int p(y_t \mid h_t, \theta)p(h_t \mid y_{t-1}, \theta) dh_t \\
\approx \int p(y_t \mid h_t, \theta) \tilde{p}(h_t \mid y_{t-1}, \theta) dh_t \\
= \sqrt{2\pi} e^{-\psi_t(y_t, h_t, \theta)} |D^2 \psi_t(y_t, \widehat{h}_t, \theta)|^{-1/2} \\
= \sqrt{2\pi} e^{-\psi_t(y_t, \widehat{h}_1, \theta) (h_{2,t}, \widehat{h}_2, \theta)} |D^2 \psi_t(y_t, \widehat{h}_1, \widehat{h}_2, \theta)|^{-1/2} \\
= \sqrt{2\pi} e^{-\psi_t(y_t, \widehat{h}_1, \theta) (h_{2,t}, \widehat{h}_2, \theta)} |D^2 \psi_t(y_t, \widehat{h}_1, \widehat{h}_2, \theta)|^{-1/2} \\
\times |D^2 \psi_t(y_t, \widehat{h}_1, \widehat{h}_2, \theta)|^{-1/2},
\]

(2.33)

where

\[
\psi_t(y_t, h_t, \theta) = -\log[p(y_t \mid h_t, \theta) \tilde{p}(h_t \mid y_{t-1}, \theta)] \\
= \frac{1}{2} \log(2\pi e^{\mu + h_{1,t} + h_{2,t}}) + \frac{1}{2} e^{-(\mu + h_{1,t} + h_{2,t})} y_t^2 + \frac{1}{2} \log(2\pi \gamma_{1,t}^2) + \\
+ \frac{1}{2\gamma_{1,t}^2} (h_{1,t} - \beta_{1,t})^2 + \frac{1}{2} \log(2\pi \gamma_{2,t}^2) + \frac{1}{2\gamma_{2,t}^2} (h_{2,t} - \beta_{2,t})^2,
\]

\[
\widehat{h}_{1,t} = \arg \min_{h_{1,t}} \psi_t(y_t, h_{1,t}, h_{2,t}, \theta), \quad \widehat{h}_{2,t} = \arg \min_{h_{2,t}} \psi_t(y_t, h_{1,t}, h_{2,t}, \theta), \quad \sigma_{1,t}^2 = |D_{h_{1,t}}^2 \psi_t(y_t, \widehat{h}_{1,t}, \widehat{h}_{2,t}, \theta)|^{-1}, \quad \sigma_{2,t}^2 = |D_{h_{2,t}}^2 \psi_t(y_t, \widehat{h}_{1,t}, \widehat{h}_{2,t}, \theta)|^{-1}
\]

with \( |D_{h_{1,t}}^2 \psi_t(.)| \) and \( |D_{h_{2,t}}^2 \psi_t(.)| \) corresponds to the second-order derivative of \( \psi_t(.) \) with respect to \( h_{1,t} \) and \( \theta = (\mu, \phi, \phi_2, \tau^2, \gamma_{2,t}^2) \).
2.4 Bayesian estimation

Priors for the parameters in the 2 independent AR(1) model are chosen accordingly to the basic SV model with additional prior specifications as follows: The second autoregressive parameter $\phi_2$ is assumed following a vague beta-distribution with $\alpha = \beta = 2$ centered around 0 and a conjugated inverse-gamma prior for $\tau_2^2$, $\tau_2^2 \sim IG(2.5, 0.25)$ has been selected.

Samples from the posterior distribution have been obtained analogous to the previous analysis of the basic SV model example by employing EKD-AD and performing an MCMC analysis by first running the sampler for 110,000 iterations and discarding the first 10,000. Finally, from the remainder a thinned output has been generated by taking every 10th sample, resulting in 10,000 kept iterations.

Table 2.2 presents results in a similar fashion as Table 2.1, including parameter estimates, CPU times and IACTs.

The parameter estimates are quite reasonable, showing that the two volatility factors capture different dynamics of the underlying time series. Furthermore, the IACTs are fairly small, which indicates that the MCMC algorithms perform efficiently.

As in previous examples, convergence diagnostics have been calculated. Again
Table 2.2: Bayesian estimates, IACTs and CPU time (in seconds) using EKF-AD for the 2 independent AR(1) model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>SD</th>
<th>MC SE</th>
<th>IACT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$</td>
<td>0.9855</td>
<td>0.0066</td>
<td>0.00012</td>
<td>31</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.0798</td>
<td>0.0137</td>
<td>0.00026</td>
<td>36</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.7150</td>
<td>0.0769</td>
<td>0.00125</td>
<td>26</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>0.2433</td>
<td>0.1665</td>
<td>0.00194</td>
<td>14</td>
</tr>
<tr>
<td>$\tau_2$</td>
<td>0.1765</td>
<td>0.0227</td>
<td>0.00039</td>
<td>30</td>
</tr>
<tr>
<td>Time (s)</td>
<td></td>
<td></td>
<td></td>
<td>4,941</td>
</tr>
</tbody>
</table>

all parameters passed the Heidelberger and Welch stationarity and halfwidth test. Geweke’s Z-scores are reasonably small in absolute values for the parameters lying in a range from 1.111 for $\tau_2$ to -2.066 for $\phi_2$. Trace and kernel density plots, shown in Figure 2.2 confirm the convergence diagnostics results.

2.5 Summary

In this section, a powerful model which has been extensively applied in recent years to various important time series, like stock market, exchange rate or interest rate data has been introduced. It originates in continuous time and its
2.5 Summary

Figure 2.2: Trace plots and kernel density estimates of the marginal posterior distribution of $\beta, \phi, \phi_2, \tau, \tau_2$, obtained by EKF-AD for the 2 independent AR(1) SV model.

discrete time counterpart can be represented in state space form. Due to the non-existent closed form likelihood, this model has been considered as a great challenge for practitioners and theorists alike.

Various specifications of the stochastic volatility model have presented and estimation methods, classical and Bayesian have been extensively reviewed. A new flexible Bayesian method combining automatic differentiation, Laplace
2.5 Summary

approximation and the extended Kalman filter has been introduced as a further alternative to already existing methods.

By explicitly calculating posterior mode and Hessian matrix using a quasi-Newton method employing automatic differentiation, this method provides an accuracy for the derivations involved up to machine precision, and therefore allows for the design of a suitable proposal density for the Metropolis-Hastings algorithm to sample from the posterior distribution of the parameters. This yields a very efficient MCMC technique with good mixing properties.

Comparisons with highly sophisticated Bayesian estimation methods have been performed and its good mixing properties have been documented by integrated autocorrelation times comparable to those of alternative highly efficient samplers. These similar autocorrelation times however can be achieved with higher computational speed.

Using exchange rate data, it has been shown how to apply this method to a financial data set that has often been analyzed in the literature. This gives a way of evaluating this method by comparing the results to those from other studies.
Chapter 3

DIC and other model checking criteria
3.1 Introduction

The SV model is a typical example of a hierarchical model, in which the number of unknowns exceeds the number of observations. In the basic SV model the unknowns include the parameters \( z = (\phi, \mu, \tau) \) and the unknown states \( (h_1, \ldots, h_n) \). The number of nominally free parameters in the model could be the number of model parameters \( (3) \) if the focus is on \( z \) only or the number of states plus the number of model parameters \( (n+3) \) if the focus is on \( z, h_1, \ldots, h_n \). The lack of definition of effective number of parameters precludes the use of the well-known Akaike (AIC, Akaike, 1973) or Bayesian (BIC, Schwarz, 1978) information criterion for model comparison over different models.

The progress in Bayesian posterior computation due to Markov chain Monte Carlo (MCMC) methods has made it possible to fit increasingly complex statistical models and entailed the wish to determine the best-fitting model in a potentially huge class of candidates. Thus, it has become more and more important to develop efficient model selection criteria. A recent proposal by Spiegelhalter, Best, Carlin and van der Linde (2002) is the Deviance Information Criterion (DIC), a Bayesian version or generalization of AIC, and related also to BIC. Similar to AIC and BIC, it trades off a measure of model adequacy against a measure of complexity. DIC is easy to calculate and applicable to a
3.1 Introduction

wide range of statistical models. It is based on the posterior distribution of the log-likelihood or the deviance, following the original suggestion of Dempster (1974) for model choice in the Bayesian framework.

In this chapter it will be shown that DIC provides an efficient and straightforward approach to defining the effective number of parameters and to identifying the most appropriate model.

This model comparison criterion has already been applied successfully to complex models in the field of medical statistics (Zhu and Carlin, 2000).

The aim here is to introduce DIC to the financial modelling community and show how to use it for the family of stochastic volatility (SV) models.

Indeed, many model checking criteria (Carlin and Louis, 2000; Gelman, Carlin, Stern and Rubin, 2003; Gilks et al., 1996; Key, Pericchi and Smith, 1999) have been proposed and discussed before the development of DIC. While Bayes factors (e.g. Kass and Raftery, 1995) have been viewed for many years as the only correct way to carry out Bayesian model comparison, they have come under increasing criticism of late (Kass and Raftery, 1995; Lavine and Schervish, 1999).

One serious drawback is that they are not well-defined when using improper priors which is typically the case in practice when employing non-informative
3.1 Introduction

priors. This led to modifications, such as the partial Bayes factor (O'Hagan, 1991), the intrinsic Bayes factor (Berger and Pericchi, 1996), and the fractional Bayes factor (O'Hagan, 1994). These modifications suffer from more or less arbitrary choices of training samples, weights for averaging training samples, and fractions, respectively.

For specifying Bayesian stochastic volatility (SV) models, however, informative and thus proper prior distributions are usually employed and Bayes factors are well defined.

Nonetheless, the number of unknown parameters in Bayesian SV models is large (exceeding the number of observations) because of the latent volatilities. Calculation of the Bayes factor for comparing any two models requires the marginal likelihoods and thus a marginalization over the parameter vectors in each model. If the dimension of the parameter space is large, these implicit, extremely high-dimensional integration problems pose a formidable computational challenge.

In the context of SV models, Kim et al. (1998) and Chib, Nardari and Shephard (2002) have shown how to compute Bayes factors using the marginal likelihood approach of Chib (1995) and evaluating the marginal likelihood at the posterior mean using particle filtering (Kitagawa, 1996; Pitt and Shephard, 1999a; Doucet, de Freitas and Gordon, 2001).

Still, it remains a computationally intensive task and is not a particularly user-
friendly tool for practising statisticians. In their review of MCMC methods for computing Bayes factors, Han and Carlin (2001, p. 1132) conclude that “all of the methods ... discussed require substantial time and effort (both human and computer) for a rather modest payoff, namely a collection of posterior model probability estimates ... As a result, one might conclude that none of the methods considered here are appropriate for everyday, rough and ready model comparison, and instead one may search for more computationally realistic alternatives”.

A well-known estimate of the marginal likelihood developed by Newton and Raftery (1994) is the harmonic mean of the likelihood values. It is easy to compute and simulation-consistent but not stable because the inverse likelihood does not possess a finite variance (Chib, 1995).

Other shortcuts to the calculation of Bayes factors that avoid multidimensional integration through large sample approximations of $-2 \ln(\text{Bayes factor})$ include the familiar BIC, also referred to as Schwarz Criterion (Schwarz, 1978), and the related penalized likelihood ratio model choice criterion, AIC.

Both, BIC and AIC have often been used in the literature to compare alternative models. AIC has been defined as $-2 \ln L(\hat{\theta}) + kn$, with $L$ being the likelihood function and $\hat{\theta}$ the maximum likelihood estimate of the parameter
of interest, \( n \) presents the number of parameters in the fitted model, \( k \) is set to 2. Alternatively, choosing \( k = \ln(n) \) leads to the definition of BIC.

Either criterion requires the specification of the number of free parameters in each model. If one considers a non-hierarchical Bayesian model with parameter \( \theta \), a flat prior would correspond to a flexible and thus complex model, whereas a tight prior constrains the model. The classical definition of model complexity as the 'number of unknown parameters' could thus be considered as a special case corresponding to a non-informative prior.

However, for a complex hierarchical model the specification of its dimensionality is rather arbitrary. This is typically the case for a SV model, where the parameters are augmented by the \( n \) latent volatilities with \( n \) being the sample size. As these are not independent but exhibit a Markovian dependence structure, they cannot be counted as \( n \) additional free parameters. Thus, neither BIC nor AIC are applicable for SV model comparison.

As detailed in the following subsection, DIC avoids this dilemma by using a complexity measure for the effective number of parameters that is based on an information-theoretic argument. This quantity is readily obtained from a MCMC analysis which makes algebraic forms and large sample approximations obsolete.

Most importantly, when modelling financial time series using SV models, the
philosophy of Box (1976) in believing that “all models are wrong but some are useful” has been adopted. Thus, it would be hard to specify prior model probabilities necessary for the calculation of posterior model probabilities. By using DIC as a formal approach to model selection, combining a measure of fit and complexity, this need can be avoided.

However, one has to caution in general against basing model choice solely on information criteria, as many other factors such as the model’s inherent plausibility, the robustness of its inferences and model diagnostics (as outlined in Kim et al., 1998, Section 4.2 and Spiegelhalter et al., 2002, Section 6) need to be taken into account.

In many instances, when none of the models is clearly superior, model averaging (Hoeting, Madigan, Raftery and Volinsky, 1999) might be more appropriate. Whether DIC can be used as a basis for model averaging is still an open question.
3.2 The Deviance Information Criterion

Assume, in general, that the distribution of the data, \( y = (y_1, \ldots, y_n) \), depends on a \( p \)-dimensional parameter vector \( \theta \). In contrast to Chapter 2, in the context of an SV model, \( \theta \) now encompasses the parameter vector \( (\mu, \phi, \tau^2) \) and the vector of log-volatilities \( h_1, \ldots, h_n \) referring to the conditional likelihood \( f(y \mid z, h) \). This is in contrast to Chapter 2 with \( \theta = z \) and referring to the marginal likelihood \( f(y \mid z) \).

From a frequentist point of view, model assessment is based on the deviance, the difference in the log-likelihoods between the fitted and the saturated model.

The saturated model refers to the model with as many parameters as observations, that yields a perfect fit to the data. By analogy, Dempster (1974) suggested examining the posterior distribution of the classical deviance defined by

\[
D(\theta) = -2 \ln f(y \mid \theta) + 2 \ln g(y) \tag{3.1}
\]

for Bayesian model selection. Here, \( f(y \mid \theta) \) is the likelihood function, i.e. the conditional joint probability density function of the observations given the un-
3.2 The Deviance Information Criterion

known parameters, and \( \ln g(y) \) denotes a fully specified standardizing term that is a function of the data alone (in the following applications, \( g(y) = 1 \)).

Dempster (1974) proposed comparing plots and potential summaries such as the posterior mean of \( D(\theta) \) and Spiegelhalter et al. (2002) followed these suggestions in the development of DIC as a model choice criterion. Based on the posterior distribution of \( D(\theta) \), DIC consists of two components, a term that measures goodness-of-fit and a penalty term for increasing model complexity:

\[
\text{DIC} = \bar{D} + p_D. \tag{3.2}
\]

1. The first term, a Bayesian measure of model fit, is defined as the posterior expectation of the deviance

\[
\bar{D} = E_{\theta | y}[D(\theta)] = E_{\theta | y}[-2 \ln f(y|\theta)]. \tag{3.3}
\]

The ‘better’ the model fits the data, the larger are the values for the likelihood. \( \bar{D} \) which is defined via \(-2\) times log-likelihood therefore attains smaller values for ‘better’ models.

2. The second component measures the complexity of the model by the effective number of parameters, \( p_D \), defined as the difference between
3.2 The Deviance Information Criterion

the posterior mean of the deviance and the deviance evaluated at the posterior mean \( \hat{\theta} \) of the parameters:

\[
p_D = \tilde{D} - D(\tilde{\theta}) = E_{\theta|y}[D(\theta)] - D(E_{\theta|y}[\theta]) = E_{\theta|y}[-2\ln f(y | \theta)] + 2\ln f(y | \tilde{\theta}).
\]

(3.4)

By defining \(-2\ln f(y | \theta)\) as the residual information in the data \( y \) conditional on \( \theta \) and interpreting it as a measure of uncertainty, (3.4) shows that \( p_D \) can be regarded as the expected excess of the true over the estimated residual information in data \( y \) conditional on \( \theta \). This means, \( p_D \) can be interpreted as the expected reduction in uncertainty due to estimation.

Rearranging (3.4), one gets \( \tilde{D} = D(\tilde{\theta}) + p_D \). Thus, the DIC defined in (3.2) can be reexpressed as

\[
\text{DIC} = D(\tilde{\theta}) + 2p_D
\]

(3.5)

which can be interpreted as a classical 'plug-in' measure of fit plus a measure of complexity. Therefore, the Bayesian measure of fit \( \tilde{D} = D(\tilde{\theta}) + p_D \) already includes a penalty term for model complexity and could thus be better thought
of as a measure of 'model adequacy' rather than pure goodness-of-fit.

Spiegelhalter et al. (2002) gave an asymptotic justification of DIC in the case where the number of observations $n$ grows with respect to the number of parameters $p$ and where the prior is non-hierarchical and completely specified (i.e., without hyperparameters). In this situation, $AIC = D(\hat{\theta}) + 2p$, where $\hat{\theta}$ denotes the maximum likelihood (ML) estimate. This is the same formula as (3.5) but with the posterior mean $\bar{\theta}$ substituted by the ML estimate $\hat{\theta}$. Thus, DIC can be seen as a generalization of AIC and also compared to the Schwarz information criterion $BIC = -2 \ln f(y|\theta) + p \ln n$. In the special case where the prior is flat, a case that corresponds to a frequentist analysis, AIC equals DIC because the ML estimate coincides with the posterior mean.

In the context of normal linear regression with uncertainty in the choice of regressors, George and Foster (2001) developed empirical Bayes alternatives to penalized likelihood criteria such as AIC and BIC and Fernandez, Ley and Steel (2001) point out links of Bayes factors with classical information criteria and provided a unifying framework.

Since DIC depends on the chosen focus, (re-)parameterisations, prior distributions and the data, an analysis using DIC for the situation in Chapter 2 may lead to different conclusions regarding model comparison. Using MCMC, the
3.2 The Deviance Information Criterion

calculation of DIC can be accomplished for the case in Chapter 2 by integrating out the states beforehand.

By combining the likelihood with the prior almost always additional restrictions (by obtaining information from the prior) on the parameter space are being imposed. This results in effectively reducing the degrees of freedom, which corresponds to a reduction in effectively free parameters, a natural indicator for reduced model complexity. By applying a logarithmic loss function, Spiegelhalter et al. (2002) give a decision-theoretic justification for DIC and show that DIC approximately describes the expected posterior loss when adopting a particular model. For additional asymptotic properties of $p_D$ and $\tilde{D}$ the interested reader is referred to Spiegelhalter et al. (2002).

In striking contrast to calculating Bayes factors, computing DIC via MCMC is almost trivial. An estimate of $\tilde{D}$ is easily calculated from the MCMC output by monitoring $D(\theta)$ and then taking the sample mean of the simulated values of $D(\theta)$. The effective number of parameters $p_D$ can be obtained by evaluating $D(\theta)$ at the sample average of the simulated values of $\theta$ and subtracting this plug-in estimate of the deviance from the estimate of $\tilde{D}$ (see also section 3.4).

So far, no efficient method has been developed for calculating reasonably accu-
rate MC standard errors of DIC. Zhu and Carlin (2000) explore this problem, but their approach using the multivariate delta method yields poor results. Their final recommendation is a “brute force” approach, which is simply replicating the calculation of DIC some N times and estimating \( \text{VAR}(\text{DIC}) \) by its sample variance

\[
\hat{\text{VAR}}(\text{DIC}) = \frac{1}{(N - 1)} \sum_{k=1}^{N} (\text{DIC}_k - \bar{\text{DIC}})^2.
\]

Although this is a painfully time-consuming approach, it at least gives an indication of the inherent variability of DIC.

### 3.3 Marginal Likelihood and Harmonic mean

To carry out a comparison of the performance of DIC with that of Chib's marginal likelihood method and the harmonic mean in the following case studies, it is worthwhile to first review Chib's method for calculating the marginal likelihood and Newton and Raftery's (1994) method for estimating the marginal likelihood by the harmonic mean of the sampled likelihood values.
By definition, the marginal likelihood \( m(y) \) is the integral of the likelihood function with respect to the prior density \( \pi(z) \)

\[
m(y) = \int f(y|z)\pi(z)dz
\]  

(3.6)

with \( z \) denoting the vector of parameters in the model. As solving this integral would require high-dimensional integration, Chib (1995) suggested evaluating the marginal likelihood by rearranging Bayes's theorem

\[
m(y) = \frac{f(y|z)\pi(z)}{\pi(z|y)}
\]

where \( \pi(z|y) \) denotes the posterior probability density function of \( z \). Thus, the log-marginal likelihood which is referred to as \( \ln L \) in the following, can be calculated by

\[
\ln L = \ln m(y) = \ln f(y|z) + \ln \pi(z) - \ln \pi(z|y),
\]  

(3.7)

where \( z \) is an appropriately selected high density point (here, simply, the posterior mean \( \bar{z} \) has been used). The first term on the right side of (3.7) is the log-likelihood evaluated at the posterior mean of the parameter vector \( z \) (marginalized over the latent volatilities \( h_t \)) and the second term is the log
3.3 Marginal Likelihood and Harmonic mean

prior density evaluated at $\tilde{z}$. The third quantity involves the posterior density which is only known up to a normalizing constant. However, an approximation can be obtained by using a multivariate kernel density estimate as suggested in Kim et al. (1998) (see also Silverman, 1986, Chapter 4) which is based on the posterior MCMC sample of $z$.

It has been previously mentioned that the log-likelihood function $\ln f(y|z)$ has no analytical form for the SV model as it is marginalized over the latent states $h_1, \ldots, h_n$, and this is why the exact maximum likelihood method is extremely difficult to implement.

However, it is possible to approximate the likelihood by making use of the so-called particle filter method. Important contributions in this area include Gordon, Salmond and Smith (1993), Kitagawa (1996), and Pitt and Shephard (1999a). By successive conditioning, the log-likelihood $\ln f(y|z)$ can be decomposed into

$$
\ln f(y|\tilde{z}) = \ln f(y_1|\tilde{z}) + \sum_{t=1}^{n-1} \ln f(y_{t+1}|Y_t, \tilde{z})
$$

(3.8)

where $Y_t = (y_1, \ldots, y_t)$ collects the available data up to time $t$. Taking the latent volatilities into account, each one-step-ahead prediction density has a
3.3 Marginal Likelihood and Harmonic mean

mixture representation as

\[ f(y_{t+1}|Y_t, \tilde{z}) = \int f(y_{t+1}|h_{t+1}, Y_t, \tilde{z}) f(h_{t+1}|Y_t, \tilde{z}) dh_{t+1} \]

\[ = \int f(y_{t+1}|h_{t+1}, Y_t, \tilde{z}) \left[ \int f(h_{t+1}|h_t, \tilde{z}) f(h_t|Y_t, \tilde{z}) dh_t \right] dh_{t+1}. \]

Then \( f(y|\tilde{z}) \) can be estimated by

\[ \frac{1}{M} \sum_{i=1}^{M} f(y_{t+1}|h_{t+1}^{(i)}) \]

where \( h_{t+1}^{(i)}|h_t^{(i)} \) is drawn from the state equation (2.5) given samples \( h_t^{(i)} \), the so-called filtered particles, from \( f(h_t|Y_t, \tilde{z}) \).

Here, Kitagawa’s algorithm for particle filtering has been utilized. This algorithm is applicable to a broad class of nonlinear non-Gaussian multi-dimensional state space models of the form

\[ \begin{align*}
  y_t &= H(x_t, u_t), \\
  x_t &= F(x_{t-1}, v_t),
\end{align*} \tag{3.9} \]

where \( x_t \) is a \( k \)-dimensional state vector (here, \( x_t = h_t \) is the one-dimensional log-volatility), \( v_t \) is an \( l \)-dimensional white noise sequence with density \( q(v) \), \( u_t \) is a one dimensional white noise sequence with density \( r(u) \) and assumed uncorrelated with \( \{v_s\}_{s=1}^{n} \), and \( H \) and \( F \) are possibly nonlinear functions.
Let \( u_t = G(y_t, x_t) \) and \( G' \) be the derivative of \( G \) as a function of \( y_t \). The density of the initial state vector is assumed to be \( p_0(x) \).

All the steps involved in Kitagawa's algorithm can now be summarized as follows:

1. Generate \( M \) \( l \)-dimensional particles from \( p_0(x) \), \( f_0^{(j)} \) for \( j = 1, \ldots, M \).

2. Repeat the following steps for \( t = 1, \ldots, n \).
   
   (a) Generate \( M \) \( l \)-dimensional particles from \( q(v) \), \( v_t^{(j)} \) for \( j = 1, \ldots, M \).
   
   (b) Compute \( p_t^{(j)} = F(f_{t-1}^{(j)}, v_t^{(j)}) \) for \( j = 1, \ldots, M \).
   
   (c) Compute \( \alpha_t^{(j)} = r(G(y_t, p_t^{(j)})) \) for \( j = 1, \ldots, M \).
   
   (d) Resample \( \{p_t^{(j)}\}_{j=1}^M \) to get \( \{f_t^{(j)}\}_{j=1}^M \) with probabilities proportional to \( \{r(G(y_t, p_t^{(j)})) \times |G'(y_t, p_t^{(j)})|\}_{j=1}^M \).

It can be seen that almost all SV models to be analysed in section 3.4 can be rewritten in the state space form (3.9); hence it is straightforward to modify the preceding algorithm to reconcile it with the SV model structure.

The only exception is MODEL 5, which violates the assumption of no correlation between \( u_t \) and \( v_{t+1} \). But, although the correlation between \( u_t \) and \( v_{t+1} \) makes Kitagawa's algorithm not directly applicable, a simple rewrite of this
3.3 Marginal Likelihood and Harmonic mean

model gives

\[
y_t | h_t = \exp(h_t/2)u_t, \quad u_t \overset{iid}{\sim} N(0,1),
\]

\[
h_t | h_{t-1}, y_{t-1}, \mu, \phi, \tau^2 = \mu + \phi(h_{t-1} - \mu) + \rho \tau \exp(-0.5h_{t-1})y_{t-1} + \tau \sqrt{1 - \rho^2} w_t,
\]

where \( u_t \overset{iid}{\sim} N(0,1) \) and \( \text{corr}(u_t, w_t) = 0 \). Based on the new representation, steps 2(a) and 2(b) in Kitagawa’s algorithm can be modified by:

2(a) generate \( M \) particles, called \( v_t^{(j)}, j = 1, \ldots, M \), from a normal distribution with mean \( \rho \tau \exp(-0.5h_{t-1})y_{t-1} \) and variance \( \tau^2(1 - \rho^2) \);

2(b) obtain \( M \) particles by setting

\[
p_t^{(j)} = F(f_t^{(j)}, v_t^{(j)}), \quad j = 1, \ldots, M.
\]

\( p_t^{(j)} \) can be regarded as independent draws from \( p(h_t|y_{t-1}) \).

It should be pointed out that more efficient particle filter algorithms are available. An example is the auxiliary particle filter introduced by Pitt and Shephard (1999a); see the implementation of this particle filter algorithm in Kim
et al. (1998), Pitt and Shephard (1999b), and Chib et al. (2002, 2004) in the context of SV models. Experience suggests that by choosing $M = 50,000$ for Kitagawa's algorithm one obtains very similar results to the auxiliary particle filter method with $M = 2,500$.

Newton and Raftery (1994) suggested the calculation of approximate Bayes factors for model comparison using the harmonic mean of the sampled likelihood values as a simulation-consistent estimator of the required marginal likelihood. Let $\theta$ denote the parameter vector (augmented by latent volatilities), that is $\theta = (z, h_1, \ldots, h_n)$, as in Chapter 3.2. Similar to (3.6), the marginal likelihood $m(y)$ can be expressed as

$$m(y) = \int f(y|\theta)f(\theta)d\theta$$

where $f(\theta)$ denotes the joint prior density function of $\theta$.

The importance sampling method for evaluating this integral is to generate a sample $\{\theta^{(i)}; i = 1, \ldots, M\}$ from a so-called importance sampling density $f^*(\theta)$.

Under quite weak assumptions, a simulation-consistent estimate of $m(y)$ is
3.3 Marginal Likelihood and Harmonic mean

given by

\[ \hat{m}(y) = \frac{\sum_{i=1}^{M} w_i f(y|\theta^{(i)})}{\sum_{i=1}^{M} w_i}, \quad (3.10) \]

where \( w_i = \frac{f(\theta^{(i)})}{f^*(\theta^{(i)})} \).

The Gibbs sampler produces a sample \( \theta^{(i)} \) approximately drawn from the posterior density \( f^*(\theta) = f(\theta|y) = \frac{f(y|\theta)f(\theta)}{m(y)} \). Using these \( \theta^{(i)} \) in (3.10) yields the harmonic mean estimator of \( m(y) \):

\[ \hat{m}_{hm}(y) = \left\{ \frac{1}{M} \sum_{i=1}^{M} \frac{1}{f(y|\theta^{(i)})} \right\}^{-1}. \quad (3.11) \]

Here, \( \hat{m}_{hm}(y) \) converges almost surely to the correct value \( m(y) \) as \( M \) goes to \( \infty \), but it does not, in general, satisfy a Gaussian central limit theorem as \( 1/f(y|\theta) \) is often not square integrable with respect to the posterior distribution. Thus, a few outlying values \( \theta^{(i)} \) with small likelihood values can have a large effect on the estimate. For this reason, Newton and Raftery (1994) also proposed modified estimators that are much more stable than the straight harmonic mean that will be used here.
3.4 Case studies

The case studies in this section encompass a simulation and an empirical study in subsections 3.4.1 and 3.4.2. They deal with parameter estimation of several SV models, most of them described in section 2.2 using Bayesian MCMC methods, and model comparison based on DIC and related methods.

A previously performed case study in subsection 2.4.2 has presented an analysis of a SV model estimation based on a new Bayesian computation using the extended Kalman filter. The performance of this technique has also been compared with other well-sophisticated Bayesian MCMC methods.

3.4.1 A simulation study

The main objective of this simulation study is to investigate whether DIC is capable of identifying the true model from which the data are generated. It needs to be remarked that regarding the interpretation of DIC as a model comparison criteria this identification scenario is strictly spoken not necessary, it has been followed, however common economic practice. Chib's marginal likelihood and the harmonic mean estimate for each model within the set of competing models have also been calculated.

First of all, it is necessary to stress, what has been pointed out by Spiegelhalter
et al. (2002, rejoinder) about the unprejudiced use of model comparison criteria. They caution against using the Bayes factor (or marginal likelihood) as a 'gold standard' against which to assess DIC. The Bayes factor addresses how well the prior has predicted the observed data, whereas DIC addresses how well the posterior might predict future data generated by the same mechanism that gave rise to the observed data. Thus, these criteria cannot in general be expected to lead to the same conclusions as they are designed to answer different questions. Especially for the practical selection of financial time series models, the consideration of this posterior predictive point of view, justified by DIC seems to be potentially more relevant.

A dataset comprising 2,000 observations has been simulated from a SV model that includes a jump component as described in Model 6. The data are plotted in the first panel of Figure 3.1. This SV + jumps model is very similar to the one proposed in the simulation analysis by Chib et al. (2002). The BUGS (Bayesian Inference Using Gibbs Sampling) software package (Spiegelhalter, Thomas, Best and Gilks, 1996), available online at

http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml

has been used for posterior computation. BUGS is an easy to learn and easy to use Bayesian software package that implements the Gibbs sampler for gen-
3.4 Case studies

Figure 3.1: Time Series Plots for Simulated Data, S&P100, VIX, Logarithm of Absolute Value of S&P100 Returns

Operating samples from a Markov chain whose equilibrium distribution is the posterior distribution. As demonstrated by Meyer and Yu (2000), it can be applied to fit stochastic volatility models. Although more efficient Markov Chain Monte Carlo techniques exist for fitting SV models (Kim et al., 1998), the use of BUGS is highly attractive due to the ease of implementation.

In the following, the choice of priors and the detailed implementation of the
DIC in WinBUGS will be described, and results of the simulation analysis will be presented.

Eight different stochastic volatility models (MODEL 1 to MODEL 8) have been fitted to the simulated data including the true model from which the data are generated (MODEL 6).

For the parameters $\phi$ and $\tau^2$ of the basic SV model, the following prior specifications of Kim et al. (1998) are employed: $\tau^2 \sim \text{Inverse-Gamma}(2.5, 0.025)$ which has a mean of 0.167 and a standard deviation of 0.024. Defining $\phi = 2\phi^* - 1$, Kim et al. (1998) specified a beta-distribution with parameters 20 and 1.5 for $\phi^*$ which implies a mean of 0.86 and a standard deviation of 0.11. Following Kim et al. (1998), an informative but reasonably flat prior distribution for the parameter $\mu$, a normal distribution with mean $-10$ and variance 25 has been chosen.

For $\alpha$ in MODEL 2 a normal distribution with mean parameter $\mu_\alpha = 0$ and variance $\sigma^2_\alpha = 10$ is specified.

For MODEL 3 the same prior for $\phi$ as for the basic SV model is chosen and the prior for $\psi$ is centered around zero using a uniform distribution on $[-1,1]$.

In MODEL 4 again the same prior for $\phi$ as for the basic SV model is used and a vague prior for $\phi_2$ is centered around zero using a beta distribution with
The correlation parameter $\rho$ in MODEL 5 is assumed to be a priori uniformly distributed with support between -1 and 1.

As the parameter $\beta$ in MODEL 6 is assumed to be small a priori, an informative normal distribution with hyperparameters $\mu_\beta = 0$ and $\sigma_\beta^2 = 0.2$ is used. The parameter $q_t$ represents the frequency of a jump occurrence with a Bernoulli distribution with parameter $\kappa$. Following Chib et al. (2002), a Beta(2,100) prior distribution is specified which implies a mean of 0.02 and suggests that a priori one would observe on average one jump in approximately every 50th observation. Finally, as in Chib et al. (2002), it is assumed that $\ln(\delta)$ follows a normal prior distribution with mean -3.07 and variance 0.149.

A well-known alternative to the direct fitting of many symmetric but non-normal error distributions is through scale mixtures of normals (Andrews and Mallows, 1974). Thus, in MODEL 8 the alternative mixture representation of a $t$-distribution by

$$ y_t \sim N(0, \exp(h_t)/w_t) $$

$$ w_t \sim \frac{1}{\nu} \chi^2_\nu = \text{Gamma}(\frac{\nu}{2}, \frac{\nu}{2}). $$
3.4 Case studies

has been used.

Here, a uniform prior distribution for $\nu$ on $[2,128]$ as in Chib et al. (2002) is chosen.

The software package used for the analyses in 3.4.1 and 3.4.2 is WinBUGS, which is the BUGS version operating under Windows. A DIC module which automatically calculates values for DIC and related parameters is implemented in the latest WinBUGS version.

However, even without the DIC module, DIC is easily obtained from any MCMC output.

The first part of DIC, $\bar{D}$, is easily calculated using the MCMC output $\theta^{(i)}, i = 1, \ldots, N$. Simply calculate $D(\theta^{(i)})$ for $i = 1, \ldots, N$ and estimate $\bar{D}$ by the sample mean $(1/N) \sum_{i=1}^{N} D(\theta^{(i)})$. In practice, using BUGS, this is accomplished by adding the variable $D(\theta)$.

For the second part, the effective number of parameters $p_D$, one only needs to evaluate $D(\theta)$ at the sample posterior mean $\bar{\theta} = (1/N) \sum_{i=1}^{N} \theta^{(i)}$. WinBUGS offers several useful convergence-checking criteria available in an attached CODA (Convergence Diagnosis and Output Analysis Software for Gibbs sampling output; Best, Cowles and Vines, 1995) module running, for example under S-plus. It is necessary to check whether convergence has been achieved because it is crucial that the sample is taken from the stationary distribution.
3.4 Case studies

The CODA package consists of a selection of model checking criteria, one of which is the Heidelberger and Welch test (Heidelberger and Welch, 1983). All the results reported in this work are based on samples which have passed the Heidelberger-Welch convergence test for all parameters.

In Table 3.1 means and standard errors (numbers in parentheses) of both prior and posterior distributions are reported, as well as the computing time to generate 100 iterations for each of the eight models. The numbers in square brackets are the true values of the parameters. In Table 3.2 Chib’s marginal likelihood, harmonic mean, and DIC together with $\bar{D}$ and $p_D$ are reported for each of the eight models as well as their associated rankings by each criterion.

For SV MODELS 1-5, after a burn-in period of 50,000 iterations and a follow-up period of 250,000, every 20th iteration has been stored. Due to higher posterior correlations amongst the parameters and thus slower convergence of the Gibbs sampler in the remaining models, a burn-in period of 100,000 iterations, a follow-up period of 900,000 has been chosen, and every 40th iteration has been stored. All calculations were performed on a Pentium-III PC, 550 MHz, running the WinBUGS 131 version updated with the DIC tool.
### 3.4 Case studies

#### Table 3.1: Parameter Estimates for Simulated Data

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mod 1</th>
<th>Mod 2</th>
<th>Mod 3</th>
<th>Mod 4</th>
<th>Mod 5</th>
<th>Mod 6</th>
<th>Mod 7</th>
<th>Mod 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>(10.0)</td>
<td>(10.0)</td>
<td>(10.0)</td>
<td>(10.0)</td>
<td>(10.0)</td>
<td>(10.0)</td>
<td>(10.0)</td>
<td>(10.0)</td>
</tr>
<tr>
<td>$\phi$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td>Posterior</td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.11)</td>
</tr>
<tr>
<td>$\tau$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>1.128</td>
<td>1.130</td>
<td>1.205</td>
<td>2.803</td>
<td>1.130</td>
<td>3.950</td>
<td>3.988</td>
<td>3.244</td>
</tr>
<tr>
<td>Posterior</td>
<td>(0.617)</td>
<td>(0.622)</td>
<td>(0.657)</td>
<td>(0.646)</td>
<td>(0.624)</td>
<td>(0.438)</td>
<td>(0.450)</td>
<td>(0.462)</td>
</tr>
<tr>
<td>$\alpha$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Posterior</td>
<td>(3.16)</td>
<td>(3.16)</td>
<td>(3.16)</td>
<td>(3.16)</td>
<td>(3.16)</td>
<td>(3.16)</td>
<td>(3.16)</td>
<td>(3.16)</td>
</tr>
<tr>
<td>$\psi$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Posterior</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Posterior</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
</tr>
<tr>
<td>$\tau_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
<td>0.38</td>
</tr>
<tr>
<td>Posterior</td>
<td>(0.16)</td>
<td>(0.16)</td>
<td>(0.16)</td>
<td>(0.16)</td>
<td>(0.16)</td>
<td>(0.16)</td>
<td>(0.16)</td>
<td>(0.16)</td>
</tr>
<tr>
<td>$\rho$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
<td>0.90</td>
</tr>
<tr>
<td>Posterior</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
<td>(0.58)</td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Posterior</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
<td>(0.45)</td>
</tr>
<tr>
<td>$\kappa$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Posterior</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>$\delta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
</tr>
<tr>
<td>Posterior</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
<td>(0.02)</td>
</tr>
<tr>
<td>$\nu$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Posterior</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time (seconds)</td>
<td>4.30</td>
<td>4.84</td>
<td>7.54</td>
<td>7.38</td>
<td>31.31</td>
<td>36.45</td>
<td>31.29</td>
<td>42.59</td>
</tr>
</tbody>
</table>
Table 3.2: Chib's Marginal Likelihood, Harmonic Mean, DIC for Simulated Data

<table>
<thead>
<tr>
<th>Model</th>
<th>ln ( L ) Value</th>
<th>Ranking</th>
<th>Harmonic Mean Value</th>
<th>Ranking</th>
<th>DIC Value</th>
<th>Ranking</th>
<th>( \bar{D} ) Value</th>
<th>Ranking</th>
<th>( p_D ) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mod 1</td>
<td>6472.67</td>
<td>7</td>
<td>6888.43</td>
<td>7</td>
<td>-13442.5</td>
<td>7</td>
<td>-14002.5</td>
<td>560.0</td>
<td></td>
</tr>
<tr>
<td>Mod 2</td>
<td>6467.51</td>
<td>8</td>
<td>6882.43</td>
<td>8</td>
<td>-13441.9</td>
<td>8</td>
<td>-14003.4</td>
<td>561.5</td>
<td></td>
</tr>
<tr>
<td>Mod 3</td>
<td>6474.38</td>
<td>5</td>
<td>6890.43</td>
<td>6</td>
<td>-13463.3</td>
<td>4</td>
<td>-14040.7</td>
<td>577.4</td>
<td></td>
</tr>
<tr>
<td>Mod 4</td>
<td>6495.51</td>
<td>4</td>
<td>6948.43</td>
<td>4</td>
<td>-13496.2</td>
<td>3</td>
<td>-14102.4</td>
<td>606.2</td>
<td></td>
</tr>
<tr>
<td>Mod 5</td>
<td>6472.82</td>
<td>6</td>
<td>6901.42</td>
<td>5</td>
<td>-13453.9</td>
<td>5</td>
<td>-14018.9</td>
<td>565.0</td>
<td></td>
</tr>
<tr>
<td>Mod 6</td>
<td>6569.16</td>
<td>1</td>
<td>7172.00</td>
<td>1</td>
<td>-14450.0</td>
<td>1</td>
<td>-14582.7</td>
<td>132.7</td>
<td></td>
</tr>
<tr>
<td>Mod 7</td>
<td>6548.27</td>
<td>2</td>
<td>7102.92</td>
<td>2</td>
<td>-14362.0</td>
<td>2</td>
<td>-14485.3</td>
<td>123.3</td>
<td></td>
</tr>
<tr>
<td>Mod 8</td>
<td>6517.44</td>
<td>3</td>
<td>6949.62</td>
<td>3</td>
<td>-13448.0</td>
<td>6</td>
<td>-14096.1</td>
<td>648.1</td>
<td></td>
</tr>
</tbody>
</table>

From the examination of these two tables it should be first noted that the correct model (Model 6) is estimated by MCMC with reasonably accurate results for all six parameters. Moreover, the correct model provides the smallest value for DIC as well as for the posterior mean of the deviance despite the fact that the effective number of parameters is not the smallest. Only a slightly larger value of DIC for the SV + jumps model without lagged observations (Model 7) has been obtained. This is because differences between this model and the correct model are very small. Not surprisingly, this model is ranked second by DIC. All the other models clearly perform worse. For example, compared with DIC values of -14,450 and -14,362 for the two models with jumps, the basic SV model provides a DIC value of -13,442.5. In fact, the DIC margins among all the models excluding the jump models are reasonably small. For example, DIC of the third best model differs from that of the worst
model by 54.3 whereas the difference between the second best and the third best is 865.8. Moreover, the effective number of parameters is much larger for all the models except the jump models and none of these models fits the data as well as the jump models, indicated by the highest values for the posterior mean of the deviance. Not surprisingly, the higher values of $\hat{D}$ and $p_D$ add up to the higher DIC values.

**Model 4**, with two independent AR(1) components gives a relatively good fit being ranked the best fitting after the jump models by DIC and the best fitting after the jump and $t$-SV models by Chib’s marginal likelihood. It can thus be considered as a good alternative to using SV models with jumps.

Another interesting result emerging from these two tables is the performance of DIC relative to Chib’s marginal likelihood and the harmonic mean. Neither DIC nor the harmonic mean provides the same model ranking as Chib’s marginal likelihood but the differences are not substantial. Differences between the two marginal likelihood methods and DIC are not surprising as the focus of DIC is different to that of the marginal likelihood methods, as previously explained in detail.

The computing time to generate 100 iterations suggests that the MCMC program runs substantially slower for the SV Models 5-8 than for the SV
3.4 Case studies

Models 1-4. This is because most of the full conditional distributions for SV Models 5-8 are no longer log-concave and a Metropolis-Hastings updating step is needed.

Comparison between DIC and Chib's marginal likelihood reveals that the mixture normal-Gamma t-SV model (MODEL 8) is the only cause of the discrepancy. Here it is helpful to divide DIC into a pure measure of fit $D(\hat{\theta})$ and a measure of complexity $2p_D$ as in (3.5) to see that the t-SV model is heavily penalized by its high effective number of parameters. Considering $D(\hat{\theta})$, gives a value of -14,715.4 for the true Model 6 and a value of -14,744.1 for t-SV Model 8. Thus the t-SV model provides a better fit but its high complexity tips the scales against it. Although not reported, the non-scale mixture t-SV model has also been estimated and it has been found that the performance of these two representations are quite different. The non-scale mixture t-SV model performs even worse than the scale mixture t-SV model according to DIC. It has been recognized that different mixture distributions can generate different DIC values, due to the fact that different mixture distributions correspond to different prediction problems, and more research and experience is needed as to the performance of DIC in the area of mixture models (Spiegelhalter et al., 2002).

Furthermore, a possible explanation for the different ranking of Model 8 in
3.4 Case studies

Table 3.2 might be led back to the fact that DIC has not been investigated regarding universal application. So far DIC has been only verified for exponential families. Since the Student t-distribution does not belong to this family, there is a certain amount of doubt regarding applicability in this case.

Table 3.3 shows the smallest and largest values for DIC and the harmonic mean, the number of effective parameters $p_D$ and the goodness-of-fit $\bar{D}$, respectively, obtained for six runs with different starting values for each of the seven models.

Table 3.3: Deviance and Harmonic Mean (HM) Summaries for Simulated Data

<table>
<thead>
<tr>
<th>Model</th>
<th>$\bar{D}_{\min}$</th>
<th>$\bar{D}_{\max}$</th>
<th>$p_{D_{\min}}$</th>
<th>$p_{D_{\max}}$</th>
<th>DIC$_{\min}$</th>
<th>DIC$_{\max}$</th>
<th>HM$_{\min}$</th>
<th>HM$_{\max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mod 1</td>
<td>-14003.5</td>
<td>-14000.1</td>
<td>557.3</td>
<td>560.5</td>
<td>-13443.8</td>
<td>-13441.5</td>
<td>6871.41</td>
<td>6890.55</td>
</tr>
<tr>
<td>Mod 2</td>
<td>-14004.4</td>
<td>-14001.1</td>
<td>561.5</td>
<td>563.4</td>
<td>-13443.9</td>
<td>-13439.7</td>
<td>6860.43</td>
<td>6886.79</td>
</tr>
<tr>
<td>Mod 3</td>
<td>-14040.7</td>
<td>-14036.8</td>
<td>574.4</td>
<td>577.8</td>
<td>-13465.4</td>
<td>-13461.2</td>
<td>6870.30</td>
<td>6893.43</td>
</tr>
<tr>
<td>Mod 4</td>
<td>-14103.9</td>
<td>-14100.2</td>
<td>603.3</td>
<td>606.4</td>
<td>-13499.7</td>
<td>-13495.6</td>
<td>6939.66</td>
<td>6997.15</td>
</tr>
<tr>
<td>Mod 5</td>
<td>-14022.9</td>
<td>-14018.9</td>
<td>565.0</td>
<td>568.1</td>
<td>-13456.5</td>
<td>-13453.9</td>
<td>6890.98</td>
<td>6912.18</td>
</tr>
<tr>
<td>Mod 6</td>
<td>-14585.7</td>
<td>-14580.0</td>
<td>131.7</td>
<td>133.4</td>
<td>-14452.0</td>
<td>-14448.6</td>
<td>7139.71</td>
<td>7172.00</td>
</tr>
<tr>
<td>Mod 7</td>
<td>-14485.3</td>
<td>-14479.5</td>
<td>123.3</td>
<td>124.7</td>
<td>-14362.0</td>
<td>-14354.8</td>
<td>7091.90</td>
<td>7111.97</td>
</tr>
<tr>
<td>Mod 8</td>
<td>-14099.3</td>
<td>-14096.0</td>
<td>648.1</td>
<td>650.2</td>
<td>-13453.1</td>
<td>-13448.1</td>
<td>6933.15</td>
<td>6949.62</td>
</tr>
</tbody>
</table>

It serves to demonstrate that DIC varies from one run to another but is reasonably stable across runs. This is in contrast to the well known instability problem of the harmonic mean, which is apparent from the large discrepancies between the smallest and largest values for the harmonic mean.

However, it should be noted that slightly modified estimators of the harmonic
mean as proposed by Newton and Raftery (1994) are much more stable and do not suffer from the lack of a central limit theorem.

### 3.4.2 An empirical study

The dataset considered here consists of 1,512 mean-corrected daily continuously compounded returns, $y_t$, in decimals, on the Standard & Poors (S&P 100) 100 index, covering the period of time between January 1993 and December 1998. S&P 100 index returns have often been used in the literature. For instance, Blair, Poon and Taylor (2001a) estimate the GJR-GARCH model proposed by Glosten et al. (1993) based on the S&P 100 index returns for four different sample periods from March 1984 to December 1998, one of which is identical to the sample period examined here.

Data from the Chicago Board Options Exchange Market Volatility Index (VIX) for the same period of time will be used as a covariate, measuring the so-called implied volatility. For a detailed explanation of the Chicago Board Options Exchange Market Volatility Index, see Hol and Koopman (2002) and Fleming, Ostdiek and Whaley (1995). Both data series are plotted in the second and third panels of Figure 3.1.
For the empirical study, the models introduced in section 2.2 are fitted to the preceding data set. MODEL 4 is dropped from the list due to a great deal of convergence problems that have been encountered (it may be possible to achieve convergence by using different parameterizations or using different MCMC algorithms, however). Instead, as an additional extension, MODEL 9 that includes implied volatility will be considered.

In the last panel of Figure 3.1, the logarithm of absolute values of S&P 100 returns which can be regarded as an approximation of unobserved log-volatility is plotted. It can be seen that VIX and the logarithm of absolute values of S&P 100 returns are highly correlated.

A priori, λ is assumed to be uniformly distributed in the interval [-1,1].

Due to the inclusion of the implied volatility, it is not clear a priori whether the log-volatility $h_t$ is still highly persistent. Instead of using a rather informative prior of a beta distribution with parameters 20 and 1.5 for $\phi^*$, a less informative prior for $\phi^*$, namely, a uniform distribution with support between 0 and 1 is chosen.

In Table 3.4 means and standard errors (numbers in parentheses) of both prior and posterior distributions are reported, as well as the computing time to generate 100 iterations for each of the eight models.
3.4 Case studies

For Models 1-5, after a burn-in period of 50,000 iterations and a follow-up period of 250,000, every 20th iteration has been stored. In the remaining models, a burn-in period of 100,000 iterations and follow-up period of 900,000 has been chosen, and every 40th iteration has been stored.

From Table 3.4 it can be seen that the estimated means and standard deviations for the parameters appear quite reasonable and comparable with previous estimates in the literature.

For instance, in Model 1, the volatility process is estimated to be highly persistent.

In Model 5 the posterior mean of $\rho$ is -0.4139 with the upper limit of the 95% posterior credibility interval less than zero. This suggests that the leverage effect is an important feature for the S&P 100 index and confirms similar results in the literature, see Polson and Stroud (2003).

The parameter estimates for the two SV + jumps models provide similar results for those parameters already covered by the SV models without jumps. As already observed in Chib et al. (2002), the jump parameters $\kappa$ and $\delta$ are not as precisely estimated as other parameters. However, they are well identified as their posterior distributions are substantially different from their prior distributions. The posterior mean of the jump intensity $\kappa$ is 0.011 which means an average daily probability of 1.1% of a jump occurring. This implies that a jump can be expected to occur roughly every 90th day. The standard devi-
### 3.4 Case studies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>Mod 1</th>
<th>Mod 2</th>
<th>Mod 3</th>
<th>Mod 5</th>
<th>Mod 6</th>
<th>Mod 7</th>
<th>Mod 8</th>
<th>Mod 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>Prior</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
<td>10.0</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Prior</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
<td>9803</td>
<td>9789</td>
<td>8375</td>
<td>9743</td>
<td>9888</td>
<td>9873</td>
<td>9923</td>
<td>7.745</td>
</tr>
<tr>
<td>$\tau$</td>
<td>Prior</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
<td>1674</td>
<td>1729</td>
<td>1886</td>
<td>1947</td>
<td>1331</td>
<td>1302</td>
<td>1005</td>
<td>0.4375</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Prior</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
<td>1.03e-4</td>
<td>1.03e-4</td>
<td>1.03e-4</td>
<td>1.03e-4</td>
<td>1.03e-4</td>
<td>1.03e-4</td>
<td>1.03e-4</td>
<td>1.03e-4</td>
</tr>
<tr>
<td>$\psi$</td>
<td>Prior</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
<td>1.413</td>
<td>1.413</td>
<td>1.413</td>
<td>1.413</td>
<td>1.413</td>
<td>1.413</td>
<td>1.413</td>
<td>1.413</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Prior</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Prior</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0050</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Prior</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
<td>0.0114</td>
<td>0.0114</td>
<td>0.0114</td>
<td>0.0114</td>
<td>0.0114</td>
<td>0.0114</td>
<td>0.0114</td>
<td>0.0114</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Prior</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
<td>0.0315</td>
<td>0.0315</td>
<td>0.0315</td>
<td>0.0315</td>
<td>0.0315</td>
<td>0.0315</td>
<td>0.0315</td>
<td>0.0315</td>
</tr>
<tr>
<td>$\nu$</td>
<td>Prior</td>
<td>85.0</td>
<td>85.0</td>
<td>85.0</td>
<td>85.0</td>
<td>85.0</td>
<td>85.0</td>
<td>85.0</td>
<td>85.0</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
<td>36.4</td>
<td>36.4</td>
<td>36.4</td>
<td>36.4</td>
<td>36.4</td>
<td>36.4</td>
<td>36.4</td>
<td>36.4</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Prior</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Posterior</td>
<td>0.058</td>
<td>0.058</td>
<td>0.058</td>
<td>0.058</td>
<td>0.058</td>
<td>0.058</td>
<td>0.058</td>
<td>0.058</td>
</tr>
</tbody>
</table>

| Time (seconds) | 3.25 | 3.66 | 5.70 | 23.67 | 27.56 | 23.65 | 32.20 | 4.27 |
ation of the jump size is about 0.03; that is, daily jumps are usually around 6%.

In Model 8 the posterior mean of $\nu$ is 7.306 and similar to the values of 7.7 and 8.9 for the S&P 500 index in Sandmann and Koopman (1998) and Chib et al. (2002), respectively.

The posterior mean of $\lambda$ in Model 9 indicates that the implied volatility contains important information about the volatility process. Interestingly, allowing for the implied volatility as a covariate induces a negative posterior mean of the autoregressive coefficient in the model. This finding is similar to what was obtained in Hol and Koopman (2002) based on a S&P 100 index for a different period.

In Table 3.5 Chib’s marginal likelihood, harmonic mean, and DIC are reported together with $D$ and $p_D$ for each of the eight models as well as their associated rankings by each criterion.

The most adequate models to describe S&P 100 data according to DIC are the jump model without lagged observations (Model 7) and the jump model with lagged observations (Model 6), followed by the implied volatility model (Model 9) and the model including the leverage effect (Model 5).
Although the posterior means of the deviance for the jump models are higher than those of most of the other models, the effective number of parameters is much lower. The effective number of parameters is around 26 for the jump models which is less than one-third of the effective number of parameters for the closest competitor (refer to Table 3.6).

MODEL 5 has the lowest posterior means of the deviance which suggests the best fit to the data. However, its effective number of parameters is much higher than that of the other models. In particular, it is more than 10 times larger than that of the jump models. This renders a larger value of DIC.

As for the simulated data, neither DIC nor the harmonic mean provides the same model ranking as Chib’s marginal likelihood. According to Chib’s marginal likelihood, the strength of evidence to distinguish between the models is much weaker for the S&P 100 data than for the simulated data. For
example, the marginal likelihood values from the second best model and the third best model only differs by 0.84 which is not worth more than a bare mention according to Jeffrey's Bayes factor scale \( \exp(0.84) = 2.316 \). Nonetheless, both DIC and Chib's marginal likelihood select MODEL 7 (i.e., the jump model without lagged observations) as the best performing model, whereas the harmonic mean picks MODEL 8 (i.e., the t-SV model).

A close look at Table 3.5 reveals that the mixture normal-Gamma t-SV model (i.e., MODEL 8) is the major cause of the discrepancy between the DIC ranking and Chib's marginal likelihood ranking. This is a similar finding to the simulated data.

Another minor discrepancy arises from the first three models. Chib's marginal likelihood ranks MODEL 2 the worst model, whereas DIC ranks MODEL 1 the worst.

Table 3.6 shows the smallest and largest values for DIC and the harmonic mean, the number of effective parameters \( p_D \) and the goodness-of-fit \( \hat{D} \), respectively, obtained for six runs for each of the seven models.

Again it demonstrates that DIC varies from one run to another but is reasonably stable across runs and DIC is more stable than the harmonic mean. Also, it can be seen that the ranges of DIC overlap with each other for the first three
### 3.4 Case studies

Table 3.6: Deviance and Harmonic Mean (HM) Summaries for S&P100 Data

<table>
<thead>
<tr>
<th>Model</th>
<th>$\bar{D}_{min}$</th>
<th>$\bar{D}_{max}$</th>
<th>$p_{D_{min}}$</th>
<th>$p_{D_{max}}$</th>
<th>DIC$_{min}$</th>
<th>DIC$_{max}$</th>
<th>HM$_{min}$</th>
<th>HM$_{max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mod 1</td>
<td>-10617.1</td>
<td>-10611.6</td>
<td>81.9</td>
<td>85.8</td>
<td>-10531.3</td>
<td>-10527.4</td>
<td>5266.96</td>
<td>5275.31</td>
</tr>
<tr>
<td>Mod 2</td>
<td>-10618.4</td>
<td>-10615.9</td>
<td>85.1</td>
<td>88.2</td>
<td>-10531.4</td>
<td>-10529.6</td>
<td>5271.53</td>
<td>5279.98</td>
</tr>
<tr>
<td>Mod 3</td>
<td>-10621.3</td>
<td>-10613.1</td>
<td>83.1</td>
<td>88.8</td>
<td>-10532.8</td>
<td>-10530.0</td>
<td>5270.33</td>
<td>5277.29</td>
</tr>
<tr>
<td>Mod 5</td>
<td>-10941.4</td>
<td>-10934.9</td>
<td>323.0</td>
<td>328.6</td>
<td>-10613.3</td>
<td>-10610.9</td>
<td>5297.65</td>
<td>5308.88</td>
</tr>
<tr>
<td>Mod 6</td>
<td>-10681.5</td>
<td>-10674.9</td>
<td>24.9</td>
<td>29.8</td>
<td>-10656.7</td>
<td>-10645.1</td>
<td>5279.58</td>
<td>5287.72</td>
</tr>
<tr>
<td>Mod 7</td>
<td>-10680.7</td>
<td>-10675.9</td>
<td>24.1</td>
<td>30.3</td>
<td>-10655.1</td>
<td>-10645.7</td>
<td>5278.56</td>
<td>5288.27</td>
</tr>
<tr>
<td>Mod 8</td>
<td>-10791.5</td>
<td>-10787.3</td>
<td>222.2</td>
<td>226.1</td>
<td>-10565.5</td>
<td>-10566.9</td>
<td>5324.47</td>
<td>5331.00</td>
</tr>
<tr>
<td>Mod 9</td>
<td>-10741.5</td>
<td>-10738.8</td>
<td>120.8</td>
<td>124.0</td>
<td>-10618.8</td>
<td>-10617.1</td>
<td>5298.71</td>
<td>5302.28</td>
</tr>
</tbody>
</table>

models. This explains why the first three models are difficult to distinguish.

Additionally, in this section the implications of alternative prior distributions on DIC and Chib's marginal likelihood will be examined. In particular, the focus is on a subset of hyperparameters, namely, $\phi$ and $\kappa$. Also, only a subset of the models, namely, the basic SV model (MODEL 1), the SV model with a leverage effect (MODEL 5), and the SV+jumps model without lagged observations (MODEL 7) are considered. Following Chib et al. (2002) the following two alternative priors are used:

- Prior 2: $\phi^* \sim U(0, 1)$.

- Prior 3: $\phi^* \sim U(0, 1), \kappa \sim \text{Beta}$ with mean 0.0385 and standard error 0.0264.
All three models with Prior 2 and Model 7 with Prior 3 are reestimated and DIC and Chib's marginal likelihood are calculated accordingly. The posterior means, standard errors, DICs, and the marginal likelihoods are reported in Table 3.7. A comparison with the results in Table 3.4 shows that Prior 2 yields a posterior distribution that is almost identical to that with the original prior and that Prior 3 yields a posterior distribution that is reasonably close to that with the original prior. More importantly, DIC seems quite robust to the change of the prior. Moreover, it preserves the ranking of the models considered and the ranking is consistent with that based on the marginal likelihood. Whereas in the previous analysis a large data set has been made available and therefore the likelihood dominates the prior one would expect only marginal changes of the posterior distribution and therefore only marginal changes regarding DIC. In other situations however, following the Bayesian principle, the prior may dominate the likelihood and changes in the posterior distribution are deliberately accepted, which has of course implications on DIC.

With recent progress in computational Bayesian analysis, it is possible to sample from posteriors with virtual any prior required. Consequently, this sparked the discussion of which prior might be an "ideal" prior. Therefore, robustness or sensitivity analysis has become an important task and it has found recently increased attention. For a discussion of various aspects of robustness under decision theoretic considerations see Berger (1994).
Table 3.7: Sensitivity of DIC and Chib’s Marginal Likelihood to the Prior

<table>
<thead>
<tr>
<th></th>
<th>Model 1 Prior 2</th>
<th>Model 5 Prior 2</th>
<th>Model 7 Prior 2</th>
<th>Model 7 Prior 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>-9.970 (0.2543)</td>
<td>-9.963 (0.2240)</td>
<td>-10.10 (0.3250)</td>
<td>-10.20 (0.3232)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>0.9806 (0.0092)</td>
<td>0.9768 (0.0098)</td>
<td>0.9886 (0.00678)</td>
<td>0.9887 (0.00694)</td>
</tr>
<tr>
<td>$\tau$</td>
<td>0.1680 (0.0327)</td>
<td>0.1865 (0.0317)</td>
<td>0.1271 (0.0271)</td>
<td>0.1266 (0.0290)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>-0.4145 (0.0883)</td>
<td>0.0107 (0.0064)</td>
<td>0.0121 (0.0080)</td>
<td>0.0298 (0.0120)</td>
</tr>
<tr>
<td>$\kappa$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\delta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td>-10530.7</td>
<td>-10618.1</td>
<td>-10646.5</td>
<td>-10659.0</td>
</tr>
<tr>
<td>$\ln L$</td>
<td>5225.39</td>
<td>5228.03</td>
<td>5242.27</td>
<td>5240.88</td>
</tr>
</tbody>
</table>

3.5 Summary

In this section, the practical performance of DIC as model selection criterion for comparing various stochastic volatility models has been explored. DIC is based on the difference between true and estimated residual information and has been developed for model assessment. It is particularly suited to compare Bayesian models whose posterior distributions have been obtained.
using MCMC simulation.

Similar to AIC and BIC, DIC comprises two parts, a goodness-of-fit measure, the posterior distribution of the deviance, and a penalty term, the effective number of parameters, measuring complexity. Using this concept of effective number of parameters, DIC can be used in complex hierarchical models where the number of unknowns often exceeds the number of observations and the number of free parameters is not well defined. This is in contrast to AIC and BIC, where the number of free parameters needs to be specified. DIC has been implemented as a tool in the BUGS software package.

A simulation study using a SV+Jumps model as the true model has been carried out. Estimation results with respect to the simulated data are quite accurate for the true model and DIC clearly identifies the correct model out of eight different alternatives. If one were to omit the mixture t SV model, DIC would give the same model ranking as Chib’s marginal likelihood.

By comparing eight different SV models for the S&P100 index, comprising 1512 observations from 1993 to 1998, the jump volatility model without lagged observations turns out to be the most adequate as indicated by both DIC and Chib’s marginal likelihood. The Monte Carlo error of DIC is fairly low for all the models, thus indicating a stable performance for model comparison pur-
3.5 Summary

Finally, DIC appears robust to the change of prior distributions.
Chapter 4

Estimation of interest rate diffusions
4.1 Models

A starting point for modelling interest rates emerged in continuous time with the arithmetic Brownian motion representation of Merton (1973). This model has the advantage of being convenient, however allows for negative interest rates. This very simplified model has been improved by taking into account more appealing economic features.

The first major contribution after Merton can be attributed to Vasicek (1977) who introduced a mean-reverting process based on an Ornstein-Uhlenbeck specification. This process assures stationarity, nests the Merton model, and the occurrence of negative interest rates is far less likely.

A process which prevents interest rates from becoming negative was introduced by Cox, Ingersoll and Ross (1985). Their square-root model (also referred to as CIR model) takes account of an observed positive relationship (which they assume to be linear) between volatility and level of interest rate, the so-called level effect.

In contrast to leverage effects often reported for stock return data, the volatility parameter has here only a scaling function, since there is no need to consider
asymmetric effects induced by new information and changes in levels are described by level parameter $\gamma$ (compare eq. (2.6) or Table 4.1).

Chan et al. (1992) presented a generalised model, including as special cases the previous mentioned models and many others. A summary of important specifications nested in their class of models (CKLS) is listed in Table 4.1. This model class is one of the most widely used classes of one factor interest rate models.

The key generalisation presented involves the relaxing of the power restrictions. In their analysis, they obtain a significantly higher point estimate for the power or level parameter $\gamma$ compared to that of $\frac{1}{2}$ assumed by Cox et al. (1985). Their estimate of approximately $\frac{3}{2}$ leads to the conclusion that volatility is much more sensitive to interest rate levels than previously assumed.

Brenner et al. (1996), however criticise the idea of level modelling to be too restrictive and suggest to incorporate a news arrival process. This idea lead to GARCH and SV model specifications which have been extensively surveyed in this thesis in Chapter 2.

Andersen and Lund (1997) include level effects and extend the standard CIR
4.1 Models

Table 4.1: One-factor short-term interest rate models as in CKLS (1992)

<table>
<thead>
<tr>
<th>Interest rate models</th>
<th>$\alpha$</th>
<th>$\beta$</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CKLS (1992)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$dr(t) = (\alpha + \beta r(t))dt + \sigma r^\gamma dW$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Merton (1973)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$dr(t) = \alpha dt + \sigma dW$</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Vasicek (1977)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$dr(t) = (\alpha + \beta r(t))dt + \sigma dW$</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cox et al. (1985)/CIR</td>
<td></td>
<td></td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>$dr(t) = (\alpha + \beta r(t))dt + \sigma r^{\frac{3}{2}} dW$</td>
<td></td>
<td></td>
<td>$\frac{1}{2}$</td>
</tr>
<tr>
<td>Dothan (1978)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$dr(t) = \sigma r dW$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Geometric Brownian motion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$dr(t) = \beta r(t)dt + \sigma r dW$</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Brennan and Schwartz (1980)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$dr(t) = (\alpha + \beta r(t))dt + \sigma r dW$</td>
<td>0</td>
<td></td>
<td>$\frac{3}{2}$</td>
</tr>
<tr>
<td>Cox et al. (1980)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$dr(t) = \sigma r^{\frac{3}{2}} dW$</td>
<td>0</td>
<td>0</td>
<td>$\frac{3}{2}$</td>
</tr>
<tr>
<td>CEV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$dr(t) = \beta r(t)dt + \sigma r^\gamma dW$</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

model with a stochastic volatility component in their model. They model interest rates through a series of alternative ARCH models, like GARCH, EGARCH, Level-GARCH or Level EGARCH models. They found the latter ARCH variation as the most suitable in terms of overall fit for their setting. They estimate $\gamma$ close to 0.5 and document high persistence in volatility. Their conclusion therefore suggests that a stochastic volatility component is crucial for modelling short term interest rate dynamics.
Earlier work incorporating stochastic volatility in interest rate analysis has been produced by Fong and Vasicek (1991) and Vetzal (1997). In a more recent study, Durham (2003) estimates various stochastic volatility models for the short term interest rate. Related studies in discrete time have been presented by Brenner et al. (1996) and Ball and Torous (1999).

Latest advances in parametric interest rate modelling have been inspired by nonparametric work and findings therein (Aït-Sahalia, 1996a; Stanton, 1997; Jiang and Knight, 1997). These refinements include nonlinear drift and diffusion terms, multi-factor models, jumps, regime shifts, heavy-tailed distributions and central tendency modelling.

Important contributions to the jump incorporation in interest rate modelling literature have been made by Duffie et al. (2000) and Johannes (2004). The latter argues that although interest rate data is known to be very non-normal, using pure diffusion models induces approximate normality over short time intervals. Therefore models that refrain from incorporating jumps cannot be reconciled with observed data.

The same problem has been tackled in a different way by proposing regime switching models. Advocates of this type of models can be found in the inter-
4.1 Models

est rate literature in Gray (1996) or Pfann, Schotman and Tschernig (1996) estimating a variety of self-exciting threshold autoregressive models.

Central tendency modelling, which allows the short rate mean to revert randomly has been promoted by, for example, Beaglehole and Tenney (1991), Balduzzi, Das and Foresi (1998) and Jones (2003a). Implementation of this feature follows closely stochastic volatility modelling procedures, as the mean is modelled as an additional factor representing separate dynamics.

Aït-Sahalia (1996b) and Conley et al. (1997), for instance, analyse one factor models with non-linear drift functions. Pritsker (1998) and Chapman and Pearson (2000) criticize their nonparametric approach and show strong inconsistencies. They argue that their findings of nonlinearity in drift are due to misleading estimates. Jones (2003a) reviews models incorporating nonlinear mean reversion and explores the discrepancies in the literature regarding its significance. He finds in his Bayesian analysis that nonlinearities in the drift can be associated with informative prior beliefs and apparent noise in high-frequency data. He furthermore concludes that likelihood based methods, frequentist (through apparent mean-reversion bias) or Bayesian (prior choice) suffer from similar vulnerabilities as nonparametric methods when investigating the problem of nonlinearity in the drift function.
Duffee and Stanton (2004) remark that for the efficient method of moments technique (see section 2.3), comparable difficulties can be encountered, since if the parameter estimate of the auxiliary model is biased in a finite sample, this bias will also be reflected in the parameter estimate of the true model.

Ahn, Dittmar and Gallant (2002), alternatively model nonlinearities in interest rates by developing a generalized positive definite quadratic form.

So far, a general conclusion about one factor models is that parsimonious models show large inconsistencies with the data. Therefore, in the recent literature one can find interest rate modelling based on combining several of the previously described features to multi-factor modelling.

For instance Gallant and Tauchen (1998) introduce a very general two factor model including stochastic volatility and allow for feedback effects from the level of return into the volatility dynamics.

Andersen, Benzoni and Lund (2004) introduce a very general continuous time class of short-term interest rate models including multiple latent factors for both, drift and diffusion coefficients, as well as jumps in the interest rate level.
4.2 Estimation methods

In a discrete-time analysis of stochastic volatility models for the short rate, Nardari (1999) analyses several multi-factor extensions, as for instance non-linear drift, and feedback effects similar as previously analyzed in continuous time by Gallant and Tauchen (1998) and the effects of a more heavy-tailed underlying t-distribution (continuous time modelling using non-Gaussian distributions is based on Lévy processes, see Barndorff-Nielsen and Shephard, 2001 and Schoutens, 2003).

These multi factor interest rate models are only an important few under an enormous amount of generalisations found in the literature. An important finding in multi factor modelling is that when introducing a more flexible (SV, nonlinear, jump etc.) representation of interest rates, the dominant level-effect introduced in one factor models becomes insignificant.

4.2 Estimation methods

When discretizing one-factor interest models, an analysis using maximum like- lihood estimation techniques is generally feasible.
4.2 Estimation methods

Analysing an Ornstein-Uhlenbeck process yields either, in this case, the existing strong solution or the models obtained from various discretization techniques, including the simple Euler-Maruyama approximation result in obtaining the form of an autoregressive model.

The proposed frequentist methods to estimate the (stationary) autoregressive parameter will be either the generalized or ordinary least squares method, which coincides with maximum likelihood in the case of normality. There is a vast literature about estimating parameters in dynamic models (see Box and Jenkins, 1976; Hendry, 1995; and Stewart and Gill, 1998).

Ball and Torous (1999) implement quasi-maximum likelihood methods (see also section 2.3 for a short summary of this method) and apply the integration based filter to perform parameter estimation for discrete time SV modelling for interest rates.

For multifactor models like the SV model (see Chapter 2), however, the lack of an explicit likelihood function prevents the direct use of maximum likelihood. Similar problems occur when considering models in continuous time. Since only few processes such as the previously mentioned Ornstein-Uhlenbeck process have analytically tractable closed-form solutions is the likelihood function
readily available only in a few cases.

A range of estimation methods have been reviewed in Chapter 1 with most of them having been applied to interest rate data.

Nonparametric estimation methods of drift and volatility parameters are presented and applied to interest rates by Aït-Sahalia (1996a), Conley et al. (1997), Jiang and Knight (1997) and Stanton (1997).

EMM, described in detail in section 2.3 has found several recent applications for interest rate estimation, exemplary in Andersen and Lund (1997) and Ahn et al. (2002).

Various discretization methods have also been reviewed in Chapter 1.

Especially for continuous time models of the short-term interest rate, a series of discretization methods have been developed. For instance, Nowman (1997) introduced a discretization method inspired by the Gaussian estimation methods of Bergstrom (1990 and previous publications) with further improvements presented by Yu and Phillips (2001).
Nowman (1997) makes the critical assumption that the volatility of the interest rate remains unchanged over each one unit observation period. This makes the transition density be independent of the volatility and therefore a closed-form description of an approximate discrete time model can easily be derived.

Moreover, with constant volatility the resulting transition density is Gaussian and therefore Bergstrom’s Gaussian estimation method which allows exact maximum likelihood estimation can be applied. Although his approximation for the Merton and Vasicek model is exact, Nowman’s constant volatility assumption has often been criticized as being too strong.

Yu and Phillips (2001) develop a Gaussian estimator to overcome these restrictions by performing a random time-change of the underlying process using the concept of subordination. Applying this technique, they show how the original process can be transformed into a Gaussian one.

Approximations and transformations discussed here have been developed to obtain more exact specifications of interest rate dynamics in discrete time and reducing discretization bias. This idea is also behind the development of the simulated method of moments (SMM), as a extension of the generalised
4.3 Bayesian estimation

method of moments (Duffie and Singleton, 1993 and applied in Dai and Singleton, 2000), with EMM as a later suggested special case.

Hereby, moment functions may be obtained by repeatedly simulating from the underlying continuous time process.

One can use Markov Chain Monte Carlo techniques in a Bayesian setting to produce draws from data augmented distributions to reduce discretization bias. This MCMC approach which is similar to how the SMM method deals with the problem will be reviewed in greater detail in Chapter 4.3.1.

4.3 Bayesian estimation

When discretizing an interest rate model from the highly popular continuous time CKLS class, the resulting strong solution has a form similar to an autoregressive process of order 1 (AR(1)). Therefore, estimation of a one factor interest rate model as described in Chan et al. (1992) reduces to a problem similar to the well-known time series estimation problem for autoregressive parameters.
A typical frequentist approach is the least squares estimation approach (see also Chapter 4.3.2).

Alternatively, Bayesian methods for estimating autoregressive processes can be found in the literature. An early application is given by McCulloch and Tsay (1994) who employ the Gibbs sampler to analyse various autoregressive processes.

Barnett, Kohn and Sheather (1996) present a Bayesian analysis of very general autoregressive processes that encompasses the choice of autoregressive order, enforcement of stationarity, outlier detection, missing values and the case of multiplicative seasonality. Their analysis is based on a reparameterisation of the autoregressive parameters in terms of partial correlations. The authors develop a Markov Chain Monte Carlo algorithm which provides a simultaneous treatment of all the features above. Their analysis builds up on previous work on Bayesian analysis of time series models like Zellner (1971) and Jacquier et al. (1994).

Stroud et al. (2003) discretize a two-factor CIR model with jumps. Hereby, the interest rate volatility itself has been modelled as well as an CIR process.
4.3 Bayesian estimation

A block sampling MCMC method (Carter and Kohn, 1994) based on an auxiliary mixture model with mixing weights dependent on the state vector has been introduced for estimation purposes.

Furthermore, Bayesian approaches can also be easily applied in far more complex models, as in Frühwirth-Schnatter and Geyer (1998) who estimate term structure models. Johannes and Polson (2002) show, how a variety of specific term structure models can be estimated using MCMC methods with examples of Vasicek, CIR and Vasicek with jumps (Baz and Das, 1996) models. Furthermore a regime switching model (Landen, 2000 and Naik and Lee, 1994) and a time-varying central density model have been estimated.

An interesting variational Bayes alternative for estimating AR coefficients is described by Roberts and Penny (2002), who introduce a non-sampling Bayesian approach by finding an approximation of the true posterior density that minimizes the Kullback-Leibler divergence, a well-established measure of the discrepancy between two distributions.
4.3 Bayesian estimation

4.3.1 Bayesian estimation with augmented data

Recently proposed estimation methods of continuous-time models are based on a Bayesian method using augmented data combined with powerful MCMC techniques (Jones, 1998; Elerian, Chib and Shephard, 2001; Eraker, 2001). These methods are developed from a frequentist background in the spirit of Pedersen (1995) who shows how a stochastic differential equation for which the likelihood function is unknown or intractable, can be estimated. Therefore he has developed an approximation assuming the time gap between observations is sufficiently small. He advocates an augmented data approach by sampling higher frequency data and presents a set of simulated functions on which classical estimation techniques may be applied.

The key behind the Bayesian augmented data method lies in introducing auxiliary data points between two discrete time points and treating them in a Bayesian Markov Chain Monte Carlo (MCMC) analysis as additional parameters. This MCMC missing value technique using augmented data is well-known since Tanner and Wong (1987) and, unlike approximate frequentist methods, provides exact inference in finite samples. Introducing a sufficient amount of latent data points should considerably reduce discretization bias to any accuracy required. This has been demonstrated in Elerian et al. (2001).
The idea of reformulating complex problems as missing data problems reaches back at least to Rubin (1974). For a review of Bayesian analyses of various missing data problems using MCMC methods see Schafer (1997).

Applied work with the Bayesian MCMC approach using augmented data, described in detail in the following, is documented within the context of term structure modeling by Sanford and Martin (2004), for the analysis of interest rate models with nonlinear drift function as in Jones (2003a) or with applications to CEV and CEV with stochastic volatility models, as in Eraker (2001) or Durham and Gallant (2002).

The data augmentation approach reviewed here, follows closely the framework of Elerian et al. (2001), which has been further generalised by Eraker (2001). Mikkelsen (2001) developed a multivariate version of this approach.

Consider the data evolution described through an Itô stochastic process (1.1). The associated Euler-Maruyama approximation gives (1.5) and the resulting transition density is Gaussian with mean \( y(t) + a(y(t), \theta) \) and variance \( b^2(y(t), \theta) \).
It will be assumed that the time gaps between two time successive points are independent of time \( t \), \( (\Delta(t) = \Delta) \). Furthermore, \( \{\tau_{t,1}, \ldots, \tau_{t,M}\} \) has been defined to represent \( M \) evenly spaced time points between \( t \) and \( t + 1 \), so that \( \tau_{t,0} = \tau_{t-1,M+1} = t \) and \( \tau_{t,M+1} = \tau_{t+1,0} = t + 1 \).

Now, including the latent data \( y^*(t) = (y^*(t,1), \ldots, y^*(t,M)) \), \( y^*(t,j) = y^*(\tau_{t,j}) \) into the analysis, the new time gap can be defined as

\[
\Delta^* = \tau_{t,k+1} - \tau_{t,k} = \frac{\Delta}{M+1}.
\]  

A new improved transition density is obtained by integrating out the latent observations:

\[
f^*(y(t+1) \mid y(t), \theta) = \int f(y(t+1) \mid y^*(t,M), \theta) \prod_{j=2}^{M} f(y^*(t,j) \mid y^*(t, j-1), \theta) \]

\[
f(y^*(t,1) \mid y(t), \theta) \ dy^*(t, M), \ldots, dy^*(t, 1).
\]

Because of independence, the inherited Markov property, and Bayes theorem:

\[
= \int f(y(t+1), y^*(t, M), y^*(t, M-1), \ldots, y^*(t, 1) \mid y(t), \theta) \ dy^*(t, i)
\]
4.3 Bayesian estimation

\[
\begin{align*}
= & \int f(y(t+1), y^*(t), y(t), \theta) \frac{1}{f(y(t), \theta)} dy^*(t) \\
= & \int f(y(t+1) | y^*(t), y(t), \theta) f(y^*(t), \theta, y(t)) f(y(t), \theta) dy^*(t) \\
= & \int f(y(t+1) | y^*(t), y(t), \theta) f(y^*(t) | \theta, y(t)) dy^*(t)
\end{align*}
\]

where the normally distributed transition density will be inherited:

\[
f(y^*(t, j) | y^*(t, j-1), \theta) = f_N(y^*(t, j-1)) + a(y^*(t, j-1), \tau_{t,j-1}, \theta) \Delta^*,
\]

\[
b^2(y^*(t, j-1), \tau_{t,j-1}\theta) \Delta^*, \ t \in (0, T].
\]

The Bayesian missing value approach has been previously introduced by Jones (1998), Elerian et al. (2001) and Eraker (2001). Their work differ from each other and the approach in this work by the design of MCMC algorithms for sampling the data.

Instead of drawing from a highly multivariate distribution Jones (1998) applies a cyclic Metropolis chain to draw samples from many univariate distributions.

This concept has been originally introduced by Jacquier et al. (1994) for draw-
4.3 Bayesian estimation

ing latent volatilities in stochastic volatility models.

Elerian et al. (2001) proposed an independence Metropolis block sampler. Compared with work of the other authors, their approach seems to be the most efficient technique. On the other hand, their approach is practically also the most limited in that they consider only univariate and one-factor diffusion processes.

Eraker (2001) uses Gibbs sampling techniques, with missing vectors of observations simulated by a hybrid rejection Metropolis-Hastings algorithm, and shows how to apply the augmentation technique to more sophisticated models, like a CEV model with stochastic volatility.

4.3.2 Estimation bias and prior sensitivity

Least squares estimates for the autoregressive parameter in stationary time series are known as consistent and approximately normal distributed, but they can be strongly biased, especially for small sample sizes and processes with near-unit-root behaviour (see for example Orcutt and Winokur, 1969; Hendry, 1995, Chapter 3.11.).
Furthermore, including a fitted intercept or a trend exacerbates the magnitude of the bias. An analysis of the estimation bias in models with even more complicated dynamics (as for example the nonlinear drift model estimated by Jones, 2003a) turns out to be complicated and the practitioner is often left in deep darkness regarding the magnitude of bias. Therefore, bias correction terms, assessed for simpler cases by Kendall (1954) for the AR(1) model with intercept using first-order approximations are desirable, but hard to obtain and, above that, rely on different arbitrarily specific forms of error expansions.

Furthermore, these correction factors can increase the variance of the estimator and therefore the mean square error (see Copas, 1966; Orcutt and Winokur, 1969).

Further work on bias correction has been done by Kiviet and Phillips (1993, corrected OLS estimation). Classical unit root discussions are given by Campbell and Perron (1991).

Applied work, investigating the small sample properties of point estimators for the AR(1) autoregressive parameter with Monte Carlo methods has been published by Copas (1966) and Thornber (1967).
EMM and ML estimation results have also been compared in a Monte Carlo study by Duffee and Stanton (2004) for an AR(1) process. They note that estimation bias is even more pronounced when applying EMM due to the divergence between sampled and simulated data.

It has to be noted, that with the exception of Jones (2003a) and Phillips and Yu (2004) interest rate models and attempts to assess estimation biases within these models have been mentioned. Bias correction methods and unbiased estimation techniques such as Andrews (1993) have been developed for models with simpler structures. However, when considering more complex models (several factors, nonlinear models), bias assessment proves more complicated and obscure.

Ball and Torous (1996) investigate the bias problem for interest rate dynamics applying a CIR model, when parameter estimation is carried out by Least-Squares and GMM.

For the continuous time Ornstein-Uhlenbeck process, least squares estimates have been calculated and a bias discussion has been given by Perron (1991).
4.3 Bayesian estimation

A detailed comparison of estimation biases in the case of a continuous time square-root diffusion model regarding the performance of MLE, QMLE, GMM and EMM is given for a high persistent, low volatility (among other scenarios) parameter constellation by Zhou (2001). In these more general cases, the median-unbiased estimation technique introduced by Andrews (1993) is not feasible any more. However, relying on bootstrap methods, Tanizaki (2000) proposed simulated mean and median-unbiased estimators for the AR(1) model and more general cases. The jackknife method developed by Phillips and Yu (2004) based on work of Quenouille (1956) proves to be very flexible and can be employed, even in fairly complicated situations, successfully. Improved results have been obtained even in cases when the underlying model is misspecified.

Alternatively to classical approaches, there is meanwhile a large amount of literature describing Bayesian attempts to deal with the estimation bias problem.

Unbiasedness is generally a criteria for point estimation with frequentist background, since by definition expectation over the sample space is involved. Applying a Bayesian approach with low-information priors can be seen as (Carlin and Louis, 2000, Appendix B)

"an effective balance between frequentist robustness and Bayesian efficiency ... often "beating the frequentist approach as its own game.”"
4.3 Bayesian estimation

From a decision-theoretic viewpoint point estimation is in most cases related to applying a squared error loss function. Differences between the Bayesian and frequentist approach in finding a decision rule lie then in the definition of the associated risk function, which is the posterior risk for the former and the frequentist risk for the latter. The posterior mean, obtained by summarizing the posterior distribution from an MCMC output minimizes the posterior risk. Hereby, it has been noted by Efron (2002) that MCMC methods can be seen as the Bayesian counterpart of the frequentist bootstrap and the theoretical basis of bootstrapping is closely connected with the discussion of how to select non-informative priors for a Bayesian analysis (Efron, 1998).

Furthermore, Bayesian posterior distributions for the parameters of autoregressive processes depend crucially on initial conditions and stationarity assumptions for the underlying process.

The key feature when performing Bayesian inference is, however, the choice of priors. Objectivity is a general aim for the statistician when prior information is not available and the main point of criticism for frequentists regarding Bayesian analyses. Nevertheless, much applied work makes use of very specific, informative priors for analyzing various problems. Reasons for their subjective choice originate in theoretical constraints, real prior knowledge or computa-
tionally convenience.

For an objective analysis, however, priors which take this aim into account are preferable. It turns out that in the case of choosing a prior for the regression parameters in autoregressive processes special care is needed.

A remarkable discussion how to perform a valid Bayesian analysis and how to define a non-informative prior distribution has been sparked regarding the investigation of unit-root detection in AR(1) processes, encompassing the whole volume 6 of *The Journal of Applied Econometrics*.

Hereby, Sims and Uhlig (1991) are concerned with the inappropriate use of classical p-values as asymptotic equivalences of Bayesian posterior tail probabilities. In their study they calculate for the AR(1) process with fixed error variance \( \sigma^2 = 1 \) the joint distribution of correlation parameter \( \rho \) and its least-square estimate \( \hat{\rho} \) and graphically investigate the conditional probability distribution function of \( \rho \) as the cross section of the joint distribution along the line for fixed least squared estimates. They apply a uniform prior, with the belief that this prior would be a good approximation of a flat prior on \( \rho \) and consider 31 different values of \( \rho \) to illustrate their results on a grid ranging from 0.8 to 1.1. Their findings show that by using a flat prior, the distribution
of the estimate given the data can be asymmetric, whereas the likelihood or probability distribution function of the data as a function of the parameter is symmetric.

In reply to this article, Phillips (1991) argues that a flat prior, often applied as an non-informative prior for standard problems such as the static linear regression models behaves informatively in a dynamic model. Therefore, he argues that a flat prior is not suitable for dynamic models. Alternatively, he supports a Bayesian analysis set up with a Jeffreys prior (Jeffreys, 1946) which is invariant to reparameterisations and takes into account the special covariance dynamics which emerges in a dynamic case. The corresponding compare these two priors yields dramatical differences in location, scale and modality of the resulting posterior distributions permitting contradicting results for the unit-root testing procedures.

But Jeffreys' prior seems not to be the final solution to the Bayesian unit-root controversy. The proposed Jeffreys' prior has been criticized as putting too much weight on the non-stationary case. Kass and Wasserman (1996) discuss this point and give a general review of philosophies behind various non-informative prior suggestions. Furthermore, the calculation of Jeffreys prior depends crucially on initial conditions and stationarity assumptions; Uhlig
4.3 Bayesian estimation

(1994a) investigates the effects of using a likelihood function based on Jeffreys priors calculated on initial fixed observations or their exact initial distributions, respectively.

Additional work is presented by Stambaugh (1999) who focuses on the case of lagged stochastic regressors with correlation between regression disturbance and the regressors' innovation. He reports properties of the OLS estimator and sensitivity performance of different priors.

Applied work has been presented by Jones (2003a) who compares results obtained from different scenarios in a nonlinear autoregressive setting by considering two classes of priors representing different assumptions regarding stationarity for the interest rate process.

Berger and Yang (1994) report their experience with reference priors (Berger and Bernardo, 1989), which shows no improvement in bias reduction in the AR(1) case. Moreover, for the stationary process, their reference prior coincides with Jeffreys prior. They conclude, that for the AR(1) problem, an operationally objective Bayesian answer may not be available.

Other prior suggestions for this particular case are based on the concept of
maximal data information (Zellner, 1977).

By concentrating on parameterisation issues, Schotman (1994) summarizes several priors and investigates their sensitivity regarding the posterior of the autoregressive parameter, taking into account dependencies from the unconditional mean. He remarks that posteriors which put some weight on the unit root case, somehow correct for the small sample bias inherited in the classical estimate of the autoregressive parameter. Jones also noted this.

Further work on the unit root problem and associated prior choice discussions are given in Uhlig (1994b).

An excellent discussion of further developments regarding the initial discussion in Sims and Uhlig (1991) and Phillips (1991) can be found in Bauwens, Lubrano and Richard (1999, Chapter 6) who compare the flat prior with 3 variants of Jeffreys prior (Phillip’s prior, Berger and Yang’s prior, and Lubrano’s prior) using risk functions in a grid of values reaching from 0.8 to 1.1. Their main result is that the flat prior is indeed a comparatively poor choice, being uniformly dominated by each of the other priors. Each of the other priors is dominant over a different region, but overall they show similar performances.
More general suggestions of prior choices have been made for parameters of the AR model by Huerta and West (1999). Their description of the AR model follows the illustration with the characteristic polynomial and the related decomposition. For real and complex root parameters dirichlet priors have been introduced, whereas standard normal or uniform priors for the AR coefficients are implemented, depending on stationary constraints.

Whereas Huerta and West (1999) are introducing flexibility in introducing a new class of priors, Barnett et al. (1996) present an approach based on a more flexible model description. They decompose the original autoregressive process in a partial regressions model, which includes a fairly large amount of partial regression parameters. Their distributions are assumed independent of each other and uncertainty is characterized by flat prior assumptions.

### 4.4 Case study

Here, the missing value approach as introduced in section 4.3.1 will be applied to a simple interest model from the CKLS class, namely the Vasicek model. A reasonable high sample frequency has been chosen \( \Delta = \frac{1}{12} \) and a replicated
4.4 Case study

dataset has been simulated with parameter constellations that are very sensitive to estimation bias.

In the case of an AR(1) process, by choosing the autoregressive parameter $\rho$ close to 1, one has to deal with a "near unit-root" situation, which makes estimation being an extremely difficult task. At $\rho = 1$, the unit root case, the key problem is that there exists a point of discontinuity, where a very different and hard to calculate asymptotic distribution (unlike the cases $\rho < 1$ and $\rho > 1$) has to be considered.

This is in contrast to previous work of Elerian et al. (2001) who chose a sample frequency of ($\Delta = 1$) and ($\Delta = 5$), respectively in their analysis of a CIR model. This represents a situation where a high amount of discretization bias can be expected. Furthermore, they have simulated a replication of datasets using a parameter setting, for which estimation can be performed with ignorable magnitude of estimation bias.

Unlike Elerian et al. (2001) who showed that the Bayesian approach with augmented data can significantly reduce discretization bias, the focus here is on estimation bias. Whether the Bayesian augmented data approach can reduce estimation bias is unclear at this time. Therefore, the impact of this
estimation method on reducing estimation bias will be investigated here, regarding sensitivity to prior choice and two different reparameterisations of the discretized model.

Throughout this analysis, the Bayesian all purpose software WinBUGS (Spiegelhalter et al., 1996; see also Chapter 3.4.1) has been used. This package provides a single-update Gibbs sampling scheme which is also able to treat missing values as latent variables and sample from the full posterior distributions of these augmented parameters. The advantage of WinBUGS is that this imputation is carried out automatically by sampling new values for the missing observation at each iteration. This procedure leads to posterior estimates of all parameters involved with a full adjustment for uncertainty in the latent variables.

Although there are more efficient problem-specific sampling techniques available, convergence can be achieved in most cases very quickly with the Gibbs sampler for the dataset analysed here. Therefore in this application, efficiency concerns are of minor interest.

Furthermore, software available at Gelmans webpage under

http://www.stat.columbia.edu/~gelman/bugsR/
for running WinBUGS from R has been used to summarize simulation results, to carry out convergence diagnostics, and as a convenient platform for further modification.

In this section the practical implementation of the method described previously will be illustrated by using simulated data from a Vasicek process and results from the combined prior and reparameterisation analysis will be presented.

Applying the Euler discretization technique for the Vasicek model

\[ dy(t) = \alpha y(t) \, dt + \sigma \, dW(t), \ t \in (0, T], \]  \hspace{1cm} (4.2)

yields

\[ y(t + \Delta) = \alpha \Delta + (1 + \beta \Delta)y(t) + \sigma \sqrt{\Delta} \epsilon_{t+\Delta}, \ \epsilon_{t+\Delta} \sim N(0, 1), \]  \hspace{1cm} (4.3)

whereas the strong solution results in:

\[ y(t + \Delta) = \frac{\alpha}{\beta} \left( \exp(\beta \Delta) - 1 \right) + \exp(\beta \Delta)y(t) + \sqrt{\frac{\sigma^2}{2\beta} \left( \exp(2\beta \Delta) - 1 \right)} \epsilon_{t+\Delta}, \]

\[ \epsilon_{t+\Delta} \sim N(0, 1). \]  \hspace{1cm} (4.4)
200 replications of 40 time series observations each of a Vasicek process have been simulated using the exact solution of the underlying process. To construct a situation where eminent small sample bias occurs, a value for the mean-reverting parameter $\beta = -1.25$ which leads to $\rho_{Vas} = \exp(\beta \Delta) = 0.901$ for the AR coefficient has been chosen. This amount of fairly high persistence has been typically observed in many series of interest rates data.

Other parameter values set for the simulation are $\sigma_{Vas} = \sqrt{\frac{\sigma^2}{2\beta}}(\exp(2\beta \Delta) - 1) = 0.04$, $\alpha = 0$ and a sample frequency of $\Delta = \frac{1}{12}$.

We will refrain from reporting estimation results for the volatility, since previous studies confirm that this parameter can be estimated fairly accurately and furthermore, it does not show any significant impact on the persistence parameter estimation procedure, on which this chapter is focused.

The model is then reestimated considering the discrete model corresponding to the Euler approximation (which is analogous to the situation when zero latent values are introduced) using both ML and a Bayesian approach. The latent variable technique will be employed and the model will be reestimated with a fixed number of 1, 3, and 5 auxiliary data points inserted between two
observed data points corresponding to sampling frequencies of $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{6}$.

When choosing the initial value $y_0$ being stochastic and further assuming the stationary process has been run for some time prior to the sample period of interest, then $y_0$ can be treated as a realisation of a normal distributed random variable with mean $-\frac{\sigma}{\beta}$ and variance $\frac{\sigma^2 \Delta}{1-(1+\Delta)^2}$. This assumption for $y_0$ proves extremely convenient since it provides internal consistency to the model (Hendry, 1995). It assigns the same mean and variance to every other $y_t, t > 0$, so that all the observations have the same distribution.

Alternatively, one could define the initial value deterministically. Often a choice of $y_0 = 0$ has been employed. The analysis here, however, is based on the former standpoint.

When the underlying time series does not follow stationary assumptions, special insights in the covariance structure prove necessary. Since there is some evidence that interest rates are stationary (Chan et al., 1992), priors that provide constraints for stationarity have been chosen. Stationarity can be introduced by defining a prior over the stationary range or by discarding in the MCMC analysis all draws falling into a non-stationary range (Congdon, 2001).
4.4 Case study

A variety of priors will be used for the Bayesian analysis to investigate its sensitivity regarding the posterior distribution and therefore regarding Bayesian estimation. To mimic two extreme situations a Normal prior (A) with a mean placed very close to the true value \( \rho = (\rho_{\text{Euler}})^{\frac{1}{11}} \) and \( \beta = -1.25 \), respectively) and small variance (0.1 for \( \rho_i \) and 1 for \( \beta \)) has been chosen to illustrate extreme prior belief regarding the parameter. In contrast, a Normal prior (B) with mean far away from the true parameter (0.2 for \( \rho_i \) and -10 for \( \beta \)) with comparatively large variance (1 for \( \rho_i \) and 10 for \( \beta \)) has been chosen to represent falsely assumed prior knowledge, but with a certain amount of doubt about this prior belief. A Normal prior (C) with mean and variance equal to the ML estimates has been employed to integrate an empirical Bayesian approach into the analysis.

Certain variants of priors (D)-(F) with the aim to mimic different philosophies of non-informativeness are corresponding to a (vague) Uniform, flat, and Jeffrey’s prior.

The flat prior (D) employed here is a built-in prior for WinBUGS and represents an improper uniform distribution. It serves as an alternative to the vague Uniform prior (E) chosen in the range between 0 and 1 for \( \rho_i \) and between -100 and 0 for \( \beta \). The concept of using vague, but proper priors has
4.4 Case study

often been criticised as a "hide-and never solve-" concept (Berger, 1997, page 247). Nevertheless, vague priors have extensively been used in various analyses in many fields and are therefore employed for comparison reasons.

Jeffreys prior (F) is defined as

$$\pi_J(\rho, \sigma) = \sqrt{\text{det}(I(\rho, \sigma))},$$

(4.5)

with $I(\rho, \sigma)$ being the Fisher Information matrix. For the AR(1) model, Jeffreys prior can be represented as (Berger and Yang, 1994):

$$\pi_J(\rho, \sigma) = \sqrt{\text{det}(I(\rho, \sigma))} = \left[\frac{2T}{\sigma^2} \left( \frac{T}{1 - \rho^2} + \frac{1 - \rho^{2T}}{1 - \rho^2} \left\{ E\left(\frac{X_0^2}{\sigma^2}\right) - \frac{1}{1 - \rho^2} \right\} \right) \right]^{\frac{1}{2}},$$

(4.6)

with sample size $T$ and initial condition $X_0$ of the AR(1) process.

The Jeffreys prior for the initial condition $y_0 = 0$ and no restriction regarding stationarity has been derived by Phillips (1991). For a variety of different initial conditions and different ranges for the autoregressive parameter, Uhlig (1994a) has derived exact Jeffreys priors. The exact marginal Jeffreys prior,
4.4 Case study

given a stochastic initial value following a Normal distribution, is improper for the stationary case. Therefore an approximate Jeffreys prior for $\rho$ assuming to be independent of nuisance parameters has been considered here, which coincides with the Beta (0.5, 0.5) distribution (Thornber, 1967).

When carrying out a Bayesian analysis of the Vasicek model, it is natural to set the priors directly on the parameter of interest (REP1) ($\rho_{Euler} = 1 + (\beta \Delta)$ for the Euler approximation). Alternatively, the flexibility of the Bayesian MCMC approach combined with the use of the WinBUGS software package can be exploited to easily draw inference of functions of parameters. Several suggestions for possible reparameterisations have been given in the literature. For an introduction of general reparameterisation issues, see Hills and Smith, 1992.

One alternative way of reparameterising would be to allow a prior specification on mean reversion parameter $\beta$ instead of $\rho_{Euler}$ and therefore to carry out inference on a parameter directly available in the transition densities (REP2). Considering a prior specification directly on $\beta$ is of special interest since accurate estimates for $\beta$ are even harder to obtain than for $\rho$ (Phillips and Yu, 2004).
4.4 Case study

The analysis is then performed by estimating $\beta$ in each model corresponding to the number of auxiliary data points introduced and by inferring from that point back to the autocorrelation parameter for this specific model $\rho_i = (1 + (\Delta_i^+ \cdot \beta))$, $i = 0, 1, 3, 5$.

Having obtained estimates for $\rho_i$ in REP1 or REP2 respectively, the transformation $\hat{\rho}_{i,Euler} = \hat{\rho}_i^{(i+1)}$ has been made to draw inference about the original parameter of interest.

Jeffrey's prior for $\beta$ in REP2 has been derived by modifying the $\text{Beta}(0.5,0.5)$ distribution for $\rho_{Euler}$ by applying the change of variable technique to obtain a marginal prior for $\beta$ with distribution

$$p(\beta) = \frac{\Delta}{B(0.5,0.5)}(\beta \Delta + 1)^{-\frac{1}{2}} (-\beta \Delta)^{-\frac{1}{2}},$$

in the range $[-\frac{1}{\Delta}, 0)$ with $B(0.5,0.5)$ is the Beta function with parameter values 0.5 and 0.5, respectively. Furthermore, priors for $\alpha$ and $\sigma$ follow Normal with mean 0 and variance 10 and Uniform (0,10) distributions, respectively.

Each choice of amount of missing values $(0, 1, 3, 5)$ has been combined with each prior (A-F) and each reparameterisation. For every analysis two chains
4.4 Case study

Table 4.2: Parameter estimates for the VASICEF-tion, 0 auxiliary data points)
\[ y_t | \theta_t = \alpha \Delta^* + (1 + (\Delta^* \cdot \beta)) \cdot y_{t-1} + u_t, \quad u_t \overset{i.i.d.}{\sim} \]

According to 0 auxiliary data points: \( \Delta^* = \frac{A}{A} = \)
True parameter values are \( \rho = 0.901 \) and \( \beta = -

<table>
<thead>
<tr>
<th>Prior</th>
<th>Reparameterisation 1</th>
<th>\hat{\beta}</th>
<th>SD</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>0.7915</td>
<td>0.0757</td>
<td>0.1323</td>
<td></td>
</tr>
<tr>
<td>(B)</td>
<td>0.7741</td>
<td>0.0898</td>
<td>0.1547</td>
<td></td>
</tr>
<tr>
<td>(C)</td>
<td>0.7832</td>
<td>0.0790</td>
<td>0.1410</td>
<td></td>
</tr>
<tr>
<td>(D)</td>
<td>0.7789</td>
<td>0.0895</td>
<td>0.1506</td>
<td></td>
</tr>
<tr>
<td>(E)</td>
<td>0.7785</td>
<td>0.0883</td>
<td>0.1502</td>
<td></td>
</tr>
<tr>
<td>(F)</td>
<td>0.7950</td>
<td>0.0896</td>
<td>0.1380</td>
<td></td>
</tr>
<tr>
<td>ML estimate</td>
<td>0.8012</td>
<td>0.1001</td>
<td>0.1406</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prior</th>
<th>( \hat{\beta} )</th>
<th>SD</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>-2.502</td>
<td>0.9082</td>
<td>1.547</td>
</tr>
<tr>
<td>(B)</td>
<td>-2.711</td>
<td>1.0780</td>
<td>1.816</td>
</tr>
<tr>
<td>(C)</td>
<td>-2.595</td>
<td>0.9479</td>
<td>1.645</td>
</tr>
<tr>
<td>(D)</td>
<td>-2.653</td>
<td>1.0742</td>
<td>1.767</td>
</tr>
<tr>
<td>(E)</td>
<td>-2.658</td>
<td>1.0599</td>
<td>1.762</td>
</tr>
<tr>
<td>(F)</td>
<td>-2.460</td>
<td>1.0254</td>
<td>1.586</td>
</tr>
</tbody>
</table>

have been constructed, for each of them, after a burn-in period of 5,000 iterations, a further 5,000 iterations have been generated.

Convergence has been monitored considering a potential scale reduction factor \( \hat{R} \) as defined in Gelman, Carlin, Stern and Rubin (2003). A large potential
4.4 Case study

Table 4.3: Parameter estimates for the VASICEK MODEL (Euler approximation, 1 auxiliary data point)

\[ y_t \mid \theta_t = \alpha \Delta^* + (1 + (\Delta^* \cdot \beta)) \cdot y_{t-1} + u_t, \quad u_t \overset{i.i.d.}{\sim} N(0, \sigma^2 \Delta^*), \quad t = 1, \ldots, n. \]

According to 1 auxiliary data point: \( \Delta^* = \frac{\Delta}{2} = \frac{1}{24} \) and \( \rho_1 = 1 + \Delta^* \beta \). True parameter values are \( \rho = 0.901 \) and \( \beta = -1.25 \).

<table>
<thead>
<tr>
<th>Prior</th>
<th>Reparameterisation 1</th>
<th>Reparameterisation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\beta} )</td>
<td>SD</td>
</tr>
<tr>
<td>(A)</td>
<td>0.7754</td>
<td>0.0877</td>
</tr>
<tr>
<td>(B)</td>
<td>0.7684</td>
<td>0.0946</td>
</tr>
<tr>
<td>(C)</td>
<td>0.7725</td>
<td>0.0889</td>
</tr>
<tr>
<td>(D)</td>
<td>0.7752</td>
<td>0.0915</td>
</tr>
<tr>
<td>(E)</td>
<td>0.7730</td>
<td>0.0915</td>
</tr>
<tr>
<td>(F)</td>
<td>0.7972</td>
<td>0.0850</td>
</tr>
</tbody>
</table>

ML estimate: 0.8012 | 0.1001 | 0.1406

<table>
<thead>
<tr>
<th>Prior</th>
<th>Reparameterisation 1</th>
<th>Reparameterisation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\beta} )</td>
<td>SD</td>
</tr>
<tr>
<td>(A)</td>
<td>-2.944</td>
<td>1.2511</td>
</tr>
<tr>
<td>(B)</td>
<td>-3.052</td>
<td>1.3735</td>
</tr>
<tr>
<td>(C)</td>
<td>-2.985</td>
<td>1.2715</td>
</tr>
<tr>
<td>(D)</td>
<td>-3.070</td>
<td>2.3296</td>
</tr>
<tr>
<td>(E)</td>
<td>-2.984</td>
<td>1.3213</td>
</tr>
<tr>
<td>(F)</td>
<td>-2.645</td>
<td>1.1970</td>
</tr>
</tbody>
</table>

scaling factor is then associated with the belief that further improvement regarding the target distribution of the underlying parameter can be achieved by increasing the number of iterations. Here, it has been followed the recommendation in Gelman et al. (2003) to discard results for which \( \hat{R} > 1.1 \) for any of the parameters' posterior distribution in the analysis.
Table 4.4: Parameter estimates for the Vasicek Model (Euler approximation, 3 auxiliary data points)

\[ y_t \mid \theta_t = \alpha \Delta^* + (1 + (\Delta^* \cdot \beta)) \cdot y_{t-1} + u_t, \quad u_t \sim i.i.d. N(0, \sigma^2 \Delta^*), \quad t = 1, \ldots, n. \]

According to 3 auxiliary data points: \( \Delta^* = \frac{A}{4} = \frac{1}{48} \) and \( \rho_3 = 1 + \Delta^* \beta. \) True parameter values are \( \rho = 0.901 \) and \( \beta = -1.25. \)

<table>
<thead>
<tr>
<th>Prior</th>
<th>Reparameterisation 1</th>
<th>Reparameterisation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\rho} )</td>
<td>SD</td>
</tr>
<tr>
<td>(A)</td>
<td>0.7699</td>
<td>0.0915</td>
</tr>
<tr>
<td>(B)</td>
<td>0.7662</td>
<td>0.0983</td>
</tr>
<tr>
<td>(C)</td>
<td>0.7717</td>
<td>0.0927</td>
</tr>
<tr>
<td>(D)</td>
<td>0.7792</td>
<td>0.0910</td>
</tr>
<tr>
<td>(E)</td>
<td>0.7752</td>
<td>0.0881</td>
</tr>
<tr>
<td>(F)</td>
<td>0.7957</td>
<td>0.0869</td>
</tr>
<tr>
<td>ML estimate</td>
<td>0.8012</td>
<td>0.1001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prior</th>
<th>Reparameterisation 1</th>
<th>Reparameterisation 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \hat{\beta} )</td>
<td>SD</td>
</tr>
<tr>
<td>(A)</td>
<td>-3.175</td>
<td>1.4442</td>
</tr>
<tr>
<td>(B)</td>
<td>-3.253</td>
<td>1.6310</td>
</tr>
<tr>
<td>(C)</td>
<td>-3.151</td>
<td>1.4831</td>
</tr>
<tr>
<td>(D)</td>
<td>-3.032</td>
<td>1.6223</td>
</tr>
<tr>
<td>(E)</td>
<td>-3.088</td>
<td>1.3817</td>
</tr>
<tr>
<td>(F)</td>
<td>-2.789</td>
<td>1.3456</td>
</tr>
</tbody>
</table>

The estimates changed only marginally when a smaller limit (1.01) or a larger sample period (15,000 iterations after a burn-in of 5,000 iterations) have been chosen.

In most of the cases, convergence could be achieved. Only for the case of REP1, 5 auxiliary data points and the flat prior, for most of the replicated
4.4 Case study

Table 4.5: Parameter estimates for the VASICEK MODEL (Euler approximation, 5 auxiliary data points)

\[ y_t | \theta_t = \alpha \Delta^* + (1 + (\Delta^* \cdot \beta)) \cdot y_{t-1} + u_t, \quad u_t \, \text{i.i.d.} \sim N(0, \sigma^2 \Delta^*), \quad t = 1, \ldots, n. \]

According to 5 auxiliary data points: \( \Delta^* = \frac{\Delta}{6} = \frac{1}{72} \) and \( \rho_5 = 1 + \Delta^* \beta \).

True parameter values are \( \rho = 0.901 \) and \( \beta = -1.25 \).

<table>
<thead>
<tr>
<th>Prior</th>
<th>( \hat{\rho} )</th>
<th>SD</th>
<th>RMSE</th>
<th>( \hat{\rho} )</th>
<th>SD</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>0.7724</td>
<td>0.0921</td>
<td>0.1574</td>
<td>0.8572</td>
<td>0.0239</td>
<td>0.0490</td>
</tr>
<tr>
<td>(B)</td>
<td>0.7760</td>
<td>0.0916</td>
<td>0.1541</td>
<td>0.6918</td>
<td>0.1092</td>
<td>0.2351</td>
</tr>
<tr>
<td>(C)</td>
<td>0.7728</td>
<td>0.0897</td>
<td>0.1556</td>
<td>0.7846</td>
<td>0.0707</td>
<td>0.1354</td>
</tr>
<tr>
<td>(D)</td>
<td>0.7743</td>
<td>0.0936</td>
<td>0.1567</td>
<td>0.7749</td>
<td>0.0887</td>
<td>0.1534</td>
</tr>
<tr>
<td>(E)</td>
<td>0.7722</td>
<td>0.0909</td>
<td>0.1569</td>
<td>0.7705</td>
<td>0.0949</td>
<td>0.1605</td>
</tr>
<tr>
<td>(F)</td>
<td>0.7936</td>
<td>0.0883</td>
<td>0.1382</td>
<td>0.7907</td>
<td>0.0919</td>
<td>0.1428</td>
</tr>
<tr>
<td>ML estimate</td>
<td>0.8012</td>
<td>0.1001</td>
<td>0.1406</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prior</th>
<th>( \beta )</th>
<th>SD</th>
<th>RMSE</th>
<th>( \beta )</th>
<th>SD</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A)</td>
<td>-3.192</td>
<td>1.5203</td>
<td>2.466</td>
<td>-1.848</td>
<td>0.3274</td>
<td>0.681</td>
</tr>
<tr>
<td>(B)</td>
<td>-3.133</td>
<td>1.5147</td>
<td>2.416</td>
<td>-4.508</td>
<td>1.8836</td>
<td>3.763</td>
</tr>
<tr>
<td>(C)</td>
<td>-3.179</td>
<td>1.4566</td>
<td>2.417</td>
<td>-2.950</td>
<td>1.0837</td>
<td>2.016</td>
</tr>
<tr>
<td>(D)</td>
<td>-3.143</td>
<td>1.5812</td>
<td>2.467</td>
<td>-3.143</td>
<td>1.5812</td>
<td>2.467</td>
</tr>
<tr>
<td>(E)</td>
<td>-3.193</td>
<td>1.4868</td>
<td>2.446</td>
<td>-3.230</td>
<td>1.5888</td>
<td>2.539</td>
</tr>
<tr>
<td>(F)</td>
<td>-2.866</td>
<td>1.4183</td>
<td>2.150</td>
<td>-2.919</td>
<td>1.4803</td>
<td>2.231</td>
</tr>
</tbody>
</table>

Datasets convergence could not be achieved, this is indicated by a blank entry in Table 4.5. For the cases of REP2 and REP1 and 0, 1 and 3 auxiliary data points introduced, in at least 192 out of the 200 simulated datasets convergence could be achieved with the 10000 iterations (obtained from applying Gelman’s scale reduction factor). In those few cases of non-convergence, results for these particular simulated datasets have been excluded from further analysis.
Estimates for root mean squared errors have been calculated using (Berger and Yang, 1994) averages over the posterior means of $(\hat{\rho}_i - \rho)^2$ calculated across the 200 replicated time series. Estimates for the autoregressive parameters $\rho_i$ have been obtained by averaging the posterior means for $\rho_{i,j}$, $j = 1,...,200$ calculated for each of the 200 simulated data sets.

Tables 4.2 - 4.5 contain parameter estimates and standard deviations of $\rho_{Euler}$ and the mean-reversion parameter $\beta$ for reparameterisation REP1 and REP2, derived by the Bayesian augmented data approach using 0, 1, 3 and 5 auxiliary data points and a range of different, more or less informative priors as introduced previously.

The maximum likelihood estimate for $\rho_{Euler}$ is reasonably accurate given the induced bias of approximately $\frac{1 + 3\rho}{40}$ (see Marriott and Pope, 1954 and Kendall, 1954).

The chosen sample size of 40 is very small. Therefore, the prior distributions will have a larger impact on the resulting posterior distributions than they would have in the case of large sample sizes where the likelihood will have a stronger impact. This has been confirmed for REP2 by observing throughout
all analyses that the tight prior around the true value which leads to the least biased estimates (between 0.85 and 0.86 for $\rho$ and between -1.85 and -1.79 for $\beta$) whereas the analyses using the prior with location far away from the true value and high dispersion provide the worst biased estimates (between 0.69 and 0.72 for $\rho$ and between -4.5 and -3.4 for $\beta$).

Considering REP1, these effects are not so clear. Estimation results obtained for these two priors become more similar to the results from priors (C)-(E) when the growing number of auxiliary data points is introduced.

Furthermore, with the exceptions of priors (A) and (B), results obtained for REP2 and REP1 are fairly similar. This can be observed throughout all numbers of auxiliary data points and both estimates of $\rho$ and $\beta$.

Generally, in nearly all the cases parameter estimates become slightly more biased than the classical maximum likelihood estimates, especially when focusing on estimation of $\beta$ with the increased dimension of the underlying parameter space (increasing the number of auxiliary data points) together with increased standard deviations.

Overall, nearly all estimates (with the exception of the extreme knowledge
case) come along with fairly high standard deviations.

Estimates obtained with priors (C)-(E) are very similar throughout the analyses but none of them lies near the maximum likelihood estimate, which is the least biased estimate together with estimates obtained by employing Jeffreys prior.

Estimates obtained by the analysis employing Jeffreys prior are closest to the ML estimate compared with results obtained from estimates provided by the remaining prior choices.

4.5 Summary

In this chapter, recent ideas regarding interest rate modelling have been reviewed. Sophisticated interest rate models originate in continuous time and during recent years they have been refined to account for specific features observed in the data. Advances in estimation methodologies provide the platform for drawing inference about models with complex dynamics.
4.5 Summary

When discretizing continuous time interest rate models, one often ends up with transition densities having the form of autoregressive processes. If these processes are very persistent, additional to the already inherited discretization bias, classical as well as Bayesian methods provide biased estimates for the autoregressive parameters for finite sample sizes.

Employing a new Bayesian method to a Vasicek and CIR process, respectively, this estimation bias has been investigated by using data augmentation. This is astonishingly easy to accomplish when employing the Bayesian software package WinBUGS, which allows to take into account all the induced uncertainty induced by the latent variables when deriving posterior distributions.

A Monte Carlo study has been carried out for a small sample and effects of different reparameterisations and prior specifications have been documented for the autoregressive parameter of the Vasicek process using the Bayesian data augmentation method.

Elerian et al. (2001) have documented a significant reduction in discretization bias when employing this method to low frequency data. When applying the Bayesian augmented data augmentation method to high frequency data and therefore diminishing the effects of discretization bias, however, it can be
concluded that a further reduction of bias, namely pure estimation bias, can not be achieved.
Bibliography


an autoregressive model using Markov chain Monte Carlo”, *Journal of
Econometrics*, 74, 237-254.

cesses implicit in PHLX deutsche mark options", *Review of Financial


structure for jump-diffusion processes: A numerical analysis", *Journal
of Fixed Income*, 6, 78-86.

interest rate-contingent claim pricing equations", *Journal of Fixed In-
come*, 1, 69-83.

erion for comparing stochastic volatility models", *Journal of Business

1-124.


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


BIBLIOGRAPHY


