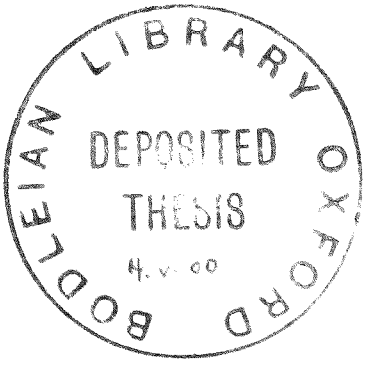
NUFFIELD COLLEGE, UNIVERSITY OF OXFORD

SIMULATION ESTIMATION OF CONTINUOUS-TIME MODELS WITH APPLICATIONS TO FINANCE

Ola Elerian[1]



[1] A thesis submitted to the Faculty of Social Studies in fulfillment of the requirements for the degree of Doctor of Philosophy of the University of Oxford, Trinity Term 1999.

Abstract

Over recent years, we have witnessed a rapid development in the body of economic theory with applications to finance. It has had great success in finding theoretical explanations to economic phenomena. Typically, theories are employed that are defined by mathematical models. Finance in particular has drawn upon and developed the theory of stochastic differential equations. These produce elegant and tractable frameworks which help us to better understand the world. To directly apply such theories, the models must be assessed and their parameters estimated. Implementation requires the estimation of the model's elements using statistical techniques. These fit the model to the observed data. Unfortunately, existing statistical methods do not work satisfactorily when applied to many financial models. These methods, when applied to complex models often yield inaccurate results. Consequently, simpler analytical models are often preferred, but these are typically unrealistic representations of the underlying process, given the stylised facts reported in the literature. In practical applications, data is observed at discrete intervals and a discretisation is typically used to approximate the continuous-time model. This can lead to biased estimates, since the true underlying model is assumed continuous.

This thesis develops new methods to estimate these types of models, with the objective of obtaining more accurate estimates of the underlying parameters present. The methods are applicable to general models. As the solution to the true continuous process is rarely known for these applications, the methods developed rely on building an Euler-Maruyama approximate model and using simulation techniques to obtain the distribution of the unknown quantities of interest. We propose to simulate the missing paths between the observed data points to reduce the bias from the approximate model. Alternatively, one could use a more sophisticated scheme to discretise the process. Unfortunately, their implementation with simulation methods require us to simulate from the density and evaluate the density at any given point. This has until now only been possible for the Euler-Maruyama scheme. One contribution of the thesis is to show the existence of a closed form solution from use of the higher order Milstein scheme. The likelihood based method is implemented within the Bayesian paradigm, as in the context of these models, Bayesian methods are often analytically easier. Concerning the estimation methodology, emphasis is placed on simulation efficiency; design and implementation of the method directly affects the accuracy and stability of the results. In conjunction with estimation, it is important to provide inference and diagnostic procedures. Meaningful information from simulation results must be extracted and summarised. This necessitates developing techniques to evaluate the plausibility and hence the fit of a particular model for a given dataset. An important aspect of model evaluation concerns the ability to compare model fit across a range of possible alternatives. The advantage with the Bayesian framework is that it allows comparison across non-nested models.

The aim of the thesis is thus to provide an efficient estimation method for these continuous-time models, that can be used to conduct meaningful inference, with their performance being assessed through the use of diagnostic tools.

ACKNOWLEDGEMENTS

I would like to thank my supervisor, Dr. Neil Shephard, for his constant patience and encouragement throughout the thesis. I am grateful for his invaluable advice and for his guidance. I have greatly benefited from his insight, ability to focus ideas and approach to research.

I am indebted to Professor David Hendry for his comments and advice on the reading of material as the thesis was developing. I am grateful for his support and encouragement. I would also like to thank Dr. Jurgen Doornik for his help and for writing Ox - the programming language used to code and implement the algorithms. I would also like to thank Dr. Doornik, in his capacity as the external examiner, and Professor Michael Sørensen for their comments and reading of the thesis for the viva. I have benefited from many helpful discussions and would like to thank Dr. Bent Nielsen, Dr. Jouni Kuha, Dr. David Myatt, Dr. Nour Meddahi, and Dr. Tina Rydberg for their comments and reading of the thesis.

In addition, I gratefully acknowledge the financial support provided by Nuffield College.

Contents

I Background Motivation and Literature

1 Introduction 3

2 Stochastic Differential Equations 11

3 Basic Simulation Estimation 36

II Contribution to the Literature

4 Single-Move Metropolis-Hastings Algorithms for the States 55

5 Block Sampling for the Auxiliary States 70

6 Parameter Estimation 97

7 Application to Interest-rate data 136

8 Estimating the likelihood for stochastic differential equations 149

III Extensions

9 The Milstein scheme and higher order approximations 170

10 A continuous-time bivariate model with unobserved stochastic volatility 186

11 Multivariate stochastic differential equations 208

12 Concluding remarks and further topics 221

IV References

Bibliography 223

Part I

Background Motivation and Literature

Background Motivation and Literature

1 Introduction **3**

1.1 Overview . . . 3

1.2 Motivation . . . 4

1.3 Structure of thesis . . . 6

1.4 Software . . . 8

1.5 Previously circulated material and related independent work . . . 8

1.6 Frequently used notation . . . 9

2 Stochastic Differential Equations **11**

2.1 Definitions and notation . . . 11

2.2 Examples of processes to be considered . . . 15

2.3 Discretizing the continuous-time process . . . 17

2.4 Augmentation: introducing auxiliary points . . . 20

2.5 Existing methods for estimating stochastic differential equations . . . 26

2.6 Conclusion . . . 32

2.7 Appendix . . . 33

3 Basic Simulation Estimation **36**

3.1 Introduction . . . 37

3.2 Markov chain Monte Carlo methods . . . 37

3.3 Sampling variates from a multivariate distribution . . . 40

3.4 Using residuals to assess model fit . . . 47

3.5 Comparison of non-nested models . . . 49

3.6 Conclusion . . . 50

CHAPTER 1

INTRODUCTION

1.1 OVERVIEW

In the world of finance, events occur rapidly and frequently. Financial variables can thus be thought of as being close to evolving continuously in time. Economic theory has developed mathematical models which incorporate this feature. Examples include models of portfolio management, term-structure modelling and the pricing of derivatives. Typically, the central element is the presence of a *non-linear diffusion.* This necessitates the estimation of a system of *stochastic differential equations.*

The application and testing of a theory requires empirical observations. These observations are data that are recorded periodically. Sometimes, they may be recorded at high frequency; for example, stock price data may be recorded every second. Other information may be recorded rather less frequently, for instance, daily or weekly. All examples, however, remain discrete in nature. Since data are discrete, existing statistical methods are designed to fit discrete-time models. Unfortunately, they cannot, in general, be directly applied to the continuous-time financial models. As these are only observed at discrete intervals, it is generally not possible to determine the analytical form of the required density functions, unless the stochastic differential equation can be solved out explicitly. Examples of systems that can be solved out analytically are rare. Previous researchers have attempted to circumvent this problem by approximating the continuous-time financial models by discrete-time statistical analogues. Such efforts have been only partially successful as results are often *biased*, and reducing this bias requires very high frequency data sets. Effectively, there is a *missing values* problem. The analyst does not observe the path followed between any given pair of observed data points. In addition, high frequency data may not be available for many series, such as interest rate data.

The research reported here constructs a general estimation method which can be used to deal with the financial and economic models discussed above. There are two main objectives; the first is to define the necessary mathematical structures. The second develops the computational tools to enable the efficient execution and implementation of the models considered. Models may not be tractable, that is, explicit solutions may not be available. In fact, economic theory frequently suggests models which involve integrals that cannot be solved explicitly. An example of this is where economic variables follow a non-linear stochastic differential equation. Term structure modelling is an example where the statistical model is often formulated using some non-linear diffusion. In addition, though most models for financial time series are based on continuous-time, data is observed at (fixed) discrete intervals. It is well known that estimation methods are then strongly biased unless the interval between those observations tends to zero. This thesis develops computational methods to bridge the gap between

continuous models and discrete data by simulating a subset of the missing data points. Having used a deterministic approximation to obtain an approximate model in the context of discrete observations, stochastic simulation is used to simulate the missing/auxiliary or latent values.[1] By simulating these auxiliary points, we are using a better model resulting in bias reduction. The method is easy to extend, for example, in cases where the data may be irregularly spaced.

In cases where one has access to both low and high frequency data for a given series, there may be reasons to work with the low frequency data, due to problems including bid-ask spread, discreteness in observed prices and irregularity in the sampling interval of data. For example, suppose the analyst is interested in summarising an aggregate level, such as the daily rate. Though the theory tells us to use high frequency data, (such as the intra-day rate), it may be more desirable to use the low frequency data, (such as the daily rate) to reduce the effects of market microstructure; high frequency data is often contaminated and vast in quantity. If we consider a model for prices, for example, there is evidence in the literature to suggest that the underlying price process follows a continuous-time model. However, prices can often only move in fixed tick sizes, and so a discrete-time model may be more realistic for more frequently observed data, to reflect points defined on a lattice structure. For example, in Engle and Russell, (1998 and 1999), the authors suggest thinking of high frequency data as being generated by processes with discrete sample paths.

1.2 Motivation

We will motivate the research by considering a specific example. The celebrated Black and Scholes model, (Black and Scholes 1973), lies behind much of the interest in modelling these continuous-time models. The model was initially proposed for option pricing by postulating a diffusion model for the underlying asset and obtaining the fair price of a European call or put option. The model satisfies the *no arbitrage* condition, so that the opportunity of making a risk free profit within the market is not possible. Previous research has focused on whether the postulated underlying distribution of returns actually corresponds to the actual distribution and how the unknown parameters present should be estimated. More complex structures can be postulated for the underlying assets, for example, by treating the unknown mean and volatility as additional random processes. These include treating the volatility as an unobserved state variable which itself follows a diffusion. Though these models may no longer be arbitrage based relations, empirical evidence suggests that the assumption of constant instantaneous volatility is unrealistic. More complex structures, however, introduce quantities that cannot be estimated using analytical methods, taking us back to the issues introduced above.

Trading will be assumed to take place in continuous-time, but only discrete observations will be available in practice. For these types of problems, there are two competing views. The first is that the stochastic differential equation can be viewed as the limiting approximation to a discrete stochastic difference equation. The second is to view a discretised model as an approximation to the underlying continuous-time Markov process, so that the observed data is only an approximation to an unobservable process. Issues concerning the validity of the continuous-time formulation against use of a discrete model are not addressed in this thesis. As a basis, the continuous-time framework is assumed for the underlying model and since many of these are intractable, a discretised approximation must be used. Hence, given a continuous-time model, the thesis addresses the question of estimation.

[1] For the estimation problem we consider, the terms missing, auxiliary and latent are taken to be equivalent and will be used interchangeably.

The underlying sample paths for the random variables are assumed to follow diffusions which are continuous, but nowhere differentiable. Diffusions are often used to specify an underlying model for financial data because of their Markovian structure, a necessary condition under the efficient market hypothesis. In an efficient market for assets, the price process must satisfy the Markov property, that is, given all past information, only current prices are needed to determine next period's prices. The mathematical models all rely on *Itô* calculus, which is used extensively in modern economics and finance. Kiyosi Itô initiated and developed the theory of *stochastic integrals* and stochastic differential equations. In his 1942 and 1951 papers, Itô first applied the theory to Kolmogorov's problem of finding the Markov process for a system of infinitely divisible distributions, (when a given probability measure can be expressed as a convolution power of another probability measure, such as the Gaussian density - see Section 2.1), given some initial conditional distribution; the problem is stated formally in Kolmogorov (1931). Sample paths of the Markov process can be constructed directly using the probabilistic approach suggested by Lévy and substantiated by Itô. These sample paths often depend on some driving process such as Brownian motion or the Wiener process, (the probability theory being developed by Wiener 1923 and Lévy 1937). The Wiener process, an example of a homogeneous Lévy process and thus having independent and stationary increments, can be thought of as a random walk in continuous time. It can be interpreted, amongst other things, as a martingale, a Gaussian process, a Markov process and a diffusion. It's flexibility means that, under a suitable series of transformations, it can be used as a building block to construct other processes, see the introduction in Rogers and Williams, (1994, Ch. 1). Both Wiener and Lévy had shown that the stochastic integral, (present in the stochastic differential equation and involving the increments of the Wiener process), could not be interpreted as an ordinary integral in the Riemann-Stieltjes sense, (since the value of the Wiener process depends on the intermediate points in the approximating sums and is nowhere differentiable for almost all sample paths). Consequently, Itô introduced the idea of a stochastic integral, so that under certain growth conditions on the drift and volatility functions, the sample path could be interpreted as the unique solution to the stochastic differential equation. A unique solution is required for the statistical inference to have meaningful interpretation. The concepts will be outlined in Part I of the thesis.

For many cases, because of the well known and attractive properties associated with maximum likelihood, the aim will be to obtain the maximum likelihood estimates of the parameters of the process. So, for a given set of measurements, the aim will be to estimate some parameter vector of interest. For some models, exact maximum likelihood estimation of the parameters is possible, as the stochastic differential equation has a strong solution that is known analytically so that the true Markovian transition density between two observed points is explicitly available. The transition density is simply the probability law linking two separate points in time. Thus, in the likelihood context, an estimate of the parameters is based on a likelihood function. However, in general, an explicit form is not available, necessitating the use of an approximation. Such approximations may cause estimates to be biased, since the approximation error depends on both the sample size and the time step between each pair of observed values. In practice, bias in the estimates can be considerable since both the sample size and the time intervals are fixed. The fact that analytic solutions are rarely available and only known in a limited number of cases, has led to growing interest in methods for estimating stochastic differential equations on the basis of discrete measurements. A great deal of recent work in the literature has focused on developing general estimation methods. Important developments have included the indirect inference method, the efficient method of moments approach, non-parametric techniques, estimating functions and filtering methods. A review will be provided in Part I of the thesis.

The contribution of this thesis lies in developing methods to perform the likelihood analysis, even when

the true transition density is not known. For any pair of observed values, we propose to estimate non-linear diffusions by *filling in* the missing data points between each pair of observations, that is, introducing a set of latent or auxiliary points between the observed values. The process is then discretised using a Taylor expansion, for example the Euler approximation. This implies an approximate conditional density. Simulated values for these latent points are obtained using Markov chain Monte Carlo (MCMC) methods, in particular, the Metropolis-Hastings (M-H) algorithm. Such methods are effective in dealing with latent variable models as they allow dependent samples to be drawn from any density of known form. Knowledge from previous information is incorporated through a prior density. The prior, in conjunction with the likelihood, (based on values of the observed quantities), enables estimation of the posterior density of interest. The method is applied to all pairs of observations in the sample, giving us a sequence of imputed values. Simulated values are then obtained from the posterior density of the parameters, forming one sweep or iteration of the algorithm. The algorithm is then repeated until convergence. The Euler approximation, a simple stochastic Taylor expansion, is known to induce some bias and so other schemes such as the Milstein approximation, (which has a higher rate of convergence than the Euler scheme), are considered. The parameters of the process are estimated within a Bayesian framework as the method is analytically easier and there is little difference when compared to classical likelihood analysis due to the size of typical financial datasets. Bayesian methods, in this context provides a powerful device to evaluate complex likelihoods, especially when the process is not entirely observed. Summary statistics, such as the posterior mean can be reported, and diagnostic measures, (which are required to assess the performance of the method), can thus be evaluated.

1.3 Structure of thesis

The thesis is composed of three parts. Part I provides the background and Part II the contribution to the literature. Part III explores possible extensions, based on the work developed in Part II. The literature review in Part I outlines previous work in the area and motivates the research. Background material to the concepts of stochastic differential equations and simulation techniques is provided in Chapters 2 and 3 respectively.

New methods for estimating stochastic differential equation systems are presented in Part II. Once an approximate model is constructed, the issue of bridging the gap between the approximate model (based on the discrete data) and the continuous-time model is addressed. Given the observed data Y, simulation of the auxiliary data Y^*, and the parameter of interest θ, is carried out using MCMC methods. This relies on constructing a suitable Markov chain. It is assumed that the process under consideration tends to an equilibrium, $f(Y^*;\theta|Y)$. The simulated quantities have a limiting invariant distribution which they are assumed to follow in equilibrium. Simulation from the posterior density of interest $f(Y^*;\theta|Y)$, is achieved by establishing the density as the limiting distribution of a Markov chain and simulating from this density until convergence, that is, when the chain has reached equilibrium. The advantage is that simulation from the (often complicated) joint distribution can be achieved by sampling from the conditional decomposition, which are easier to sample from. If the resulting density can additionally be evaluated, a Gibbs sampler can be implemented to directly obtain variates from the posterior distribution, $f(Y^*;\theta|Y)$. An example of a model which can be sampled using Gibbs, considered in Chapter 4 is the Ornstein-Uhlenbeck process, since the strong solution is known and a closed form exists for the density of the auxiliary points, given both neighbours. For most cases this is not feasible and the M-H algorithm is used instead.

The M-H algorithm requires constructing a proposal density and an acceptance probability. The proposal is used to suggest the next step in the path of the chain. The acceptance probability ensures that the proposed move is in the same direction as that of the limiting distribution. Chapter 4 concentrates on sampling the paths between the observed data by employing a single-move algorithm. Various approximations are considered, including a zero-order and first-order Taylor approximation of the auxiliary point around its neighbours and extending it to consider a Taylor approximation of the log-likelihood. The trajectory of the Markov chain provides a correlated sequence of draws form the posterior density $f(Y^*;\theta|Y)$. These simulated values can be used to plot the mean of the simulations and for certain models checked against the strong solution. To obtain reliable estimates, it is important to run the chain until convergence has been reached. One measure which can be used to assess convergence and hence the efficiency of the sampler, is through plots of the correlogram; if the chain has reached equilibrium, the correlogram should exhibit little activity. The number of auxiliary points introduced between each pair of observations, assuming these are the same over the pairs, is denoted by M. An increase in M improves the approximation $f^M(y_{t+1}|y_t;\theta)$, the augmented one-step ahead transition density, to $g(y_{t+1}|y_t;\theta)$ and hence reduces the inherent bias associated with the discretised version. However, for simulation purposes, an increase in M increases the dimension of the simulation state space and can increase the correlation between the draws in the simulation procedure. Rather than increase simulation time, it becomes important to construct efficient algorithms. Efficiency is measured in terms of speed of convergence and the ease in obtaining simulated samples. Considerations include the choice of the proposal, the need for transforming the random variables, the cost and time of an iteration and the stability of solutions. This is dealt with in Chapter 5, which extends the analysis to consider a blocking strategy for updating the auxiliary states. Initially, the algorithm is modified to incorporate a fixed block size. Comparisons of different block sizes show that small values induce correlation in the sampler, while high values, (such as sampling all M points at once), can lead to high levels of rejection for proposed moves. Performance of the algorithm is additionally increased through use of a random block size, based on a Poisson random variable, which is one example of a randomising device that can be used. Numerical convergence rates can only be obtained for processes with a known strong solution and a method for obtaining numerical convergence rates is proposed.

Chapter 6 addresses issues of parameter estimation when the likelihood is available in closed form and when it has to be computed using simulation techniques. Simulated values of θ are drawn from the marginal $f(\theta|Y)$, obtainable through $f(\theta|Y,Y^*)$. These can be used to conduct posterior inference about θ, by summarising the draws from a given model and examining diagnostic statistics. The sampled mean and standard deviations can be used as draws from the corresponding posteriors. Consistency of these estimates is established by appealing to certain ergodic laws for Markov chains outlined in Part I of the thesis. These can be grouped into intervals and plotted using histograms based on the relative frequencies. As the number of iterations of the algorithm increase, the resulting histogram approximates the density more closely. The approximation error is also in our control, through choice of the number of iterations of the algorithm.

An application to the daily Eurodollar exchange rate, a series used extensively in the literature is provided in Chapter 7. Techniques to evaluate the plausibility of a particular model are shown in the context of this example. Standardised forecast errors and autocorrelation functions of the residuals and their reflected counterparts can be used as visual aids to assess the fit. A formal method for comparing fit across models is provided through use of Bayes factors. This involves computing the ratio of marginal likelihoods, evaluated at some suitable point, such as the mode. This requires computation of the log-likelihood for each model, since the marginal likelihood is simply the integral of the likelihood with

respect to the prior density. The method additionally allows comparison across non-nested models.

Part II concludes with Chapter 8 on obtaining an estimate of the likelihood, useful for inference and necessary from a classical perspective. Existing methods for obtaining an expression for the log-likelihood are first outlined and a new method is presented. The existing methods are based on simulated likelihood methods and have been used in numerous approaches, including filtering techniques and methods based on estimating functions. One method uses the one-step ahead conditional transition density, and is shown to be inefficient as it induces high Monte Carlo variance; conditioning is only considered with respect to the previous state. An alternative method is based on a simulated likelihood ratio statistic. The new method uses the concept of importance sampling for obtaining the likelihood. Using simulated data, a comparison between the three approaches is presented; the importance sampling approach is shown to dominate the other methods in finding the MLE.

Part III explores extensions to the work examining alternative schemes that can be used to approximate the transition density in Chapter 9 and extends the method developed in Chapter 5 to a bivariate but partially latent system in Chapter 10 and general multivariate models in Chapter 11. The volatility functions in the univariate models considered in earlier parts of the thesis specify constant volatility coefficients. Constant instantaneous volatility is unrealistic, given the vast amount of empirical evidence that exists against this hypothesis. The models applied to the Eurodollar deposit spot-rate in Chapter 7, were shown to provide an inadequate fit to the data. Volatility clustering in the residuals suggest that these models should be extended to incorporate an underlying model for the volatility, such as a heavy tailed latent stochastic volatility component. The methods developed in Chapter 5 are extended to consider a bivariate system of stochastic differential equations, where one of the processes is unobserved. Specifically, the geometric Brownian motion and Generalised Cox, Ingersoll and Ross model are considered, where both are extended to incorporate a latent stochastic volatility structure. The techniques are illustrated in the context of a bivariate stochastic volatility model, extended to incorporate a Markov switching structure. The general formulation is addressed in Chapter 11. The general structure can accommodate many types of models. These include multi-factor models, where the state space represents a system of p series, (for example in a term structure model, where each series represents a different maturity level, or a series of exchange rates across different currencies), or a combination of series, each with a multi-factor model associated with it. Its specification is easily extended to incorporate a Markov switching regime or a jump diffusion component, often suggested for asset price models. The thesis concludes with Chapter 12.

1.4 Software

A variety of high level and low level programming languages have been used. The code for the algorithms was programmed in Ox, an object-oriented matrix programming language developed by Doornik (1996). The syntax for Ox is similar to that of C++. The speed of the algorithms can be increased, by linking them to C++ libraries. Certain graphics were drawn using Givewin, (see Doornik and Hendry 1999), S-Plus and Adobe Illustrator. The thesis was written using LaTeX.

1.5 Previously circulated material and related independent work

Some of the material presented in Part II of the thesis was used in the paper "*Likelihood Inference for Discretely Observed Non-Linear Diffusions*" (1998), with co-authors Siddhartha Chib (John M. Olin

School of Business, Washington University) and Neil Shephard (Nuffield College), [2] and has been submitted for publication. Additionally, work contained in Chapter 9, in Part III of the thesis, was used as a basis for the material in a sole-authored paper "*A Note on the Existence of a Closed Form Conditional Transition Density for the Milstein Scheme*" (1998).[3] These have been presented at various conferences including: the ESRC Econometric Study Group, annual conference in Bristol, 9-11 July 1998, the workshop on Statistical Inference for Stochastic Processes: theoretical and computational aspects, held at the Department of Theoretical Statistics at the University of Copenhagen during August 25-29, 1998, the National Bureau of Economic Review (NBER)/NSF Time Series conference held in Chicago, 4-5 September, 1998, at a workshop on Econometrics and Financial Time Series, at the Isaac Newton Institute at Cambridge University, 12-16 October, 1998 and at the Econometric Society European Meeting, to be held at Santiago de Compostela, 28th August- 1st September, 1999.

The idea of augmentation has also been discussed independently by Eraker(1998), in the context of the stochastic volatility model. The work contained in the thesis differs from Eraker (1998) in several crucial respects. The method is based on a tuned MCMC algorithm for carrying out the calculations, rather than a single move method and includes techniques for computing the model marginal likelihood and diagnostic checks on the fitted models. Related ideas have been developed by Billio, Monfort, and Robert (1998), (in their work applying the Geyer (1999) simulated likelihood ratio method to diffusions and other econometric problems), and by Roberts and Stramer (1999), (using MCMC methods in conjunction with the Girsanov transformation). An outline of their methods can be found in the literature review in Chapter 2.5.

1.6 Frequently used notation

Global notation used throughout the thesis is given below. Specific notation is defined in the corresponding parts of the thesis.

$y_t|y_s$ denotes y_t given the state y_s;

$P(y_t \in B|y_s)$ is the probability of the transition from y_s at time s, to the state y_t in the set B at time t.

$g(y_{t+1}|y_t;\theta)$ denotes the true one step ahead conditional transition density obtained from the strong solution of the SDE;

$f(y_t|y_s;\theta)$ denotes an approximation to $g(y_t|y_s;\theta)$;

$f^M(y_t|y_s)$ denotes an approximation to $g(y_t|y_s;\theta)$, based on introducing M additional auxiliary points between the time interval (s,t);

$p(\theta)$ denotes the prior density of θ;

y_t, Y, x_t and X are used to denote observed quantities;

y_t^*, Y^*, x_t^* and X^* are used to represent the latent or auxiliary data;

c is a generic constant used in the log-likelihood expressions, unless stated otherwise;

[2] Economics Discussion paper, 146, Nuffield College.
[3] Nuffield Economics Discussion List, W18.

M_t refers to the number of auxiliary points introduced between the data points y_t and y_{t+1}. In most cases, $M_t = M$ for all t;

N refers to the number of iterations of the sampler;

B_N refers to the number of pre-specified lags used to estimate the autocorrelation function;

E is the expectation operator and var denotes the variance;

$\approx$ is approximately equal to;

$\sim$ is distributed as;

$\overset{\text{approx.}}{\sim}$ is approximately distributed as;

$\propto$ is proportional to;

$h \to 0^+$ denotes the convergence of h to zero from above;

$\xrightarrow{\text{a.s.}}$ tends to almost surely;

$\xrightarrow{\text{p}}$ tends to in probability;

$\xrightarrow{\text{d}}$ tends to in distribution;

for notational clarity, bracket ordering will be $[\{(\cdot)\}]$;

for two functions $f_1 = f_1(t)$ and $f_2 = f_2(t)$, such that $f_1 \leq Cf_2$ as $t \to 0$, for some constant C, the order notation can be used to express this as $f_1 = O(f_2)$. If $f_1/f_2 \to 0$ as $t \to 0$, then we use $f_1 = o(f_2)$;

$(\Omega, \mathcal{F}, P)$ is a probability space;

$L_p[t_0, T]$ is the set of functions $f(s)$, measurable with respect to $(\Omega, \mathcal{F}, P)$, such that $\int_{t_0}^{T} |f(s)|^p ds < \infty$;

$L^p = L^p(\Omega, \mathcal{F}, P)$ denote random variables satisfying $\mathrm{E}|X|^p < \infty$;

$\mathcal{B}^d$ denotes the $\sigma-$algebra of Borel sets in $\mathbb{R}^d$;

$\mathbb{R}^d$ denotes the $d-$dimensional Euclidean space with distance function $|x - y|$ and we assume that $\mathbb{R}^d$ is endowed with the $\sigma-$algebra $\mathcal{B}^d$ of *Borel* sets.

The rest of Part I focuses on the background material on stochastic differential equation theory and simulation estimation methods, which are used in the remaining parts of the thesis.

CHAPTER 2

STOCHASTIC DIFFERENTIAL EQUATIONS

This Chapter presents the background material on stochastic differential equation theory and their estimation. General definitions and notations are outlined in Section 2.1. It will be important to introduce additional concepts used in alternative approaches to estimation of these equations. Martingales, for example, are used in many of the methods presented in the literature review, though they are not specifically used as a tool in the proposed approach. The exposition is informal; formal treatments can be found in the references provided below. Section 2.2 gives examples of some of the processes to be considered in the thesis, while Section 2.3 discusses issues surrounding discretisation of the continuous-time process. Augmentation issues are covered in Section 2.4 and a literature review of existing estimation methods is presented in Section 2.5. The Chapter concludes with Section 2.6.

The underlying sample paths for the random variables will be assumed to follow diffusions. These are continuous, but nowhere differentiable and are often used to specify an underlying model for financial data because of their Markovian structure. The theory behind Markov processes was initiated in 1906 by A.A. Markov, describing a process that depends only on the current information and so has no memory. If the state of a system at time t is known, then any additional information concerning the state at time $s < t$ has no effect on our knowledge of the probabilistic development of the system at $s > t$. For a comprehensive overview of models that have been considered in the finance literature, see for example, Merton (1990), Dixit (1993), Rossi (1996) and Campbell, Lo, and Mackinlay (1997). Issues concerned with the application of continuous-time analysis are discussed in Karatzas and Shreve, (1991, pp. 371-386), Cootner, (1964, pp. 79-83 and 189-197), Fama (1970) and briefly in Merton (1990).

2.1 DEFINITIONS AND NOTATION

The outline provided below is included as a preliminary and is not central to the contribution of the thesis. The content serves as a background, introducing both assumptions underlying a non-linear stochastic differential equation and the concepts used in the literature review. For those interested only in the method, this Section can be skipped on first reading.

Diffusion processes can either be defined through conditions on the transition densities or in terms of the random variable and its variation with respect to time. We are interested in the latter, as this leads to a stochastic differential equation in the random variable and we rely on probabilistic methods used in the calculus of stochastic differential equations. A full treatment of the theory of diffusions and stochastic differential equations, amongst others, is given in Itô and Mckean (1964), Arnold (1974), Itô (1987), and Rogers and Williams, (1994 and 1996). Treatments of the theory of stochastic processes can be found in

Doob (1953), Cox and Miller (1968), Karlin and Taylor (1975) and Bhattacharya and Waymire (1990). For issues concerning the statistical treatment of stochastic processes, see Basawa and Rao (1980) and for limit theorems for stochastic processes, see Jacod and Shiryayev (1987). For references on general probability concepts, see Loève, (1977, 1978) or Feller, (1968 and 1971) and Shiryayev (1996), which are often cited as classic references. For concepts of stochastic integration, Protter (1995) and Chung and Williams (1983) can be consulted; concise treatments can be found in Dachunha-Castelle and Duflo, (1986, pp. 331-389) and Karatzas and Shreve, (1991, Ch. 3).

Suppose we take the probability space $(\Omega, \mathcal{F}, P)$ and a non-decreasing family $\{\mathcal{F}_t,\ t \geq 0\}$ of σ-algebras in $\mathcal{F}$, the filtration up to time t. A random variable that is measurable is assumed to be measurable with respect to the family of $\sigma-$algebras just defined. For a thorough treatment of these concepts, see Billingsley (1961).

We define a continuous-time stochastic process on the state space $\mathbb{R}^d$ as a sequence of random variables $\{y(t)\}$, where the index set is $[t_0, T] \subset [0, \infty) = \mathbb{R}^+$, for finite and non-negative t_0. The time set under consideration is $[t_0, T]$, so that the stochastic process is defined as $y(t) = \{y(t) : t \in [t_0, T]\} : [t_0, T] \to \mathbb{R}^d$. Hence $y(t) = y(t, \cdot)$ is a random variable for each $t \in [t_0, T]$ and $y(\cdot, \omega) : [t_0, T] \to \mathbb{R}^d$ is a realization of the stochastic process for each $\omega \in \Omega$.[1] P is a probability measure on $\mathcal{F}$, the space of all $\mathbb{R}^d-$valued functions defined on $[t_0, T]$. Further information on the analytical properties of $y(t)$ allows us to choose a subspace of $(\mathbb{R}^d)^{[t_0,T]}$ for Ω – for example the space of uniformly continuous functions.

2.1.1 The Markov process as a diffusion process

A good treatment of probability theory for Markov processes as diffusion processes can be found in Arnold, (1974, Ch. 1 and 2). For a reference on stochastic differential equations and diffusions, see Rogers and Williams, (1996, Ch. 5).

An *Itô process* can be defined as a stochastic process that satisfies a *stochastic differential equation*, (SDE), of the form

$$dy(t) = a\{y(t), t; \theta\}dt + b\{y(t), t; \theta\}dW(t), \tag{2.1}$$

where the SDE is assumed to be a homogeneous Markov process. The non-anticipative drift function $a\{y(t), t; \theta\} : \mathbb{R}^d \times \mathbb{R}^+ \times \mathbb{R}^p \to \mathbb{R}^{d\times 1}$ and the volatility function $b\{y(t), t; \theta\} : \mathbb{R}^d \times \mathbb{R}^+ \times \mathbb{R}^p \to \mathbb{R}^{d\times d}$ are measurable functions, depending on the paths $y(t)$, time t, and an unknown parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$. In most cases, we will concentrate on the $d = 1$ case.

2.1.2 The solution to the SDE with a homogeneous Lévy driving process

Equation (2.1) is often written as

$$y(t) = y(t_0) + \int_{t_0}^{t} a\{y(\cdot), s; \theta\}ds + \int_{t_0}^{t} b\{y(\cdot), s; \theta\}dW(s). \tag{2.2}$$

Two processes defined on the same space are said to be equivalent if their finite-dimensional distributions coincide, so that each process is simply a *version* of the other. Note that dependence on t implies that the

[1] A realization of the random variable is also termed a trajectory, a sample function or a path.

paths of these sample functions can have very different properties. Equation (2.1) defines $y(t)$ implicitly; under certain conditions imposed on (2.1), $y(t)$ is the solution of the equation for an arbitrary initial value and a Markov process on $[t_0, T]$, with distribution function $G\{y(t)|y(s)\}$ and an initial probability P_{t_0}.

Unless otherwise stated, we will assume the following Lipschitz and linear growth conditions on the given drift and volatility functions at an arbitrary, but fixed time t,

Measurability condition

$$a(y,t) \text{ and } b(y,t) \text{ are jointly } \mathcal{L}^2 - \text{measurable in } (y,t). \tag{2.3}$$

Lipshitz condition There exists a constant $K > 0$ such that

$$\begin{aligned} |a(x,t) - a(y,t)| &\leq K|x-y| \\ |b(x,t) - b(y,t)| &\leq K|x-y| \end{aligned} \tag{2.4}$$

for all $t \in [t_0, T]$ and x, $y \in \mathbb{R}$.

Linear growth bound There exists a constant $K > 0$ such that

$$\begin{aligned} |a(x,t)|^2 &\leq K^2(1+|x|^2) \\ |b(x,t)|^2 &\leq K^2(1+|x|^2) \end{aligned} \tag{2.5}$$

for all $t \in [t_0, T]$ and x, $y \in \mathbb{R}$.

Under (2.3–2.4) pathwise uniqueness can be shown for solutions to (2.2) corresponding to the same initial value and the same Wiener process, assuming that the strong solution exists, (see, Kloeden and Platen, 1992, pp. 128-129). If, in addition to (2.3–2.5), we assume that the initial value is $\mathcal{F}_{t_0}$-measurable, with finite absolute second moment, then the SDE defined in (2.1) has a pathwise unique strong solution $y(t)$ on $[t_0, T]$ with $\sup_{t_0 \leq t \leq T} \mathrm{E}\{|y(t)|^2\} < \infty$, see Kloeden and Platen, (1992, p. 131).

Weak solutions

A weak solution of (2.1) is any stochastic process that obeys the same probability laws as the strong solution. A strong solution implies that there is a solution for each given Wiener process and we can consider a separable version which has almost surely continuous sample paths. Formally, this requires that for a given Wiener process, the solution to (2.1) is $\mathcal{F}_t$-adapted, where $\mathcal{F}_t$ is the σ-algebra generated by the Wiener process. When no strong solution is available, we can find a weak solution in the sense that we are free to select a particular Wiener process and then find a solution corresponding to that process. For additional details, see Kloeden and Platen, (1992, Chapter 4.5) or Rogers and Williams, (1996, Ch. 5, Section 5).

The driving process

We will focus on the Wiener process as the driving process for the SDE, denoted by $W(t)$. This is an example of a homogeneous Lévy process, that is, a process with independent increments (Lévy 1937),

which is continuous in probability, see for example Rogers and Williams, (1996, pp. 308-313) or Meyer (1975). For a brief introduction to the theory of Lévy processes, see Rogers and Williams, (1996, pp. 73-84). By Lévy's theorem, (see for example Itô and Mckean, 1964, p. 146), the driving process can contain jumps. A special case of a homogeneous Lévy process which does not include a jump component, is the Wiener or Brownian motion process.[2] For a treatment on Brownian motion, see Borodin and Salminen, (1996, Ch. 4). We shall use Brownian motion and Wiener processes interchangeably. The two conditions under which the Wiener process is defined are given by:

For $0 \leq s < t < \infty$, $W(t) - W(s)$ is independent of $\mathcal{F}_s$; that is, the increments are independent of the past.

For $0 < s < t$, $W(t) - W(s)$ is a Gaussian random variable with mean 0 and variance $(t-s)C$, where C is a non-random matrix.

Existence of the Wiener process was proved by Ciesielski (1961), using Kolmogorov's extension theory and a comprehensive treatment is given in Protter (1995).[3] We will assume that the conditions under which the SDE exists and is unique are satisfied. We will additionally assume that it can equivalently be represented as a diffusion process, (see Øksendal 1995, page 64 for more detail). Functional central limit theorems for continuous processes can be found in Dachunha-Castelle and Duflo, (1986, Ch. 7.4).

Given a family of probability measures on $\{(\mathbb{R})^d\}^{\mathbb{Z}^+}$, the stochastic process can be constructed with those measures as its finite-dimensional distribution under the consistency conditions given by *Kolmogorov's extension theory*; see for example, Rogers and Williams, (1994, Ch. 1, Section 28). For a given probability measure or law μ, with characteristic function $\mathbb{F}_\mu$, μ is said to be ***infinitely divisible*** if, for any positive integer n, there exists a probability measure μ_n such that $\mathbb{F}_\mu = (\mathbb{F}_{\mu_n})^n$. That is, μ can be expressed as a convolution power of μ_n. The Gaussian density is one of many examples, see Rogers and Williams, (1994, pp. 73-75) or Feller, (1971, Ch. 17). Feller, (1971, pp. 557-558) gives an application of the concept of infinitely divisibility to processes with independent increments. In brief, if W is a Lévy process, the law of $W(1)$ is infinitely divisible and conversely, any infinitely divisible law is the law of $W(1)$ for some Lévy process. This result allows us to pass from an analytical description of the process, (through e.g. the characteristic function), to a description involving sample paths. If the increments are independent, but non-stationary, the result still generalises. This means that we are not restricted to considering stationary processes.

Issues that typically arise, concern changes in the model if the time interval dt changes and the assumption of Gaussianity on $dW(t)$. If the drift term is linear in $y(t)$ and the volatility is a constant, for example, σ^2, then a closed form solution exists and the model is the same whether we change the time interval or the distribution on the increments of $W(t)$. If $W(t)$ is a homogeneous Lévy process, then whether the model's distribution changes as dt changes depends on the form of $b\{y(t), t; \theta\}$; no statements can be made for a general Lévy process. The dependence of a solution on the conditions of linearity on the drift, the distribution of $dW(t)$ and the form of $b\{y(t), t; \theta\}$ can be seen by writing the solution as

$$\begin{aligned} y(t+1) &= y(t) + \int_t^{t+1} \mathrm{E}\left[a\{y(s), s; \theta\} | \mathcal{F}_t\right] ds + \int_t^{t+1} \left(a\{y(s), s; \theta\} - \mathrm{E}\left[a\{y(s), s; \theta\} | \mathcal{F}_t\right]\right) ds \\ &\quad + \int_t^{t+1} b\{y(s), s; \theta\} \, dW s. \end{aligned}$$

[2] By Lévy's theorem, Brownian motion with drift is the only continuous Lévy process without jumps.

[3] An initial proof was given by Wiener, who proved its existence by representing it as a random Fourier series.

2.1.3 Martingales, compensators and the infinitesimal generator

For a rigorous treatment on concepts in martingale theory, see Feller, (1971, Ch. 6, Section 12) or Dachunha-Castelle and Duflo, (1986, Ch. 7.2). Martingales and compensators are used predominantly in the estimating functions literature, outlined in Section 2.5.2.

2.1.4 The Radon-Nikodym derivative and Girsanov's theorem

These concepts are used in the literature review and in particular, Section 2.5.1. The concepts are not central to the thesis and can be skipped by the reader on first reading.

The *Radon-Nikodym* (R-N) derivative and *Girsanov's* theorem allow the change of a probability measure in continuous-time models. An important application relates to computing arbitrage-free prices, through the use of *absolute continuous* measure transformations. Under an equivalent martingale measure, the price of any financial derivative can be expressed as the expectation of a discounted payoff function. Their existence ensures an arbitrage-free market and vice versa. Use of the R-N derivative and Girsanov's theorem can also be used to obtain the maximum likelihood estimates. Formal treatments can be found in Karatzas and Shreve (1991) and Ikeda and Watanabe (1981). Alternative references are given by Arnold, (1974, pp. 12 and 111) and Kloeden and Platen, (1992, pp. 58, 60, 154 and 512). Concise references for Girsanov's theorem are given in Øksendal, (1995, pp. 146-151).

2.1.5 Finite difference methods

This Section is not used in the original contribution to the thesis and is only relevant to the literature review in Section 2.5.7.

Many existing methods are based on finding numerical solutions to a diffusion equation and by a change of variable converting these into numerical solutions of the original system, (see for example, Wilmott, Howison, and Dewynne 1995, 1995, Section 8). The finite-difference methods entail obtaining numerical solutions to a system of partial differential equations. Direct application of these methods is not generally possible. Essentially, the method requires replacing the partial derivatives by approximations involving Taylor expansions. For further references on finite difference methods, see Richtmyer and Morton (1967) and Smith (1985a). For an application in option pricing, see Brennan and Schwartz (1978).

2.2 Examples of processes to be considered

Specific examples include the

- *Geometric Brownian motion* (GBM) is often motivated by considering independent and identically distributed (i.i.d.) percentage changes that is, $r(t)/r(t-1)$ are close to being i.i.d. Then $\log r(t) = \sum_{i=1}^{t} \log\left\{\frac{r(i)}{r(i-1)}\right\}$ normalised, is approximately Brownian motion and so $r(t)$ is approximately GBM. For example, suppose one has the option but not the obligation of purchasing one unit of a stock at a fixed price K at time T in the future. If prices vary according to a BM with unit volatility and the present value of the stock is denoted S_p, then the expected worth of owning the

option can be calculated as

$$\begin{aligned} \mathrm{E}\left[\max\{S(T)-K,0\}\right] &= \int_0^\infty P\left\{S(T)-K>a\right\}da \\ &\quad \text{since the option is only exercised if the price at time } T \text{ is} \geq K \\ &= \int_0^\infty \int_{\log\left(\frac{K+a}{S_p}\right)}^\infty \frac{1}{\sqrt{2\pi T}} \exp\left(-\frac{x^2}{2T}\right) dxda. \end{aligned}$$

The drift and volatility functions are given by $a\{y(t),t;\theta\} = \mu y(t)$ and $b\{y(t),t;\theta\} = \sigma y(t)$, for $\theta = (\mu,\sigma)$. The strong solution is explicitly available.

- *Ornstein-Uhlenbeck* (OU) process with $\theta = (\mu,\sigma)$, where $a\{y(t),t;\theta\} = \mu y(t)$, and $b\{y(t),t;\theta\} = \sigma$. This process has strong solution which is available explicitly.

- *Quadratic drift* process with $\theta = (\mu,\sigma)$, where drift is a quadratic function of $y(t)$, ie. $a\{y(t),t;\theta\} = \mu y(t)^2$ and the volatility, $b\{y(t),t;\theta\} = \sigma$. No analytical expression for the strong solution exists.

- In the literature on interest rates, the OU process with a non-constant mean, with $\theta = (\alpha,\beta,\sigma)$, where $a\{y(t),t;\theta\} = \alpha - \beta y(t)$ and $b\{y(t),t;\theta\} = \sigma$, is called the *Vasicek* model. The parameters α,β and σ are all positive and constant. As this is just the mean reverting version of the OU process, an analytical form for the strong solution exists.

- The *Cox, Ingersoll and Ross* (CIR) model of Cox, Ingersoll, and Ross (1985), with $\theta = (\alpha,\beta,\sigma)$ where $a\{y(t),t;\theta\} = \alpha - \beta y(t)$ and $b\{y(t),t;\theta\} = \sigma\sqrt{y(t)}$. The parameters α,β and σ are all positive and constant. This process is always positive unless absorbed at zero. It has a known strong solution. The conditional transition density is given by a non-central chi-squared distribution. The process oscillates stochastically around $-\frac{\alpha}{\beta}$, but tends to be stationary. It is often used to value bonds where the price of a bond is $\exp\{-\int_0^T y(s)ds\}$, where T is a fixed time point. $\mathrm{E}[\exp\{-\int_0^T y(s)ds\}]$ can be calculated explicitly as shown in Cox, Ingersoll, and Ross (1985) or using techniques by Revuz and Yor (1991). It is typically cited as an example of a *Constant Elasticity of Variance* (CEV) model which can be solved explicitly.

- The *Hull-White model*, (Hull and White 1990 and Taylor 1982), also known as the Generalised CIR, with $\theta = (\alpha,\beta,\sigma)$ where $a\{y(t),t;\theta\} = \alpha - \beta y(t)$ and $b\{y(t),t;\theta\} = \sigma y(t)^\gamma$. The parameters are constrained by α,β,γ and σ are all positive constants. No analytical expression for the strong solution is available. An example of a *Constant Elasticity of Variance* (CEV) model, where $\gamma \neq (1/2)$.

- The *Affine Cox, Ingersoll and Ross* model, with $\theta = (\alpha_0,\alpha_1,\beta_0,\beta_1)$ where $a\{y(t),t;\theta\} = \alpha_0 + \alpha_1 y(t)$ and $b\{y(t),t;\theta\} = \{\beta_0 + \beta_1 y(t)\}^{\frac{1}{2}}$. The parameters $\alpha_0 \geq 0$, $\alpha_1 \leq 0$, $\beta_0 > 0$ and $\beta_1 \geq 0$. The strong solution cannot be written down explicitly.

- The *general parametric* model used, studied in Aït-Sahalia (1996b), with $\theta = (\alpha_0,\alpha_1,\alpha_2,\alpha_3,\beta_0,\beta_1,\beta_2,\beta_3)$, specifies $a\{y(t),t;\theta\} = \alpha_0 + \alpha_1 y(t) + \alpha_2 y(t)^2 + \frac{\alpha_3}{y(t)}$ and $b\{y(t),t;\theta\} = \beta_0 + \beta_1 y(t) + \beta_2 y(t)^{\beta_3}$. Con-

ditions on the parameters are given by

$$\begin{aligned}
&\beta_0 \geq 0 \text{ (and } \beta_2 > 0 \text{ if } \beta_0 = 0 \text{ and } 0 < \beta_3 < 1, \text{ or } \beta_1 > 0 \text{ if } \beta_0 = 0 \text{ and } \beta_3 > 1) \\
&\beta_2 > 0 \text{ if either } \beta_3 > 1 \text{ or } \beta_1 = 0, \text{ and } \beta_1 > 0 \text{ if either } 0 < \beta_3 < 1 \text{ or } \beta_2 = 0 \\
&\alpha_2 \leq 0 \text{ and } \alpha_1 < 0 \text{ if } \alpha_2 = 0 \\
&\alpha_3 > 0 \text{ and } 2\alpha_3 \geq \beta_0 \geq 0, \text{ or } \alpha_3 = 0, \alpha_0 > 0, \ \beta_0 = 0, \ \beta_3 > 1 \text{ and } 2\alpha_0 \geq \beta_1 > 0.
\end{aligned}$$

The first two equations are necessary for $b^2(\cdot)$ to be positive, the third ensures that the drift is mean reverting at high interest values while the last equation rules out the zero case. No analytical form for the strong solution exists.

- The *hyperbolic diffusion*, see Barndorff-Nielsen (1977), is defined as

$$dS(t) = S(t)\left(\left[\kappa + \frac{1}{2}v^2\{\log S(t) - \kappa t\}\right]dt + v\{\log S(t) - \kappa t\}dW(t)\right),$$

where $S(t)$ denotes the stock price. One can obtain different models on appropriate choice of $v(x)$; we choose the diffusion

$$v^2(x) = \sigma^2 \exp\left\{\alpha\sqrt{\delta^2 + (x-\mu)^2} - \beta(x-\mu)\right\},$$

which corresponds to the logarithm of the stock price (eventually) approximately following a hyperbolic distribution. The density of the hyperbolic is given by

$$h(x;\alpha,\beta,\delta,\mu) = \frac{\sqrt{\alpha^2-\beta^2}}{2\alpha\delta K_1(\delta\sqrt{\alpha^2-\beta^2})} \exp\left\{-\alpha\sqrt{\delta^2+(x-\mu)^2} + \beta(x-\mu)\right\}, \quad x \in \mathbb{R}.$$

The parameters $(\alpha, \beta, \delta, \mu)$ are such that $\mu \in \mathbb{R}$, $\delta > 0$ and $0 \leq |\beta| < \alpha$. K_τ denotes the modified Bessel function of the third kind with index τ. The generalised hyperbolic distribution is the mixture of a normal and a generalised inverse Gaussian distribution. The parameter α is interpreted as a gradient parameter, (the distribution is steeper the larger the α), β is the parameter of asymmetry, ($\beta = 0$ represents symmetry around the mean), δ is a scale parameter and μ the location parameter.

2.3 Discretizing the continuous-time process

The aim is to be able to estimate the parameters of the process accurately, when we have observed data $y_t = y(\tau_t)$ at time points $\tau_1, \ldots, \tau_T$, based on the log-likelihood function $l_T(\theta) = \log L(y_2, \ldots, y_T | y_1; \theta) = \sum_{t=1}^{T-1} g(y_{t+1}|y_t;\theta)$, where $g(y_{t+1}|y_t;\theta)$ denotes the one-step ahead conditional transition density obtained using the strong solution. If the strong solution of the underlying process is explicitly available, then the SDE in equation (2.1) can be solved analytically using conventional Itô calculus, implying that $g(y_{t+1}|y_t;\theta)$ is available in closed form and maximum likelihood methods can be applied directly. In general, however, closed-form solutions are rarely available for most SDE's and recent developments in the econometric literature have focused on methods for estimating non-linear diffusions when only discrete observations are available. The method updates the states conditional on θ and then updates θ conditional on the observed and auxiliary points, so that θ can be assumed known when considering

the updating of the states.

To invoke the discretisation scheme, we first assume that the Itô process under consideration in general satisfies the stochastic differential equation, given in (2.2), in Itô form, for $t \in (t_0, T]$. We shall use a stochastic Taylor expansion to derive a time discrete approximation with respect to the strong convergence criterion, which will be defined in (2.6). This implies the strong convergence of the approximation based on a discretised version, denoted by $y_t = y(\tau_t)$, to the state at time t based on the Itô process. For proofs on convergence results, see Kloeden and Platen, (1992, pp. 323-331, pp. 340-350).

2.3.1 Convergence of the discretised version to the Itô process

A time-discrete approximation y_T, with maximum step size $\delta = \min_{t=1}^{T} \Delta_t^{\dagger}$, *converges strongly* to $y(T)$, with order $\gamma > 0$, if there exists a positive constant C, which does not depend on γ, and a $\delta_0 > 0$ such that,

$$\epsilon(\delta) = E\{|y(T) - y_T|\} \leq C\delta^{\gamma} \tag{2.6}$$

for each $\delta \in (0, \delta_0)$, where the expectation is taken with respect to the random variable $y(T)$. This is known as the *absolute error* of the discretisation scheme. If there exists a non-negative function, $c = c(\delta)$ with $\lim_{\delta \to 0+} c(\delta) = 0$, such that

convergence in mean

$$\mathrm{E}\left\{\left|\mathrm{E}\left(\frac{y_{t+1} - y_t}{\Delta_t}\middle| \mathcal{F}_t\right) - a(y_t, \tau_t; \theta)\right|^2\right\} \leq c(\delta) \tag{2.7}$$

variance of difference converges to zero

$$\mathrm{E}\left\{\frac{1}{\Delta_t}\left|\{y_{t+1} - y_t - \mathrm{E}(y_{t+1} - y_t|\mathcal{F}_t)\}^2 - b(y_t, \tau_t; \theta)(W_{t+1} - W_t)\right|^2\right\} \leq c(\delta) \tag{2.8}$$

for all discretisations τ_t and points y_t, then the discrete-time approximation is *strongly consistent.* If we let $z(t)$ represent the upperbound between the squared absolute value of the Itô process and the approximation at time t, then for $y_0 = y(0)$,

$$\begin{aligned} z(t) &= \sup_{0 \leq s \leq t} \mathrm{E}\left(|y_s - y(s)|^2\right) \\ &= \sup_{0 \leq s \leq t} \mathrm{E}\left\{\left|\sum_{t=0}^{s-1}(y_{t+1} - y_t) - \int_0^s a(y_r, \tau_r; \theta)\, d\tau_r - \int_0^s b(y_r, \tau_r; \theta)\, dW_r\right|^2\right\} \\ &\leq C\{\delta + c(\delta)\} \end{aligned} \tag{2.9}$$

using the Lipschitz and linear growth conditions on the drift and volatility coefficients in (2.3–2.5), an estimate of the second moment of $y(t)$, the conditions (2.7–2.8) and the Gronwall inequality in Appendix 2.7.1.

To ensure convergence of the density under the discretisation scheme, we consider a time-discrete approximation y_T, with maximum step size $\delta = \min_{t=1}^{T} \Delta_t^{\dagger}$, *converges weakly* to $y(T)$ as $\lim \delta \to 0^+$, if for

all continuous functions g

$$\lim_{\delta \to 0^+} |\mathrm{E}[g\{y(T)\}] - \mathrm{E}\{g(y_T)\}]| = 0.$$

Comparisons for different approximations can be assessed through their respective rates of convergence. The discrete-time approximation y_t converges weakly with order $\beta > 0$ to $y(T)$ as $\delta \to 0^+$, if for each $g \in C^{2(\beta+1)}$, (the space of twice continuously differentiable functions, which with partial derivatives of orders up to and including the second, have polynomial growth), there exists a positive constant C, which is independent of δ and $\delta_0 > 0$, such that

$$|\mathrm{E}[g\{y(T)\}] - \mathrm{E}\{g(y_T)\}| \leq C\delta^{\beta},$$

for each $\delta \in (0, \delta_0)$, see Kloeden and Platen, (1992, pp. 326-327).

A time-discrete approximation y_T, with maximum step size $\delta = \min_{t=1}^{T} \Delta_t^{\dagger}$ is weakly consistent, if there exists a non-negative function $c = c(\delta)$ with $\lim_{\delta \to 0^+} c(\delta) = 0$, such that (2.7) holds and

$$\mathrm{E}\left[\left|\mathrm{E}\left\{\frac{1}{\delta}(y_{t+1} - y_t)^2 | \mathcal{F}_t\right\} - b^2(y_t, \tau_t; \theta)\right|^2\right] \leq c(\delta), \tag{2.10}$$

for all y_t and discretisations τ_t. Condition (2.10) is weaker than (2.8) as only the variance of the approximate scheme needs to be close to that of the Itô process.

2.3.2 The Euler-Maruyama scheme

The Euler-Maruyama scheme is defined as

$$y_{t+1} = y_t + a(y_t, t; \theta)\Delta_t^{\dagger} + b(y_t, t; \theta)(W_{t+1} - W_t), \tag{2.11}$$

where $\Delta_t^{\dagger} = \tau_{t+1} - \tau_t$ represents the time difference between y_t and y_{t+1}. This can also be expressed as

$$y_{t+1}|y_t \overset{approx.}{\sim} N\{y_t + a(y_t, t; \theta)\Delta_t^{\dagger}, b^2(y_t, t; \theta)\Delta_t^{\dagger}\}. \tag{2.12}$$

The approximate Gaussian conditional transition density implied by invoking the Euler scheme will be denoted by $f(y_{t+1}|y_t; \theta)$. If the volatility function is given by $b(y_t, \tau_t; \theta) = b(\tau_t; \theta)$, so that $\partial b/\partial y = 0$ and noise is additive, then the Euler attains an order of convergence $\gamma = 1$.

Under the assumptions given in Appendix 2.7.1,

$$\mathrm{E}\{|y(T) - y_T\} \leq K\delta^{(1/2)},$$

where K is independent of δ. Using (2.9), it can be shown that

$$z(t) \leq C\left[|y(0) - y_0|^2 + \{1 + |y(0)|\}^2\delta\right]$$

for a constant C and $t \in [0, T]$, see Kloeden and Platen, (1992, pp. 343-344). The Euler approximation attains a weak order of convergence of $\beta = 1$, if in addition to the assumptions in Appendix 2.7.1, the drift and volatility functions are four times continuously differentiable.

2.3.3 The Milstein scheme and higher order Taylor expansions

The Milstein scheme can be expressed in terms of the Euler approximation with an additional quadratic term which depends on the square-root of the volatility, its derivative with respect to the states, the squared increment of the W_t process and the time step, $\Delta^\dagger$. Use of the Milstein scheme gives us a non-central chi-squared distribution.

The Milstein approximation attains an order of strong convergence of $\gamma = 1.0$ in general. If $\partial b(y_t, \tau_t; \theta)/\partial y_t = 0$, the Milstein scheme collapses to the Euler so that the latter attains the same order of convergence as the Milstein scheme. In addition, the Milstein approximation attains a weak order of convergence of $\beta = 2$, if in addition to the assumptions in Appendix 2.7.1, the conditional moments of $y_{t+1}|y_t$ closely approximate those of the Itô process, the drift and volatility functions are sufficiently smooth and a regularity condition on the time discretisation, see Kloeden and Platen, (1992, pp. 374-375) for more detail. More detail will be provided in Section 9.1.

For this approximation we have

$$\begin{aligned} y_{t+1} &= y_t + a\Delta + b(\Delta W) + \frac{1}{2}bb'\{(\Delta W)^2 - \Delta\} + a'b(\Delta Z) + \frac{1}{2}(aa' + \frac{1}{2}b^2a'')\Delta^2 \\ &+ (ab' + \frac{1}{2}b^2b'')\{\Delta(\Delta W) - (\Delta Z)\} + \frac{1}{2}b(bb'' + (b')^2)\{\frac{1}{3}(\Delta W)^2 - \Delta\}(\Delta W) \end{aligned}$$

where $(\Delta Z) = \int_{\tau_t}^{\tau_{t+1}} \int_{\tau_t}^{s_2} dW_{s_1}\, ds_2$, that is

$$(\Delta Z) \sim N(0, \frac{1}{3}\Delta^3)$$

and $E\{(\delta Z)(\Delta W)\} = \frac{1}{2}\Delta^2$. Here we have a polynomial in $(\Delta W)^3$, which implies that we can sample from the density, but unless we have a closed form, we cannot evaluate the density explicitly.

2.4 Augmentation: introducing auxiliary points

Suppose we are presented with a discrete set of observations from some financial time series, $y_1, \ldots, y_T$. If the strong solution of the underlying process is known explicitly, $g(y_{t+1}|y_t; \theta)$ will be known. In general, the strong solution may not be known and so we take a discretisation of the time interval between each observed value. Now between any two points $(t, t+1)$, as in Figure 2.1, consider introducing auxiliary points $y_t^* = (y_{t,1}^*, \ldots, y_{t,M_t}^*) \in \mathbb{R}^M$ at times $\tau_{t,1} < \ldots < \tau_{t,M_t}$, where $y_{t,j}^* = y^*(\tau_{t,j})$, for $j = 1, \ldots, M_t$.

The aim will be to construct a method of sampling the values of the unobserved states of the process, conditional on the available observations, $y_{t,0}$ and $y_{t,M+1}$. Our approach is to generate M auxiliary variables y_t^* between each pair of successive observed values y_t and y_{t+1}. The general approach is given in 2.4.1. This idea was initially used by Pedersen, (1995b, 1995a) in the context of likelihoods. We will then generate the parameters from their posterior distributions, given the available observations $Y = (y_1, \ldots y_T)'$ and $Y^* = (y_1^{*\prime}, \ldots, y_{T-1}^{*\prime})'$, which for these models will often be relatively easy to sample from. At first we consider only two observed values, y_t and y_{t+1} in our analysis for sampling Y^*, which is described in Sections 4 and 5. The method is then extended in Section 5, where additional issues such as block sampling are considered. Estimation of the parameter of interest and simulation from its posterior density are covered in 6.

To distinguish between methods which use additional auxiliary points and those that do not, we will let $f(y_{t+1}|y_t; \theta)$ denote the approximation to $g(y_{t+1}|y_t; \theta)$ without auxiliary points and $f^M(y_{t+1}|y_t; \theta)$

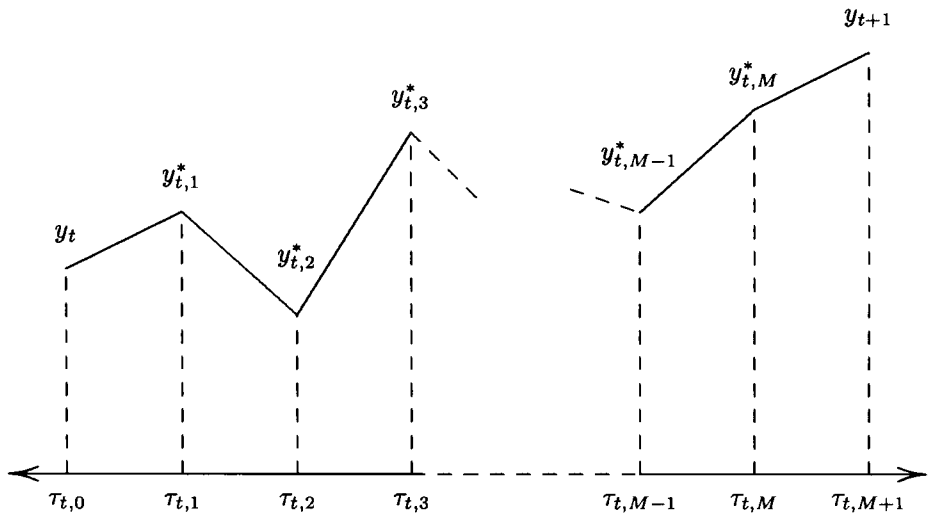


Figure 2.1: *Discretisation scheme is shown where the auxiliary points y_t^* are introduced between observed values y_t and y_{t+1}.*

denote the approximation with M additional points between each pair of observations. The introduction of auxiliary points will be shown to provide a better approximation of the conditional transition density. Convergence of the approximate conditional transition density to the true density is discussed in 2.4.2, where we observe that $f^M(y_{t+1}|y_t;\theta) \xrightarrow{\text{a.s.}} g(y_{t+1}|y_t;\theta)$ as $M \to \infty$. Section 2.4.3 considers the OU process due to the availability of a closed-form expression for the strong solution and analytical tractability of the transition density of y_t^*, given both neighbouring points. Further, we are now in a position to control the degree of discretisation bias through our choice of M. Since for a linear process, such as the OU, the strong solution is known, direct sampling from the density is possible. This provides us with a benchmark, which we can use to compare the performance of our method using a discretised version of the process. This is formalised in section 4.2, where we discuss use of the simulation smoother. In the case of a linear process, we can also perform Gibbs sampling, which involves sampling from the full conditional distribution. In general, however, it is intractable and we use the M-H algorithm.

2.4.1 General approach

The general approach will be to approximate a transition density $g(y_{t+1}|y_t;\theta)$ by

$$\begin{aligned} f^M(y_{t+1}|y_t;\theta) &= \int f(y_{t+1}|y^*_{t,M_t};\theta)\left\{\prod_{j=2}^{M_t} f(y^*_{t,j}|y^*_{t,j-1};\theta)\right\} f(y^*_{t,1}|y_t;\theta)dy^*_{t,M_t},\ldots,dy^*_{t,1} \\ &= \int f(y_{t+1}|y_t^*;\theta)f(y_t^*|y_t;\theta)dy_t^* \end{aligned} \tag{2.13}$$

where $y_t^* = (y^{*\prime}_{t,1},\ldots,y^{*\prime}_{t,M_t})' \in \mathbb{R}^M$ are the auxiliary variables which are introduced to the system. As $M_t \to \infty$, for each $t = 1,\ldots,T$, this transition density will converge to the true transition density $g(y_{t+1}|y_t;\theta)$, (as shown in Pedersen 1995a and outlined in Section 2.4.2). For notational simplicity, we will assume that $\Delta_t^\dagger = \Delta^\dagger$ and $M_t = M$ are constants, and that the time points $\tau_{t,k}$ are evenly spaced for all t and k. Hence, $\tau_{t,k+1} - \tau_{t,k} = \frac{\Delta^\dagger}{M+1} = \Delta$ is constant for all t,k. The approach can easily be

generalised; allowing M_t to vary is more relevant in practice, particularly when $\Delta_t^\dagger$ varies with t. The analysis put forward here can be applied to both stationary and non-stationary cases. In the examples presented below, $d = 1$ and $W(t)$ is the standard Wiener process with increment variance equal to $\Delta_t^\dagger$, the time step between $y(t+1)$ and $y(t)$.[4]

2.4.2 Convergence of the approximate conditional transition density

Bally and Talay (1996a) show that for N independent trajectories of $y_t^\star \in \mathbb{R}^M$, the approximation to $f^M(y_{t+1}|y_t;\theta)$

$$\frac{1}{N}\left[\sum_{i=1}^{N}\prod_{j=0}^{M} f\{y_{t,j+1}^{*(i)}|y_{t,j}^{*(i)};\theta\}\right] \rightarrow \mathrm{E}\{f^M(y_{t+1}|y_t;\theta) = g(y_{t+1}|y_t;\theta)$$

as N and $M \rightarrow \infty$. The global error of the algorithm can be measured by $\mathrm{E}|f^M(y_{t+1}|y_t;\theta) - g(y_{t+1}|y_t;\theta)|$, using a special case of the result by Talay and Tubaro (1990). In fact, in general,

$$\mathrm{E}|f^{M_t}(y_{t+1}|y_t;\theta) - g(y_{t+1}|y_t;\theta)| \leq C\left(\frac{1}{\sqrt{N}} + \frac{\Delta_t^\dagger}{M_{t+1}}\right), \quad \text{where } C < \infty.$$

We will restrict attention to $M_t = M$ and $\Delta_t^\dagger = \Delta^\dagger$. Talay and Tubaro (1990) show that for smooth g, the error can be expanded in t terms of powers of M^{-1}, see Talay (1995) for a discussion. These results extend to situations where W_t is a more general Lévy process, see Protter and Talay (1997). The same expansion can also be established under weaker conditions, see Bally and Talay (1996b). These can be generalised to establishing the convergence of the log-density. The discretisation error is thus a function of M, rather than $\Delta^\dagger$ alone and can thus be controlled through choice of M. Additional proofs using Malliavin calculus can be found in Bally and Talay (1995).

Kanagawa (1995) studies the error in the solution to the SDE, using the Euler approximation, by using the distance between the standard Normal distribution function and the distribution function of quasi-Normal random numbers. Upperbounds on the discretisation bias for various expressions, including f^M are also studied in Milstein (1974), Pardoux and Talay (1985) and Talay (1995). Pointwise convergence and Malliavin calculus is used by Pedersen (1995b) to establish the behavioural limit of f^M. Specifically, in theorem 3 of Pedersen (1995b), conditions under which $f^M(y_{t+1}|y_t;\theta)$ converges to $g(y_{t+1}|y_t;\theta)$ are established. Kohatsu-Higa and Ogawa (1997) analyse rates of convergence using the Euler scheme; the proof relies on techniques of integration by parts from Malliavin calculus. The behaviour of estimators based on discretised versions of the SDE are also studied in Dorogovtçev (1982) using generalised least squares and Stoyanov (1984). Pardoux and Talay (1985) look at pathwise and mean-square convergence of several approximation schemes and estimate the speed of convergence through analysis of mean-square errors. They conclude that two schemes may compare differently depending on the criterion chosen. In their studies, the Milstein scheme seemed to perform relatively well.

[4] A more general formulation of (2.1) is given by (11.1), in Section 11, where the dimension of $y(t)$, (corresponding to the number of different SDE's), and the dimension of $W(t)$ are not necessarily the same. Specification (11.1) also allows us to incorporate dependence between the SDE's and between the Wiener processes.

2.4.3 The advantages of introducing auxiliary points

The Ornstein-Uhlenbeck (OU) process can be used to illustrate the effects of introducing auxiliary points due to its analytical tractability, [5] so that we can compare the performance of our method using a discretised version of the process against results obtained using the true process. Since the strong solution can be expressed in closed form, the one-step ahead conditional transition density implied by the strong solution is available analytically. Furthermore, the transition density can still be expressed in closed form after conditioning on both neighbours. The expected path between any two fixed observations can therefore be evaluated for different values of M and compared to the path implied by the strong solution and the discretisation bias can be shown graphically. The approximate and true likelihoods can also be compared, for different values of $\Delta^\dagger$ and M.

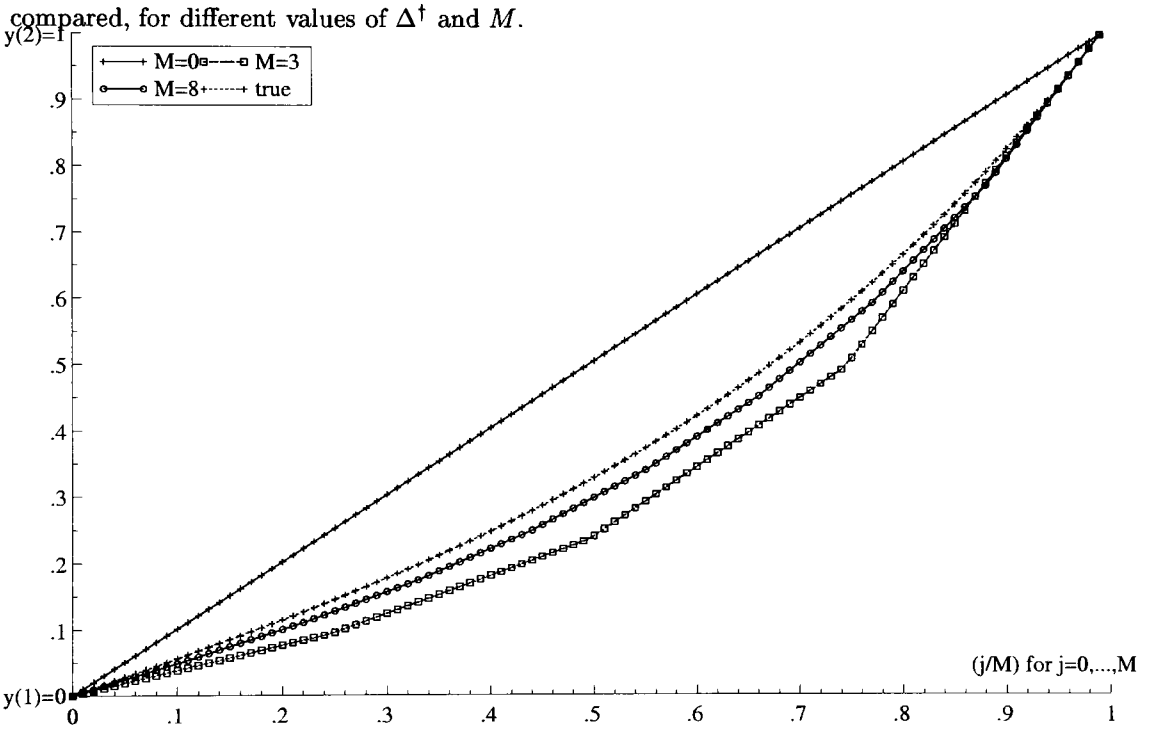


Figure 2.2: *The expected path between two fixed points, $E(y_1^*|y_1, y_2)$, and hence the discretisation bias is shown for the OU process using different values of M, (where $M = 0, 3, 10$ and 20), under the Euler scheme. The path under the true process is denoted by 'true'. The value of μ is taken to be -1 and $\Delta^\dagger = 2$ in all computations, with fixed points $y_1 = 0$ and $y_2 = y_{1,M+1} = 1$ and auxiliary variables $y_1^* = (y_{1,1}^*, \ldots, y_{1,M}^*)$.*

Intuitively, for a linear process with constant volatility, one can think of the path between the observations as a bridge. Consider the fixed observations, y_0 and y_{M+1} in the case of the OU process. We condition the process to start at y_0 at time 0 and arrive at y_{M+1} at time 1. We are therefore dealing with an OU process, tied down at each end inside a time interval $[0, 1]$, or alternatively, an OU Bridge. To illustrate the advantages of introducing auxiliary variables, consider the OU process for which $\theta = (\mu, \sigma^2)$, $a(y_t, t; \theta) = \mu y_t$ and $b^2(y_t, t; \theta) = \sigma^2$ in equation (2.1). The conditional distribution of $y_{t+1}|y_t; \theta$ under the Euler scheme is Normal with a mean of $\rho_E^\dagger y_t$ and a variance of $\sigma^2\Delta^\dagger$, where $\rho_E^\dagger = 1 + \mu\Delta^\dagger$. For the strong solution, also normal, the mean of $y_{t+1}|y_t$ is $\rho_S^\dagger y_t$ with variance $\frac{\sigma^2}{2\mu}(\rho_S^{2\dagger} - 1)$, where $\rho_S^\dagger = \exp(\mu\Delta^\dagger)$.

[5] For linear processes, it is possible to sample directly from the density, since the true conditional denisty is available in closed-form.

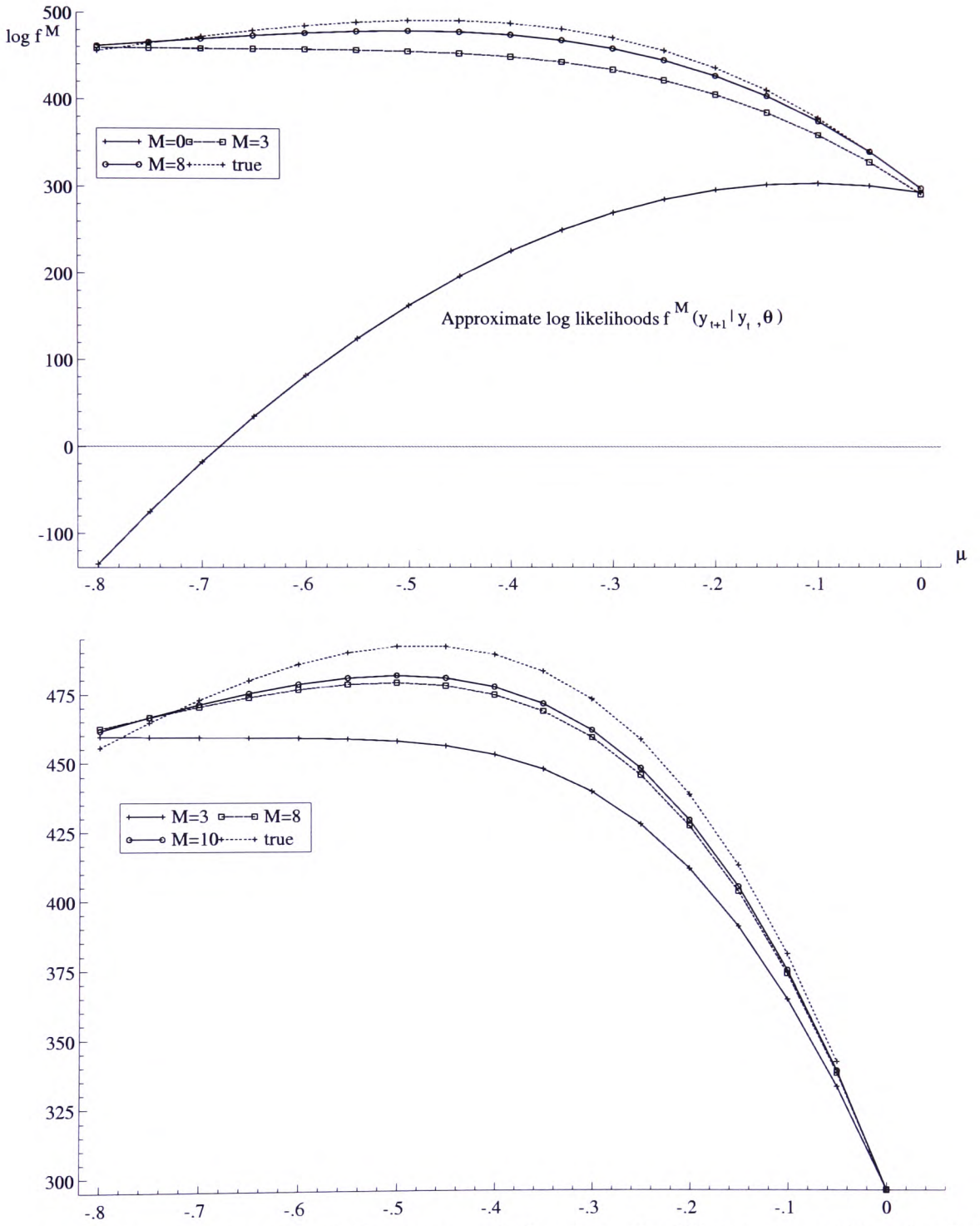


Figure 2.3: *Approximate log-likelihoods are graphed against μ for different values of M under the Euler scheme using 500 observed values. The number of latent points are $M = 0, 3$ and 8 (top graph) and $M = 3, 8$ and 10 (bottom graph). The strong solution is denoted by 'true'. Each approximation is evaluated for different values of $\mu \in [-0.8, 0]$. For the DGP, $\mu = -0.5$, $\sigma^2 = 0.01$ and $\Delta^\dagger = 4$.*

Suppose, we condition each y_t on its neighbouring points, y_{t-1} and y_{t+1}. Then the distribution of the resulting bridge process under the Euler scheme can be expressed as

$$y_t|y_{t-1}, y_{t+1}, \theta \sim N\left\{\frac{\rho_E^\dagger}{1+\rho_E^{2\dagger}}(y_{t-1}+y_{t+1}),\ \frac{\sigma^2\Delta^\dagger}{1+\rho_E^{2\dagger}}\right\}. \tag{2.14}$$

Using the strong solution of the process it is,

$$y_t|y_{t-1}, y_{t+1}, \theta \sim N\left[\frac{\rho_S^\dagger}{1+\rho_S^{2\dagger}}(y_{t-1}+y_{t+1}),\ \frac{\sigma^2\Delta^\dagger}{\left\{2\mu(\rho_S^{2\dagger}-1)/(1+\rho_S^{2\dagger})\right\}}\right]. \tag{2.15}$$

Similarly, if we consider a block of M latent points $y_t^* = (y_{t,1}^*, \ldots, y_{t,M}^*)$, then we can express $f(y_t^*|y_t, y_{t+1})$ as a Gaussian distribution, with expected value

$$\mathrm{E}(y_t^*|y_t, y_{t+1}; \theta) = \frac{1}{1+\rho^2+\ldots+\rho^{2M}} \begin{pmatrix} \rho\sum_{i=0}^{M-1}\rho^{2i}y_t + \rho^M y_{t+1} \\ \rho^2\sum_{i=0}^{M-2}\rho^{2i}y_t + \rho^{M-1}\sum_{i=0}^{1}\rho^{2i}y_{t+1} \\ \vdots \\ \rho^M y_t + \rho\sum_{i=0}^{M-1}\rho^{2i}y_{t+1} \end{pmatrix} \tag{2.16}$$

where ρ is $\rho_S = \exp(\mu\Delta)$ under the true process and $\rho_E = 1+\mu\Delta$ under the process using the Euler approximation.

Given the analytically tractable form of the conditional density for the OU process, we can therefore draw the expected path of the process between two fixed points using the strong solution of the process and illustrate the curvature bias in the paths for different discretisations under the Euler scheme. Figure 2.2 illustrates the curvature bias using $M = 0, 3$, and 8 latent points, $y_1^* = (y_{1,1}^*, \ldots, y_{1,M}^*)$ for the Euler approximation. The actual path, denoted by 'true', was computed using (2.15) and (2.16) with $\Delta^\dagger = 2$. Notice how $M = 0$ is the linear interpolation between the two fixed points, as there are no latent points, so that it completely misses the curvature of the strong solution. Expected paths for non-zero values of M produce a downwards bias, converging to the strong solution of the process, depicted by the second curve from the left. These clearly over-emphasise the curvature, though even a small M is a substantial improvement as compared to the case $M = 0$.

The Data for both graphs in Figure 2.3 is obtained using $\Delta^\dagger = 4$ with $T = 500$ observations. Parameter values used for the data generating process (DGP) are $\mu = -0.5$ and $\sigma^2 = 0.01$. Interval lengths of $M = 0, 3, 8$ and 10 are used and the log-likelihood obtained from the true process is denoted by 'true'. These are evaluated for different values of $\mu \in [-0.8, 0]$. The true OU log-likelihood of μ given by $\sum_{t=1}^{T-1}\log g(y_{t+1}|y_t;\theta)$ can thus be compared to the approximate conditional likelihood $\sum_{t=1}^{T-1}\log f^M(y_{t+1}|y_t,\theta)$ where

$$g(y_{t+1}|y_t,\theta) = \phi\left[y_{t+1}\ \middle|\ \exp(\mu\Delta^\dagger)y_t,\ \frac{\sigma^2}{2\mu}\left\{\exp(2\mu\Delta^\dagger)-1\right\}\right]$$

and

$$f^M(y_{t+1}|y_t,\theta) = \phi\left[y_{t+1}\ \middle|\ \left(1+\frac{\mu\Delta^\dagger}{M+1}\right)^{M+1}y_t,\ \sigma^2\left(\frac{\Delta^\dagger}{M+1}\right)\left\{\frac{1-\left(1+\frac{\mu\Delta^\dagger}{M+1}\right)^{2(M+1)}}{1-\left(1+\frac{\mu\Delta^\dagger}{M+1}\right)^2}\right\}\right],$$

where $\phi(y_{t+1}|m,v)$ denotes the Gaussian density of y_{t+1}, with mean m and variance v. The approximate log-likelihoods for the OU under the Euler scheme are shown in the top graph in Figure 2.3 for different values of M. The bottom graph shows a replicated version, with $M = 10$ also shown and $M = 0$ removed to depict the convergence to the strong solution log-likelihood as M increases. One can compare current methods which do not employ auxiliary variables to basing estimation with $M = 0$ which is clearly biased.

The auxiliary points do not add any information as they are not part of the observed data. Nevertheless, they are useful because they allow us to construct an approximate model with smaller bias. Augmentation therefore provides a better approximation to the conditional transition density and the degree of discretisation bias can be controlled through the choice of M.

2.5 Existing methods for estimating stochastic differential equations

Most methods which do not introduce additional auxiliary points have concentrated on approximating the continuous-time model with a discretised version. However, as shown above for the OU, using some approximate conditional transition density, $f(y_{t+1}|y_t;\theta)$ with discrete observations can result in biased estimators. The bias, a function of $\Delta^\dagger$ and T, tends to zero when the time interval between the observations tends to zero, which is not the case in practice.

General existing research can be separated into two categories; one involves finding a good approximation to the likelihood, while the other is based on alternative methods. The first further separates into methods that introduce auxiliary points and procedures that do not. Previous work, using a discretised version without augmentation is first presented in 2.5.1. Martingale methods based on estimating functions are reviewed in 2.5.2 and filtering methods are outlined in 2.5.3. Generalised method of moments estimators and efficient method of moments based estimators, in particular, the indirect inference method is covered in Section 2.5.4. A summary of simulated likelihood methods using auxiliary points is then outlined in 2.5.5. Non-parametric methods, which do not rely on a discretisation version of the SDE are the subject of Section 2.5.6. Methods partially based on finite difference equation techniques, such as the Crank-Nicholsen method are covered in 2.5.7.

2.5.1 Discretised likelihoods

When the transition densities are known, under certain regularity conditions, the log-likelihood

$$l_T(\theta) = \sum_{t=1}^{T-1} \log g(y_{t+1}|y_t;\theta)$$

can be evaluated and consistency of the estimate $\hat{\theta}_T \xrightarrow{\text{p}} \theta$ and asymptotic normality, as $T \to \infty$ can be established. In general, the transition density is not known and so an approximate log-likelihood $\hat{l}_T(\theta)$ is constructed. Florens-Zmirou (1989) showed that for fixed $\Delta_t^\dagger$, estimation based on $\hat{l}_T(\theta)$ is inconsistent and that the resulting parameter estimate is strongly biased, unless $\Delta_t^\dagger \to 0$ for all t.

Le Breton (1976) discretises the continuous-time likelihood function for the linear drift case when the diffusion coefficient does not depend on θ. Lo (1988) considers a parametric estimation problem for Itô processes using maximum likelihood with discretely sampled data. The method is computationally expensive, as it requires a partial differential equation to be solved numerically, for each maximum

likelihood iteration, unless the strong solution of the process is known. He concludes that general consistent estimators cannot be obtained, so that the use of discrete approximations, when $\Delta^{\dagger}$ is fixed, is not appropriate. Dacunha-Castelle and Florens-Zmirou (1986) analyse the loss of information in estimating parameters of the diffusion due to discretisation and emphasise the differences depending on whether the coefficient of the volatility component is known or unknown. Dohnal (1987) looks at parameter estimates obtained for the diffusion coefficient when the observations are made at discrete time points; only equi-distant sampling points are considered. Florens-Zmirou (1993) investigates estimation of the diffusion coefficient in non-parametric settings, based on discrete observation in a given finite time interval; consistency and asymptotic normality are considered. Florens-Zmirou (1989), Genon-Catalot and Jacod (1993) and Kessler (1995) all consider use of the Euler scheme to obtain a Gaussian transition density. Approximation based on discrete observations is shown to be biased and inconsistent, unless the time step between the observations, $\Delta^{\dagger}$, tends to zero in the limit.

An application is provided in a paper by Shoji and Ozaki (1997), where they compare the performance of five estimation methods for continuous-time stochastic processes from discrete observations. These include the Euler approximation, two local linearization methods, a likelihood-ratio maximum likelihood method and generalised methods of moments, (GMM). The local linearization method involves using an approximation of the drift function, the motivation being that the local behaviour of the function is primarily of interest. A combination of Itô's lemma, local constancy of the derivatives of the drift function and a linear approximation is used to obtain a discretised Gaussian process. Tests are performed for a model with linear drift and coefficients that depend on the states and for models with non-linear drift. They conclude that when the discrete-time interval is small, the difference between the methods, for the linear drift case, (except for the GMM method, which is inferior in performance compared to the other methods), is not large. For the non-linear case, GMM performance is reasonable, while the other methods have large bias. When differences are large however, the local linearization method outperforms the others. Biscay, Jimenez, Riera, and Valdes (1996) also investigate local linearization methods applied to numerical solutions of stochastic differential equations.

The method proposed by Roberts and Stramer (1999) uses Girsanov's theorem to write down an expression of the likelihood of the trajectory, which is approximated using a discretisation of the process. Augmentation is used to reduce the discretisation bias, by simulating the missing paths between the observed points. MCMC methods are used, updating the auxiliary states in one block. The candidate proposed, in the context of a M-H independence sampler, is either based on a Brownian bridge or a linear Brownian bridge. The marginal likelihood for the data can be expressed in terms of the likelihood for the parameters (through the R-N derivative, outlined in Section 2.1.4) and Brownian bridge functionals, an idea first proposed by Dacunha-Castelle and Florens-Zmirou (1986) and extended by Yoshida (1992) and Florens-Zmirou (1989). However, this closed-form expression only seems applicable to cases where the volatility coefficient is multiplicative, so that it can be separated from terms depending on the states of the process, unless the transformation that removes dependency of the volatility coefficient in $b(y_t, \tau_t; \theta)dW_t$, is known. The use of Itô's lemma in this context is equivalent to integrating out the volatility coefficient - in the case of the OU process, a multivariate$-t$ distribution is obtained for the drift parameters. The method seems difficult to generalise for processes outside the class of volatilities with a multiplicative constant coefficient structure. Reducing the SDE additionally complicates the drift term for highly non-linear processes.

An interesting point which is raised, concerns the difficulty with the performance of the M-H sampler, when a large number of latent states are imputed. In the O-U case, the quadratic variation formula is used to estimate the volatility coefficient, σ^2, which is known almost surely. As the number of latent

points go to infinity, the posterior distribution of σ^2 converges to a point mass, since it relies almost solely on the latent neighboring states and not on the information in the data, provided by the fixed endpoints. The fact that the diffusion coefficient has point mass on the previous draw is important in the limit, however, most applications use around 20 imputed points between each pair of observations for highly non-linear processes. In addition, there are additional issues arising, (see for example Section 6.3), when all latent states are updated in one block. As M becomes large, discretisation bias decreases, but the MCMC problem becomes harder and tailoring the proposal becomes very important. Results in Section 6.3 suggest a decline in the mixing for large M, as essentially the sampler stops moving around the support of the target density.

2.5.2 Estimating functions

An alternative approach to estimating discretely observed diffusions utilises martingale methods, through estimating functions. This approach has been developed by authors such as Bibby and Sørensen (1995), Kessler and Sørensen (1999) and Sørensen (1997). In Sørensen (1997), the likelihood function of interest is given by $L_T(\theta) = \prod_{t=1}^{T-1} g(y_{t+1}|y_t;\theta)$ obtained from (2.1). The idea is to replace $g(y_{t+1}|y_t;\theta)$ by some approximation, $f(y_{t+1}|y_t;\theta)$ using the Euler scheme. If the time step $\Delta^\dagger$ is small, then the Euler is used to exploit the normal conditional form. Differentiating the log-likelihood with respect to θ yields an approximate Gaussian score function

$$\begin{aligned} H_T(\theta) &= \sum_{t=1}^{T-1} \left[\frac{\partial_\theta a(y_t,t;\theta)}{b^2(y_t,t;\theta)} \{y_{t+1} - y_t - a(y_t,t,\theta)\Delta^\dagger\} \right] \\ &\quad + \frac{\partial_\theta b^2(y_t,t,\theta)}{2b^4(y_t,t;\theta)\Delta^\dagger} \left[\{y_{t+1} - y_t - a(y_t,t;\theta)\Delta^\dagger\}^2 - b^2(y_t,t,\theta)\Delta^\dagger \right] \end{aligned} \tag{2.17}$$

where $\partial_\theta a = \partial a/\partial\theta$ and $\partial_\theta b^2 = \partial b^2/\partial\theta$. Estimates are obtained using (2.17). Consistent and asymptotically normal estimates are obtained for $\Delta^\dagger \to 0$ as $T \to \infty$. For studies where $\Delta^\dagger$ is small, see Kloeden and Platen (1992). Simulation studies by Pedersen (1995a) and Bibby and Sørensen (1995) show that the estimators are strongly biased if $\Delta^\dagger$ is not small.

The estimating function given in (2.17) can be biased and expressions for the magnitude of the bias can be found in Florens-Zmirou (1989) and Kessler (1995). The bias depends on the order $T\Delta^\dagger$ when the diffusion coefficient depends on θ. The Euler approximation implies a normal conditional transition density, while the true conditioal distribution is in fact a Gamma distribution. Martingale estimating functions have been proposed to deal with the inherent bias in the estimating functions defined in (2.17). The Euler likelihood is essentially transformed by subtracting the compensator from the discretised likelihood to give a martingale estimating function and so removes the condition that $\Delta^\dagger \to 0$. Kessler (1996) considers expressions that do not have the martingale form. Quadratic estimating functions have been considered in Bibby (1994) and Bibby and Sørensen (1997). Pedersen (1994a) and Kessler (1995) consider higher order polynomial terms. Estimating functions based on eigenfunctions were considered by Kessler and Sørensen (1999).

2.5.3 Filtering methods

The state space form (SSF) representation, described by a *measurement* and *transition* equation, has the advantage that the measurement equation is a Markov process. The SSF is also flexible, for example it

can easily handle missing observations. For a review of the SSF representation, see Harvey, (1981, Ch. 5). Filtering methods to estimate SDE's have been proposed by Baadsgaard, Nielsen, Spliid, Madsen, and Preisel (1997), Nielsen 1998, Nielsen, Vestergaard, and Madsen (1998) and Nielsen, Madsen, Melgaard, and Baadsgaard (1998). A review of filtering and prediction theory in the context of diffusions is provided in Madsen, Nielsen, and Baadsgaard, (1998, pp. 225-237).

Typically, the first and second moments of the posterior are only considered to obtain conditional moment estimators (using for example GMM) – this in fact minimises the Bayesian risk, see Madsen, Nielsen, and Baadsgaard, (1998, p. 230). For linear SDE's, the Kalman Filter (KF) can be used. Approximate filters for non-linear SDE's include the truncated second order filter, a linearised KF, the extended KF and a statistically linearised filter. When the volatility function depends on the states, a transformation of the SDE to reduce to an SDE where the volatility term is independent of y_t is required for the methods to work. For an application where the volatility function depends on the states, see Baadsgaard, Nielsen, Spliid, Madsen, and Preisel (1997) and in the context of the stochastic volatility model, see Nielsen (1998) and Nielsen, Vestergaard, and Madsen (1998).

2.5.4 Efficient method of moments based estimators and indirect inference

Focus in the literature has been on developing simulation based method of moments estimators for estimating non-linear diffusions. The *Indirect Inference* method, (see Smith 1993 and Gourieroux et al. 1993) has been recast into the influential Efficient Method of Moments (EMM) methodology by papers such as Gallant and Tauchen (1996), Gallant and Long (1997) and Tauchen (1997).[6] A paper by Gallant and Davidien (1993) uses the methodology of EMM, also referred to as Minimum chi-squared estimation and GMM. Additional literature on this subject has been provided by Hansen (1982) on the theory of GMM.

The general outline of the method involves finding an auxiliary model which produces simulated data exhibiting the same patterns and properties as those in the observed data. It is often advisable to base the choice of the auxiliary model on non-parametric considerations, see Gallant and Tauchen (1997b). Minimum chi-squared methods are shown in Tauchen (1997), to systematically provide the moment restrictions needed to obtain an auxiliary model. In practice, the aim is to estimate θ and specify the drift and volatility functions in (2.1) and test the adequacy of various models applied to a particular dataset. The Euler scheme is used to discretise the SDE, which in turn characterises the joint and conditional distributions of the state and volatility process. However, these do not lead to a closed-form expression for the likelihood, so that simulation methods must be employed. A good review of the method and its assumptions, in the context of the stochastic volatility model, is given in Engle and Lee, (1996, pp. 336-339).

The technique developed in the thesis does not require stationarity or smoothness assumptions and easily allows multivariate generalisations. The method adopt a similar basis of approach except that it is based on likelihood inference, while the aforementioned authors propose moment based estimators. One can view our method as an *amelioration* approach. An additional disadvantage with the EMM methodology is that it is difficult to gauge sensitivity of the method to the choice of M.

[6]The EMM based method can be thought of as a special form of indirect inference.

2.5.5 Simulated likelihood and likelihood-ratio methods

The next two approaches use simulation of the augmented variables to compute a smooth function of the parameters. Simulation methods will be reviewed in Section 3. The function is then maximised over the parameters to obtain a maximum likelihood or contrast type estimator. One method is based on the idea of Pedersen, (1995b and 1995a) which uses simulated likelihoods. An alternative approach is to use the method proposed by Geyer (1999) with auxiliary points. Both methods are outlined below; a detailed exposition is given in Sections 8.3 and 8.4, when we compare alternative methods on obtaining closed-form expressions for the likelihood.

The simulated likelihood method, as in the case of Pedersen, (1995b, 1995a) can be used to obtain an estimate of θ. The method is based on treating the values on a trajectory connecting two discrete observations as missing points. An approximation of the log-likelihood is then obtained using the Euler scheme. By definition, the approximate likelihood can be expressed as

$$f^M(y_{t+1}|y_t,\theta) = \int f(y_{t+1}|y_t^*,y_t,\theta) f(y_t^*|y_t,\theta) dy_t^*.$$

Pedersen suggests that this transition density be estimated using standard Monte Carlo integration. Each $y_{t,j}^*$ is drawn from $f(y_{t,j}^*|y_{t,j-1}^*;\theta)$, for $j = 1,\ldots,M$, starting with $y_{t,1}^*$ and simulating forwards. The resulting value $y_{t,M}^*$ is thus drawn from $f(y_{t,M}^*|y_t;\theta)$ and the procedure is repeated over N iterations. An estimate of $f^M(y_{t+1}|y_t;\theta)$ is then given by

$$\hat{f}^M(y_{t+1}|y_t;\theta) = \frac{1}{N}\sum_{k=1}^{N} f\left\{y_{t+1}|y_{t,M}^{(k)*};\theta\right\}$$

where $k = 1,\ldots,N$, so that an estimate of the log-likelihood is based on

$$\hat{l}_T^{(M)}(\theta) = \sum_{t=1}^{T} \log \hat{f}^M(y_{t+1}|y_t;\theta). \tag{2.18}$$

If $f^M(y_{t+1}|y_t;\theta)$, under suitable conditions, converges pointwise to $g(y_{t+1}|y_t;\theta)$ as $M \to \infty$, then the discretised log-likelihood, $l_T^M(\theta) \xrightarrow{\text{p}} l_T(\theta)$ for all $\theta \in \Theta$ and T. As the transition density is not known in general, proofs are based on generalised expectations and functionals of the Wiener process from Malliavin calculus. Though the density is initially estimated unbiasedly, taking a non-linear transformation, such as the log results in a biased, though consistent log-likelihood.

Alternatively, an importance sampler can be used to estimate $f^M(y_{t+1}|y_t;\theta)$, such as,

$$\frac{1}{J}\sum_{j=1}^{J} \frac{f\left\{y_{t+1},y_t^{*(j)}|y_t;\theta\right\}}{f(y_t^{*(j)}|y_t;\theta)} = \frac{1}{J}\sum_{j=1}^{J} f\left\{y_{t+1}|y_{t,M}^{*(j)};\theta\right\} \tag{2.19}$$

where $y_t^{*(j)}$ is drawn by iterating one-step ahead, the discretised version of (2.1), given in (2.11). For further discussion of importance sampling methods, see Kloek and van Dijk (1978), Ripley, (1987, pp. 122-3) and Geweke (1989). An alternative importance sampling function is $f(y_t^*|y_t,y_{t+1};\theta)$, which then minimises the Monte Carlo variance of (2.19), since then

$$\frac{f(y_{t+1},y_t^*|y_t;\theta)}{f(y_t^*|y_t,y_{t+1};\theta)} = f^M(y_{t+1}|y_t;\theta).$$

This means that $f^M(y_{t+1}|y_t,\theta)$ can be estimated exactly using one simulated value of y_t^*. However, this approach is generally not feasible, as the importance sampling function $f(y_t^*|y_t,y_{t+1};\theta)$ is rarely available in closed form and so instead, we have to search for an approximation to $f(y_t^*|y_t,y_{t+1};\theta)$.

The estimate in (2.18), although easily available, can be poorly behaved if M is large.Essentially, since the information contained in y_{t+1} is not being used, as M increases, the simulated paths can wander further and further away from the fixed end point, (a graphical explanation is given in Figures 8.1–8.3). Hence, the likelihoods (which are based on the simulated paths) resulting from paths that terminate close to the fixed endpoint form the bulk of the contribution to the estimate of the Monte Carlo integral. This means that as M increases, though the Monte Carlo bias decreases, the Monte Carlo variance can increase substantially. This highlights the importance of conditioning on both fixed endpoints and motivates our approach.

An alternative approach to is to consider a simulated likelihood ratio (SLR) method based on the ideas of Geyer (1999). The method has also been considered by authors such as Billio, Monfort, and Robert (1998) in the context of diffusions, in the econometrics literature. The method is based on estimating the likelihood-ratio statistic, based on a fixed parameter θ_0,[7]

$$\frac{f^M(y_{t+1}|y_t;\theta)}{f^M(y_{t+1}|y_t;\theta_0)} = \int \frac{f(y_{t+1},y_t^*|y_t;\theta)}{f(y_{t+1},y_t^*|y_t;\theta_0)} f(y_t^*|y_t,y_{t+1};\theta_0)dy_t^*$$

by

$$\frac{1}{N}\sum_{j=1}^{N} \frac{f(y_{t+1},y_t^{*(j)}|y_t;\theta)}{f(y_{t+1},y_t^{*(j)}|y_t;\theta_0)}, \quad \text{where} \quad y_t^{*(j)} \sim f(y_t^*|y_t,y_{t+1};\theta_0).$$

The sequence of simulations $\left\{y_t^{*(1)},...,y_t^{*(N)}\right\}$ represents the MCMC output, after burn-in, with equilibrium distribution $f(y_t^*|y_t,y_{t+1};\theta_0)$.

Though an efficient estimate of the likelihood function for θ in a neighbourhood of θ_0 can be obtained, evaluating the statistic is computationally intensive and sample iterates may be highly correlated. Computations become increasingly less reliable for values of θ some distance from θ_0, (since values for $y_t^{*(j)}$ which are likely under a θ far from θ_0, are unlikely to be sampled under θ_0). The method is also sensitive to initial values. Billio, Monfort, and Robert (1998) consider maximising the likelihood ratio for a range of models. They state that provided N is large, the SLR method can be considered as asymptotically efficient unlike the Method of Simulated Moments (MSM), the Simulated Pseudo Likelihood Method (SPLM) and the Indirect Inference method (II). It also only requires one optimisation (provided θ_0 is close to the exact MLE) which is not the case with Simulated EM methods.

2.5.6 Non-parametric methods

Aït-Sahalia (1996a) develops non-parametric methods based on the continuous-time formulation of (2.1), when only discrete observations are available. An application to the spot rate is considered in Aït-Sahalia (1996b). Methods are developed to obtain a non-parametric estimate of the volatility function, through *density matching*. As in the indirect inference case, the aim is to derive drift and volatility functions

[7]It can also be applied to situations where two models are compared, (possibly non-nested), say $\mathcal{M}_1$ and $\mathcal{M}_2$ which we assume have no unknown parameters.

that are consistent with the observed distribution of the data. The main advantage with this approach, is that the method does not rely on the sampling interval, $\Delta^{\dagger}$, going to zero.

The data is used to construct a non-parametric estimate of the densities. Non-parametric estimates of the drift and volatility functions are then constructed by matching up the densities. This is the reverse procedure of the parametric based methods which first specify a drift and volatility, which specify the marginal and transition densities.

2.5.7 Methods based on the Fokker-Planck equation and Crank-Nicholsen method

The Crank-Nicholson method is a second order numerical technique of implicit finite difference type for partial differential equations, see Mitchell and Griffiths (1980) and Strikwerda (1990) for theory. For applications in finance, see Wilmott, Dewynne, and Howison (1993) and Duffie (1996). A brief discussion of finite-difference methods was given in Section 2.1.5.

Poulsen (1999) constructs an approximation to the log-likelihood using the Fokker-Planck partial differential equation and numerically obtains an estimate of θ using the Crank-Nicholson method. In particular, the assumption is that the log-likelihood can be written as

$$\begin{aligned}
\hat{l}_T(\theta) &= \sum_{t=1}^{T-1} \log f(y_{t+1}|y_t;\theta) \\
&= \sum_{t=1}^{T-1} \left\{ \log g(y_{t+1}|y_t;\theta) + h^2 a(y_t,\tau_t;\theta) + o(h^2) b(y_t,\tau_t;\theta) \right\} \\
&= l_T(\theta) + \sum_{t=1}^{T-1} \left\{ h^2 a(y_t,\tau_t;\theta) + o(h^2) b(y_t,\tau_t;\theta) \right\},
\end{aligned}$$

where h is the discretisation step-size. The estimator, denoted AML, is shown to be asymptotically equivalent to the maximum likelihood estimator, under certain regularity conditions.No simulation is involved, as the analytical approximation involves finding the numerical estimate of the drift and volatility functions.

Simulated data is used to compare the method to an Euler approximation based on $M = 0$ and the GMM method, applied to the Hull-White model, given in Section 2.2. The model considered is actually given by,

$$dy_t = \kappa(\mu - y_t)dt + \sigma y_t^{\gamma} dW_t$$

which corresponds to $\alpha = \kappa\mu$, $\beta = \kappa$ in the version presented in 2.2. Estimates of μ are similar for all three methods, but differences can be seen in the estimate of κ. The Euler and AML appear to have similar behaviour, with the latter having lower standard errors. The main disadvantage with the method is that it can only be applied to one-dimensional SDE's.

2.6 Conclusion

This chapter has outlined some of the concepts used in the thesis relating to stochastic differential equations. Issues surrounding discretisation of the continuous-time process and augmentation have been covered, so that to address the issue of sampling the auxiliary states y_t^*, and the parameter of interest

θ, we need to define concepts of simulation estimation. This is addressed in Chapter 3. The methods will allow us to obtain sampled values from the posterior distribution of interest.

2.7 Appendix

2.7.1 Assumptions under the Euler and Milstein schemes

We will assume the following conditions in Section 2.3, (and additionally in Section 9.2.1), under the Euler scheme, using $y(0)$ under the itô process and y_0 under the approximation scheme, with maximum step size $\delta = \min_{t=1}^{T} \Delta_t^{\dagger}$, as starting points,

$$\begin{aligned}
\mathrm{E}\left\{|y(0)|^2\right\} &< \infty && \text{(A.2.1)}\\
\mathrm{E}\left\{|y(0)-y_0|^2\right\}^{(1/2)} &\leq K_1\delta^{(1/2)} && \text{(A.2.2)}\\
|a(y,t)-a(x,t)|+|b(y,t)-b(x,t)| &\leq K_2|y-x| && \text{(A.2.3)}\\
|a(y,t)|+|b(y,t)| &\leq K_3(1+|y|) && \text{(A.2.4)}\\
|a(y,s)-a(y,s)|+|b(y,s)-b(x,s)| &\leq K_4(|1+|y|))|s-t|^{(1/2)} && \text{(A.2.5)}
\end{aligned}$$

where the constants $K_1, \ldots, K_4$ do not depend on δ, $s,t \in [0,T]$ and $y,x \in \mathbb{R}^d$ then for the Euler approximation y_T, we obtain

$$\mathrm{E}\{|y(T)-y_T|\} \leq K_5\delta^{(1/2)}$$

where the constant K_5 does not depend on δ. Given conditions A.2.4 and A.2.5, the result follows from applying Doob's inequality and the Gronwall inequality. A proof of this assertion can be found in Kloeden and Platen (Section 10.2, 1992). For α, $\beta : [t_0,T] \to \mathbb{R}$ integrable, with

$$0 \leq \alpha(t) \leq \beta(t) + C\int_{t_0}^{t} \alpha(s)\,ds$$

for $t \in [t_0,T]$, where $C > 0$, the Gronwall inequality is given by,

$$\alpha(t) \leq \beta(t) + C\int_{t_0}^{t} \exp\{C(t-s)\}\beta(s)\,ds.$$

See Kloeden and Platen, (1992, p. 129).

Under the Milstein scheme we make the following additional assumptions:

$$
\begin{aligned}
|a^*(y,t) - a^*(x,t)| &\leq K_2|y-x| && \text{(A.2.6)}\\
|b(y,t) - b(x,t)| &\leq K_2|y-x| && \text{(A.2.7)}\\
|Lb(y,t) - Lb(x,t)| &\leq K_2|y-x| && \text{(A.2.8)}\\
|a^*(y,t)| + |La(y,t)| &\leq K_3(1+|y|) && \text{(A.2.9)}\\
|b(y,t)| + |Lb(y,t)| &\leq K_3(1+|y|) && \text{(A.2.10)}\\
|L^2 b(y,t)| &\leq K_3(1+|y|) && \text{(A.2.11)}\\
|a^*(y,s) - a^*(y,t)| &\leq K_4(1+|y|)|s-t|^{(1/2)} && \text{(A.2.12)}\\
|b(y,s) - b(y,t)| &\leq K_4(1+|y|)|s-t|^{(1/2)} && \text{(A.2.13)}\\
|Lb(y,s) - Lb(y,t)| &\leq K_4(1+|y|)|s-t|^{(1/2)} && \text{(A.2.14)}
\end{aligned}
$$

for all $s,t \in [0,T]$, $y,x \in \mathbb{R}^d$ where

$$
\begin{aligned}
a^* &= a - \frac{1}{2}bb'\\
L &= b\frac{\partial}{\partial y},
\end{aligned}
$$

and the constants $K_2,\ldots,K_4$ do not depend on δ.

2.7.2 Strong solution of OU process

To obtain the strong solution of the linear OU process:

$$dy(t) = \mu y(t)dt + \sigma dW(t)$$

Suppose we multiply both sides by $\exp(-\mu t)$, and look at the differential:

$$
\begin{aligned}
d\{\exp(-\mu t)y(t)\} &= -\mu\exp(-\mu t)y(t)\,dt + \exp(-\mu t)dy(t)\\
&= \exp(-\mu t)\sigma\,dW(t).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\exp(-\mu\tau_{t+1})y(\tau_{t+1}) &= \exp(-\mu\tau_t)y(\tau_t) + \sigma\int_{\tau_t}^{\tau_{t+1}}\exp(-\mu s)\,dW(s)\\
y(\tau_{t+1}) &= \exp\{\mu(\tau_{t+1}-\tau_t)\}y(\tau_t) + \sigma\int_{\tau_t}^{\tau_{t+1}}\exp\{-\mu(s-\tau_{t+1})\}\,dW(s)\\
&= \exp(\mu\Delta_t)y(\tau_t) + \sigma\int_{\tau_t}^{\tau_{t+1}}\exp\{-\mu(s-\tau_{t+1})\}\,dW(s).
\end{aligned}
$$

So,

$$\begin{aligned} E\{y(\tau_{t+1})|y(\tau_t)\} &= \exp(\mu\Delta_t)y(\tau_t) \\ \text{var}\{y(\tau_{t+1})|y(\tau_t)\} &= \frac{\sigma^2}{2\mu}\{\exp(2\mu\Delta_t) - 1\} \\ y(\tau_{t+1})|y(\tau_t) &\sim N[\exp(\mu\Delta_t)y(\tau_t), \frac{\sigma^2}{2\mu}\{\exp(2\mu\Delta_t) - 1\}]. \end{aligned}$$

CHAPTER 3

BASIC SIMULATION ESTIMATION

Efficient numerical integration relies on simulation techniques, which use pseudo random number generators; pseudo in the sense that the sequence generated should retain certain properties of random numbers, even though the computer generates these numbers deterministically. Ideally, the numbers should have the distribution properties associated with independent realisations. Simulation methods are used in both the Bayesian and the classical inference literature. Papers by Kloek and van Dijk (1978), Zellner and Rossi (1984), Geweke (1989) and Gelfand and Smith (1990) discuss issues of Bayesian methodology. The methods rely on obtaining an estimate of some marginal density, by numerically integrating a posterior density. Simulation methods in the classical inference literature focuses on simulating a criterion function to obtain an estimate of the quantity of interest. Examples include McFadden (1989) and Laroque and Salanie (1989). These are discussed in detail in Gourieroux and Monfort (1996) and briefly in Monfort and van Dijk (1995).

The outline of the Chapter is as follows: Section 3.1 introduces the basic concepts used in simulation methodology. In the context of the SDE defined in (2.1), given a dataset, the aim will be to obtain the paths of the missing or augmented data, Y^*, between the observed points, Y and estimate the parameters of interest, by simulating from the posterior density $f(Y^*; \theta|Y)$ using Bayesian methods. For many of these models, the likelihood is intractable, and we will use Markov chain Monte Carlo methods. Examples include the Gibbs sampler and the Metropolis–Hastings algorithm. The combination of a non-informative prior with the likelihood gives a posterior that can be analysed by simulation methods. By constructing a Markov chain with an invariant distribution, sample path averages can then be used to estimate the properties of that distribution. The effectiveness of the method depends on the characteristics of the Markov chain. We focus on the variability in the posterior mean and the accuracy of the estimation procedure. Simulation consistency of these estimates is established using laws of large numbers for sample path averages and central limit theorems applicable to Markov chains on continuous state spaces. This leads to full likelihood inference for the model, even though the likelihood is not available. Sufficient and necessary conditions to guarantee the convergence of the Markov chain to the invariant distribution, are presented in 3.2. Sampling variates from a multivariate distribution and assessing the MCMC output through convergence diagnostics is discussed in Section 3.3. Residuals can be used to assess the model fit, an important technique in assessing whether a particular model fits the data. These can also be used in order to assess the performance of our method. Diagnostics based on the residuals are outlined in Section 3.4. A method allowing comparison of non-nested models is presented in Section 3.5 and the Chapter concludes with Section 3.6.

3.1 INTRODUCTION

This thesis adopts the Bayesian approach, as the methods are often analytically easier in the context of these diffusions. The Bayesian approach to conducting inference on a quantity of interest, θ, is to construct a posterior distribution, given an information set. The information set could include data containing information about θ in the context of estimation, or the arrival of new information, if θ is to be updated. As an example, suppose we are interested in estimating a parameter θ, having access to some information Z. We assume, under certain regularity conditions on θ and a density f, that $\theta f(\theta|Z)$ is integrable. This ensures the existence of certain integrals that may be of interest to compute. Let $p(\cdot)$ denote the prior density. The posterior $f(\theta|Z)$ can be expressed in terms of the prior, $p(\theta)$ and the likelihood function, $f(Z|\theta)$, using *Bayes* theorem, as

$$f(\theta|Z) = \frac{f(Z|\theta)\,p(\theta)}{p(Z)} \propto \underbrace{f(Z|\theta)}_{\textit{likelihood}}\ \underbrace{p(\theta)}_{\text{prior}}.$$

We will often be interested in the posterior mean, that is the mean of $f(\theta|Z)$. In the case where θ is assumed to have a continuous distribution on the real line, we can define

$$\mathrm{E}(\theta|Z) = \int_{-\infty}^{\infty} \theta f(\theta|Z)\,d\theta = \int_a^b \frac{\theta f(\theta|Z)\,d\theta}{\int_a^b f(\theta|Z)\,d\theta} \tag{3.1}$$

for $f(\theta|Z)$ defined on $a \leq \theta \leq b$. If there is little prior information about θ, so that a flat prior is chosen for θ, the posterior kernel $f(\theta|Z)$ is determined completely by the likelihood function. If the support of θ is assumed to be compact and $f(\theta|Z) > 0$ for all permissible values of θ, then an estimate of the integrals in (3.1) can be computed using numerical integration or Monte Carlo integration methods. We use the fact that the integral can be interpreted as the expectation of a random variable, so that for a function h, $\int h(\theta)f(\theta|Z)\,d\theta$ can be approximated by $(1/N)\sum_{i=1}^{N} h\{\theta^{(i)}\}$. The estimate is unbiased and provided the variance of θ is finite, the estimate has a variance proportional to $1/N$. Though $f(\theta|Z)$ is proportional to $p(\theta)\,f(Z|\theta)$,the integrating factor may not be known. However, it is still possible to obtain an estimate of the integrals in equation (3.1) by using importance sampling methods or Markov chain Monte Carlo methods. Simulation precision can be made arbitrarily accurate and assessed using standard statistical techniques.

3.2 MARKOV CHAIN MONTE CARLO METHODS

Suppose that we wish to generate random variates X from a distribution with the (multivariate) density $f(X)$. Often it is difficult to sample from $f(X)$ directly. A generally applicable approach is to define a Markov chain which involves operations which are easier to compute, and has $f(X)$ as its stationary distribution. We can then generate observations from this chain, iterating steps where we move from the most recent value $X^{(j)}$ to an updated value $X^{(j+1)}$. Once the chain has converged, the sampled values $\{X^{(j)}\}$ can be treated as a dependent sample from $f(X)$. It turns out that there are many practical and computationally efficient ways of defining a Markov chain with the required stationary distribution. Simulation methods based on this idea are known as *Markov chain Monte Carlo* (MCMC) methods.

If the full conditional distributions, but not $f(X)$ itself, can be sampled from directly, the Gibbs sampler provides an effective technique for updating the iterates. In general, this is not always possible and the

M-H algorithm is used instead. The candidate generating density is used to propose the next values in the chain; these are accepted according to a probability, which is based on the properties of the Markov chain. MCMC methods are widely used in modern statistics. For a fuller description of the methodology and its applications, see Gilks, Richardson, and Spiegelhalter (1996). Here we will consider one popular MCMC method, the *Metropolis–Hastings* (M-H) algorithm. The technique was initially proposed by Metropolis, Rosenbluth, Rosenbluth, Teller, and Teller (1953) and later generalised by Hastings (1970) and Peskun (1973). It is a flexible algorithm, that does not require us to know the normalising constant, for producing variates from a given density, see also Kalos and Whitlock (1986). We will apply it to the estimation problem described in Section 2.3. Practical issues for MCMC methods are reviewed in Geyer (1992) and Gelman and Rubin (1992a).

3.2.1 Definitions and notation

Suppose $X = (X_1, \ldots, X_n)' \in \mathbb{R}^n$, where the components X_i, $i = 1, \ldots, n$, may be scalars or vectors. The *posterior* or *target* f is defined on a point X, where $X \in A \subset \mathbb{R}^n$ and is assumed to have a density $f(X)$. The subsets $A \subset \mathbb{R}^n$ are assumed to be a countably generated finite $\sigma-$algebra. The distribution of the (homogeneous) Markov chain is defined by the transition kernel

$$P\{X^{(j+1)} \leq y | X^{(j)} = x\}$$

and an initial distribution P_x. For $y \in A$, we can also obtain the conditional density as

$$p(x,y) = p\{X^{(j+1)} = y | X^{(j)} = x\} = \frac{\partial P\{X^{(j+1)} \leq y | X^{(j)} = x\}}{\partial y}$$

The *marginal posterior* density of X_i is defined as

$$f(X_i) = \int f(X_1, \ldots, X_n)\, dX_{\setminus i}$$

where $X_{\setminus i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. The full conditional density of X_i is represented by $X_i | X_{\setminus i}$ and is denoted $f(X_i | X_{\setminus i})$. We will often use the property of *conjugacy* in obtaining the distribution of the posterior density. If, for a given model, the prior and posterior distributions belong to the same class of distributions, (for example the normal), then the resulting inferential procedure is simplified. This is especially useful for updating, when new information on X arrives. Formally, a family of distributions P is conjugate to an observational model H if for every prior $p \in P$ and model $h \in H$, the posterior $f \in P$. Examples can be found in Section 6.1. Results on the theory of general state Markov chains are used to derive the properties of the algorithms. A brief and informal summary is given below; formal treatments can be found in see Revuz (1975), Tierney, (1994, 1995) and Nummelin (1984).

A Markov chain is said to have an *invariant* or stationary distribution f if

$$f(y) = \int_{\mathbb{R}^n} f(x) p(x,y)\, dx.$$

For a Markov chain with invariant distribution f and transition kernel P, for any function h, such that

$\mathrm{E}_f\{h(X)\} < \infty$, where E_f denotes the expectation under f,

$$\frac{1}{N}\sum_{j=1}^{N} h\left\{X^{(j)}\right\} \xrightarrow{p} \mathrm{E}_f\{h(X)\} \quad \text{for any initial } P_x.$$

This assumes that the chain is *irreducible*, that is, it must be able to reach all parts of the state space on which f is defined. A sufficient condition to ensure irreducibility is that $f(X)$ is positive for all X, see Tierney (1994). A *recurrent* chain has the property that all sets in the state space will be reached infinitely often. This allows us to make statements about the convergence of the averages of probabilities and implies a strong law of large numbers. More formally,

$$P_x\{X^{(j+1)} \in A \text{ infinitely often }\} > 0 \quad \text{for all } x \in A \subset \mathbb{R}^n \tag{3.2}$$

$$P_x\{X^{(j+1)} \in A \text{ infinitely often }\} = 1 \quad \text{for almost all } x \in A \subset \mathbb{R}^n \tag{3.3}$$

If a proper invariant distribution exists, then the chain is said to be *recurrent*. *Harris recurrence* is a stronger condition that requires (3.2) and (3.3) to hold for *all* $x \in A \subset \mathbb{R}^n$. An irreducible chain with invariant distribution f is *strongly aperiodic* if there exists a probability distribution on $A \subset \mathbb{R}^n$, denoted by ν and a constant $\beta > 0$ and a set B, such that $\nu(B) > 0$ and $P(x, A) \geq \beta\nu(A)$ for all $x \in B$ and $A \subset \mathbb{R}^n$, see Tierney (1995). Strong aperiodicity implies aperiodicity. If the chain has a proper invariant distribution f, (which is irreducible, recurrent and aperiodic), then f is the *unique invariant distribution* and the equilibrium distribution of the chain. For definitions on irreducibility, recurrence, aperiodicity and convergence of the chain, see Tierney (1995), ((1995), pp. 62-64). Most methods that are used to construct suitable Markov chains, use time-reversibility, that is, the chains have reversible versions. A Markov chain is said to be *reversible* if it is positive recurrent with stationary distribution f and $f(x)p(x,y) = f(y)p(y,x)$, where the latter is known as the *detailed balance equation*. Replacing the probabilities with densities extends the result to general state space Markov chains, see Tierney (1995). In summary,

- recurrence ensures that the law of large numbers holds;
- irreducibility, Harris recurrence and aperiodicity of the chain is equivalent to the chain being ergodic and ensures that f is the unique limiting distribution of the chain;
- ergodic averages converge to their expectation under f and further, we can apply central limit theorems.

This allows us to condition on previous information to construct a Markov chain whose invariant distribution is f, (for further references, see Tanner and Wong 1987, Geweke 1996, Gelfand and Smith 1990, Chib and Greenberg 1995 and Chib and Greenberg 1996). Unless otherwise mentioned, the processes to be considered are assumed to satisfy all the properties mentioned above.

3.2.2 Convergence of the Markov chain

How fast the chain converges to the stationary distribution and initial values become irrelevant are important to assess the efficiency of the MCMC method. The rate of convergence is determined through the correlation between the states and the dimensionality of the state vector, see Smith (1991). Further

references include Nummelin (1984) and Kalos and Whitlock (1986) for a full study on the rate of convergence, (for which the M-H has a geometric rate of convergence).

In the discrete setting, the transition probability obtained from moving from $X^{(j)}$ to $X^{(j+1)}$ can be written in matrix form as

$$P\left\{X^{(j)}|X^{(j+1)}\right\} = B\,P\left\{X^{(j+1)}|X^{(j)}\right\}$$

to show that the rate of convergence depends on the magnitude of the second largest eigenvalue of B, see Shanmugan and Breipohl (1988). Roberts and Sahu (1997) also showed that

$$\mathrm{E}\left\{X^{(j+1)}|X^{(j)}\right\} = BX^{(j)} + b$$

where b is constant. A definition of the rate of convergence, ρ, is given in Roberts and Sahu (1997), given by the following. Suppose X is a Markov chain with stationary density function $f(X)$, and h is a square $f-$integrable function of X. Let $\mathrm{E}_f(h)$ denote the expectation of h under the target density and ρ be defined as the minimum number such that, for all h and $r < \rho$, there exists a function $V \geq 1$ such that $f(V) < \infty$ and

$$\left|\mathrm{E}\left[h\left\{X^{(j)}\right\} - \mathrm{E}_f(h)\right]\right| \leq V\{X^{(0)}\}r^j. \tag{3.4}$$

If there exists a $\rho < 1$ satisfying (3.4), then the Markov chain is said to be *geometrically ergodic*, with rate of convergence ρ. For a further reference, see Chan and Geyer (1994). An interesting point to note is that geometric ergodicity is linked to mixing in standard asymptotic theory. In fact, a geometrically ergodic Markov chain, started at its invariant distribution is $\alpha-$mixing at a geometric rate, that is,

$$\alpha(N) = \sup_{A,B\subset\mathbb{R}^n}\left|P_f\left\{X^{(0)} \in A,\, X^{(N)} \in B\right\} - P\left\{X^{(0)} \in A\right\}\,P\left\{X^{(N)} \in B\right\}\right| = O(r^N),$$

for some $r > 1$, see Tierney, (1995, p. 67).

In general, it is not possible to compute convergence rates. When the target is Gaussian, for example, Roberts and Sahu (1997) provide exact computable rates of convergence for the Gibbs sampler. An application is considered in Section 5.4. The Gibbs sampler can be used in the case of the OU and bounds for the convergence rate can be computed, see Section 5.4.1. For most models, convergence rates have to be found numerically; this is considered in Section 5.4.2.

3.3 Sampling variates from a multivariate distribution

The Metropolis-Hastings algorithm proceeds by updating the components one at a time. Let $X_{\setminus i} = (X_1^{(j+1)}, \ldots, X_{i-1}^{(j+1)}, X_{i+1}^{(j)}, \ldots, X_n^{(j)})$ be a vector of all the components except X_i, where the components before X_i have already been updated, while those after X_i are at their values after the jth iteration. First, a *candidate point* Y is sampled from a *proposal distribution* $q(Y|X_i^{(j)}, X_{\setminus i})$. The candidate point

is then *accepted* with probability

$$\alpha\left\{X_i^{(j)}, Y\right\} = \min\left[1, \frac{f(Y|X_{\setminus i})q\{X_i^{(j)}|Y, X_{\setminus i}\}}{f\{X_i^{(j)}|X_{\setminus i}\}q\{Y|X_i^{(j)}, X_{\setminus i}\}}\right]$$
$$= \min\left[1, \frac{f(Y, X_{\setminus i})q\{X_i^{(j)}|Y, X_{\setminus i}\}}{f\{X_i^{(j)}, X_{\setminus i}\}q\{Y|X_i^{(j)}, X_{\setminus i}\}}\right]. \tag{3.5}$$

If the candidate point is accepted, we set $X_i^{(j+1)} = Y$; otherwise, $X_i^{(j+1)} = X_i^{(j)}$. The next component X_{i+1} is then updated similarly. One sweep of the algorithm consists of updating all components of X from $X_i^{(j)}$ to $X_i^{(j+1)}$. Note that it is only necessary to be able to evaluate $f(\cdot|\cdot)$ up to a constant of proportionality, since such constants will cancel in (3.5).

In general, the proposal distribution may depend on the current value $X_i^{(j)}$. Here, however, we will concentrate on the *independence sampler* where $q(Y|X_i^{(j)}, X_{\setminus i}) = q(Y|X_{\setminus i})$. A special case of the independence sampler is the *Gibbs sampler* where $q(Y|X_{\setminus i}) = f(Y|X_{\setminus i})$ and the acceptance probability $\alpha\{X_i^{(j)}, Y\}$ is thus 1 for all $X_i^{(j)}, Y$. An excellent review is given in Neal (1993). Geman and Geman (1984) initially proposed it for image processing and Gelfand and Smith (1990) for data analysis. It essentially breaks down the problem of sampling from a multivariate density by drawing successive samples from densities with smaller dimensions, (e.g. univariate).

The stationary distribution of the chain $\{X^{(j)}\}$ from an M-H algorithm is $f(X)$, irrespective of the form of the proposal distributions $q(\cdot|\cdot)$. However, the performance of the algorithm in finite runs depends crucially on the choice of the proposal. This affects both the rate of convergence of the chain to the stationary distribution and the rate of *mixing*, that is how rapidly the chain moves around the support of $f(X)$ after convergence. Mixing is slow when the proposed moves are short (Y is close to $X_i^{(j)}$), or when acceptance rates are low, in which case the candidate points are often rejected and the chain does not move. It is particularly important to avoid samplers where the acceptance probabilities are very low in some parts of the support of $f(X)$, as the chain may then get stuck in these areas for long periods.

A good proposal distribution $q(Y|X_{\setminus i})$ should be, in some sense, as close as possible to the target distribution $f(Y|X_{\setminus i})$. This is clearly the case for the Gibbs sampler where $q(Y|X_{\setminus i}) = f(Y|X_{\setminus i})$. In general, however, it is not possible to sample directly from the target distribution. It is then necessary to consider how the proposal distribution should be chosen so as to be 'close' to the target. Chib and Greenberg (1995) note that decreasing the variance of the proposal distribution increases the overall acceptance rate but reduces the length of the moves when the candidate value is accepted. It is also safest to select a proposal distribution which has heavier tails than the target distribution. This prevents the sampler from getting stuck in the tails of $f(X)$, see Gilks, Richardson, and Spiegelhalter (1996). To illustrate the selection of the proposal distribution, consider an extremely simple example.

Example 3.1. *Let the target distribution be univariate Gaussian with mean 0 and variance 1, and let the proposal be Gaussian with mean 0 and variance σ^2. Letting X denote the old value of the variable and Z the candidate value and omitting all other variables for simplicity, the acceptance probability is given by*

$$\alpha(X, Z) = \min\left[1, \exp\left\{-\frac{1}{2}\left(\frac{\sigma^2 - 1}{\sigma^2}\right)(Z^2 - X^2)\right\}\right] = \min\{1, \beta(X, Z)\}.$$

If $\sigma^2 = 1$, then $\alpha(X, Z) = 1$ and the algorithm is a Gibbs sampler for $f(X)$. If $\sigma^2 > 1$, the proposal distribution has a larger variance and heavier tails than the target distribution. In this case $\alpha(X, Z) = 1$

if $|Z| \leq |X|$. *The candidate value is thus always accepted if the proposed move is towards the mode of* $f(\cdot)$, *since* $f(Z)/q(Z) > f(X)/q(X)$. *For fixed* X, *the expected acceptance probability is*

$$\begin{aligned}\pi(X) &= \int \alpha(X,Z)q(Z)\,dZ = P[\alpha(X,Z)=1] + \{1 - P[\alpha(X,Z)=1]\}\int \beta(X,Z)q(Z)\,dZ \\ &= 1 - 2\Phi\left(-\frac{|X|}{\sigma}\right)\left\{1 - \frac{1}{\sigma}\exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}-1\right)X^2\right]\right\} = 1 - a(1-b).\end{aligned}$$

where $\Phi(\cdot)$ *is the standard normal distribution function. It can be seen that* $\pi(X) \to 1$ *as* $X \to \pm\infty$, *while* $\pi(X) = 1/\sigma$ *when* $X = 0$. *Candidate values are almost certain to be accepted when* X *is far from the mode of* $f(\cdot)$, *at the expense of reduced probability of acceptance around the mode.*

If $\sigma^2 < 1$, *moves away from the mode of* $f(X)$ *(*$|Z| \geq |X|$*) are always accepted, and the expected acceptance rate is* $\pi(X) = a + (1-a)b$. *In this case* $\pi(X) = 1$ *when* $X = 0$, *but* $\pi(X) \to 0$ *as* $X \to \pm\infty$. *When the proposal has lighter tails than the target density, the acceptance rates are large around the mode of* $f(\cdot)$ *but become arbitrarily small when* X *is far away from the mode.*

Both the Gibbs sampler and the M-H algorithm satisfy the detailed balance equations. That is, if one assumes that equilibrium of the chain has been reached, then the probabilities of transition must hold in both directions, see Geman and Geman (1984) for discussion on the Gibbs sampler. The rate of convergence is determined through the correlation between the states and the dimensionality of the state vector, see Smith (1991). To ensure convergence to the stationary distribution, consider the probability of a move in the case of the M-H algorithm,

$$\alpha\left\{X_i^{(j)}, Y\right\} = \min\left[1, \frac{f(Y, X_{\setminus i})q\{X_i^{(j)}|Y, X_{\setminus i}\}}{f\{X_i^{(j)}, X_{\setminus i}\}q\{Y|X_i^{(j)}, X_{\setminus i}\}}\right],$$

where the proposed draw $Y \sim q\{X_i^{(j)}|Y, X_{\setminus i}\}$ and f is the stationary distribution. The transition kernel for the M-H is given by

$$\begin{aligned}P\left\{Y|X^{(j)}, X_{\setminus i}\right\} &= q\left\{Y|X_i^{(j)}, X_{\setminus i}\right\}\alpha\left\{X_i^{(j)}, Y\right\} \\ &\quad + I\left\{Y = X_i^{(j)}\right\}\left[1 - \int q\left\{U|X_i^{(j)}, X_{\setminus i}\right\}\alpha\left\{X_i^{(j)}, U\right\}dU\right]\end{aligned}$$

where I denotes the indicator function. Using

$$f(Y, X_{\setminus i})\,q\{Y|X_i^{(j)}, X_{\setminus i}\}\,\alpha\left\{X_i^{(j)}, Y\right\} = f\{X_i^{(j)}, X_{\setminus i}\}\,q\{X_i^{(j)}|Y, X_{\setminus i}\}\,\alpha\left\{Y, X_i^{(j)}\right\}$$

the detailed balance equation is obtained from (3.5)

$$f(Y, X_{\setminus i})\,P\{Y|X_i^{(k)}, X_{\setminus i}\} = f\{X_i^{(j)}, X_{\setminus i}\}\,P\{X_i^{(k)}|Y, X_{\setminus i}\}$$

and integrating both sides with respect to $X_i^{(j)}$ gives

$$\int f(Y, X_{\setminus i})\,P\{Y|X_i^{(k)}, X_{\setminus i}\}\,dY = f\{X_i^{(j)}, X_{\setminus i}\}$$

which implies that if $X_i^{(j)}$ is from f, then so will be Y. Convergence results are given in Roberts (1995). An important condition is irreducibility, which is guaranteed, provided f is everywhere positive.

3.3.1 Metropolis-Hastings for sampling the auxiliary states

The M-H algorithm, within a general MCMC method, will be used to sample both the auxiliary states Y^* and the parameter θ, (except in cases where the posterior for some components of θ is available in closed-form). A general outline will be given here and covered in detail in Sections 4 and 5. The key step in the method is obtaining the posterior density, $f(Y^*;\theta|Y)$. To analyse the posterior density we can utilise Markov chain Monte Carlo (MCMC) simulation methods. These methods allow one to sample the target density (in this case the augmented posterior density) by simulating a Markov chain $\{\theta^j, Y^{*j}\}$ constructed to have $f^M(\theta, Y^*|Y)$ as the limiting invariant distribution. The trajectory of the Markov chain, after an initial transient or burn-in stage, provides a sequence of (correlated) draws from $f^M(\theta, Y^*|Y)$. Furthermore, the draws $\{\theta^j\}$ are automatically from the marginal distribution

$$f^M(\theta|Y) = \int f^M(\theta, Y^*|Y)dY^*,$$

and can be used to conduct posterior inferences about θ. For example, the sample mean and the sample standard deviation of the sampled draws are estimates of the corresponding posterior mean and posterior standard deviations; simulation consistency of these estimates is established by ergodic laws of large numbers for Markov chains on continuous state spaces. This leads to full likelihood inference for the model, even though the likelihood is not available.

For illustration, consider the issue of sampling m points, denoted by $w = w_1, \ldots, w_m$, conditional on fixed end points w_0 and w_{m+1}. so that we simulate w from some density $f(w|w_0, w_{m+1})$. The proposal density to approximate $f(w|w_0, w_{m+1})$ is chosen so that it is easy to sample from, this gives us a proposal $q(w)$, where we suggest values $z \sim q(w)$, such that

$$q(w) = q(w|w_0, w_{m+1}) \sim N(\tilde{\mu}, \tilde{\Sigma}),$$

and is accepted with probability

$$\alpha(z, w^{(j)}) = \min\left\{1, \frac{f(z|w_0^{(j+1)}, w_{m+1}^{(j)})q(w^{(j)})}{f(w^{(j)}|w_0^{(j+1)}, w_{m+1}^{(j)})q(z)}\right\}.$$

where j represents the iterations. The neighbours w_0 and w_{m+1} are always the most recently sampled values. The idea of random block sizes is discussed in Section 5 and used in a time series context by Shephard and Pitt (1997) . The M-H algorithms proposed here are independence samplers, (see for example Chib and Greenberg 1995), where the proposal does not depend on the current value of w. However, because the mean of the proposal depends on the immediate neighbours of w and because the underlying process for the observations is continuous, the mean of the proposal is typically close to w. This has the important consequence that the algorithms proposed here will normally have high acceptance probabilities and are very unlikely to get stuck in the tails of the target distribution. To illustrate this, consider an independence sampler where the mean of the proposal distribution is exactly equal to w. Let z be the candidate value for w. In this case $q(z)/q(w) < 1$ for all z. If z is closer to the mode of $f(w|w_0, w_{m+1})$ than w, $f(z|\cdot)/f(w) > 0$ and the acceptance probability is 1, (assuming the target distribution is unimodal) . Proposed moves away from the tails of the target distribution are thus always accepted. Acceptance probabilities for moves away from the mode are also likely to be high since $q(z)/q(w)$ is small.

Different choices for the proposal distributions and for the division of w into sampling components are

considered in sections 4 and 5. Sections 4.1 and 4.1.3 examine three single-move samplers. Other work in this area has included Smith and Roberts (1993) and Jacquier, Polson, and Rossi (1994) who use a cyclic Metropolis algorithm to analyse the stochastic volatility model. Once the states have been sampled, the sampling of the parameters θ is straightforward. Methods for the parameters are outlined in section 6. In a Bayesian framework, the discretized version of the likelihood of the states will be combined with any prior knowledge on θ to obtain $f(\theta|Y, Y^*)$. The posterior will often be available in closed form implying that we can use a Gibbs sampler to generate the parameter values.

3.3.2 Taylor approximations of the log-likelihood

Under some circumstances, a Gaussian form can well approximate the posterior density. As the number of observations increases, by the central limit theorem, the posterior density tends to a normal distribution. Consider the Taylor approximation of the log-likelihood,

$$l(X) = l(\hat{X}) + (X - \hat{X})' \left.\frac{\partial l}{\partial X}\right|_{X=\hat{X}} + \frac{1}{2}(X - \hat{X})' \left.\frac{\partial^2 l}{\partial X \partial X'}\right|_{X=\hat{X}} (X - \hat{X}).$$

If $\hat{X}$ is the mode, then

$$\left.\frac{\partial l}{\partial X}\right|_{X=\hat{X}} = 0$$

and taking the exponential,

$$L(X) = L(\hat{X}) \exp\left\{\frac{1}{2}(X - \hat{X})' \left(\left.\frac{\partial^2 l}{\partial X \partial X'}\right|_{X=\hat{X}}\right)(X - \hat{X})\right\},$$

the form of a Gaussian density. Gaussianity can also be reinforced by reparameterization. Accuracy is limited to a small area around the mode. As T, the number of observations, increases, the curvature of the mode around the log-likelihood increases at the rate $O(\sqrt{T})$. See Walker (1969) for issues concerning the accuracy of the Taylor approximation. Expansion of the log-likelihood using a Taylor approximation will be considered in this thesis when choosing the proposal. Iterating the second order expansion to the mode results in a *Laplace* approximation.

3.3.3 Convergence diagnostics

Influential work in this area has been carried out by Gelman (1995), Gelman and Rubin (1992b), Gelman and Rubin (1992a), Gilks and Roberts (1995), Hills and Smith (1992), Raftery (1995), Raftery and Lewis (1995a) and Raftery and Lewis (1992) which should be consulted for further reading. Diagnostics for assessing convergence are discussed in more detail in Raftery and Lewis (1995b) and Gelman (1995).

3.3.4 Summarising the MCMC output

The posterior distribution can convey the information through a number of key quantities to reflect the location and variability of the posterior. We will concentrate on the posterior mean as we are interested in its properties as a classical estimator. The posterior mean is defined as the expected value of θ under the posterior distribution $f(\theta|X)$. The covariance and correlation matrices measure the association between

the random variables and so reflect the precision of the parameters and their dependence structure. *Credible* regions will often be reported which are based on the marginal distribution. These reflect the dispersion of each parameter and are summarised through the *Highest Probability Density* (HPD), region. The HPD region represents the shortest interval that contains a given mass of the posterior. A definition is provided in Besag, Green, Higdon, and Mengersen (1995).

To assess how reliable the MCMC output is, *Inefficiency factors* or the *autocorrelation time* can be computed. The inefficiency factor of each posterior estimate (computed as a sample average over the simulated values) is defined as the spectrum at zero, $1+2\sum_{j=1}^{\infty}\rho_j$, where $\{\rho_j\}$ denotes the autocorrelation function of the simulated values. It is equal to square of the numerical standard error divided by the variance of the posterior estimate under (hypothetical) i.i.d. sampling from the posterior. Geweke, (1989, 1992), defines the alternative measure of relative numerical efficiency which is the inverse of the inefficiency factor. The idea, (see Ripley 1987 and Neal 1993) is to base the standard error on some effective sample size, which is found by dividing the number of iterations by the autocorrelation time or inefficiency factor, defined in Section 4.2.

In simulation studies, the bias, mean square error and the mean of the standard error can additionally be used to supplement diagnostics. The bias is defined as the difference between the expectation of a biased estimator and the parameter θ which is being estimated. The mean square error (MSE) is used to evaluate the quality of an estimator. We focus on those that provide the minimum MSE, where the MSE is defined by

$$\frac{1}{N}\sum_{j=1}^{N}\{\hat{\theta}^{(j)}-\theta\}^2,$$

where N is the total number of iterations and $\hat{\theta}^{(j)}$ represents the value of θ obtained at the j'th iterate.

Graphical summaries are provided through *histograms*, *paths* and *correlograms*. The histogram, based on the Gaussian kernel, provides an estimate of the marginal density of each parameter. In this context, histograms are a useful graphical aid for presenting the MCMC output, highlighting properties such as multimodality. Further details can be found in Doornik (1996) and Silverman (1986).

3.3.5 Numerical accuracy

By increasing the number of iterations, one can assess the accuracy of the approximation of $\mathrm{E}\{g(\theta)\}$, for any function g, by

$$\tilde{g}(\theta)=\frac{1}{N}\sum_{j=1}^{N}g\{\theta^{(j)}\}$$

and under certain regularity conditions the estimator will be consistent. The posterior variance is given by

$$\mathrm{var}\{g(\theta)\}=\mathrm{E}g^2(\theta)-\{\mathrm{E}g(\theta)\}^2=\sigma^2.$$

One can measure the numerical accuracy of the approximation as

$$\mathrm{var}\{\tilde{g}(\theta)\}=\tilde{\sigma}^2=\frac{1}{N}\sum_{j=1}^{N}g^2\{\theta^{(j)}\}-\tilde{g}^2.$$

The *absolute numerical error* is then given by $\tilde{\sigma}/\sqrt{N}$ and the *relative numerical error* is given by $\tilde{\sigma}/\sigma\sqrt{N}$, that is, the absolute numerical error scaled with the standard deviation of the random variable $g(\theta)$. For further reading, consult Mood, Graybill, and Boes, (1974, Chapters 2 and 6), Cramèr, (1946, Chapters 25 and 27) and Billingsley (1979).

3.3.6 Efficiency in designing MCMC algorithms and convergence diagnostics

It is difficult to determine how large the number of iterations should be when carrying out MCMC analysis. Authors that have looked at this issue include Meyn and Tweedie (1994), Polson (1996), Roberts and Polson (1994) and Roberts and Tweedie (1994). A good discussion on convergence diagnostics is given in the discussion paper Carlin, Gelman, and Neal (1998). We will rely on diagnostics discussed above and refer the reader to the aforementioned authors for further reading on this issue. Often the posterior structure leads to high correlation between some of the components of θ. The chain may take many iterations to cover the sample space and so convergence may be slow. Issues to be considered include how many iterations should be discarded at the initial stage. More efficient use of information from the Monte Carlo simulation can improve inference. Typically, efficiency depends on the scale and shape of the target distribution. The most efficient proposal should typically be shaped like the target distribution; suggestions for various M-H samplers can be found in Gelman (1995), Müller (1993) and Roberts (1995).

Transformations of the parameter space may produce samplers that are close to being independent and identically distributed (i.i.d), and so increase convergence of the chain. For densities that are non-Gaussian, reparameterization can make densities look more normal, for example, in reducing thickness in the tails. In general, a parameterization scheme should be chosen to make the correlations between the coordinate directions (of the invariant distribution) as small as possible. Simple linear approximations can be used to produce an approximately uncorrelated parameterization, see Wakefield (1993) and Hills and Smith (1992). One ad-hoc method is given by a linear transformation that has a diagonal variance matrix. These issues will be considered with the Hyperbolic diffusion in Section 6.5 and with use of the Milstein density in Section 9. Examples can also be found in Roberts and Sahu (1997) and Vines, Gilks, and Wild (1994). We also consider log-transformations for the models, (which is recommended by Smith (1991) for random variables that must be positive, for example, variance parameters). Roberts (1992) showed that reparameterization may not necessarily improve efficiency for non-normal target distributions.

Another issue which may help in the design of efficient algorithms is the use of *blocking*. We consider single-state sampling and updating all components of the state vector in one sweep and find that faster convergence is obtained by using blocks of size m, where $1 \leq m \leq M$. Computationally, this can be very beneficial where there is high correlation between the components. Other authors who have considered this are Liu, Wong, and Kong (1994) and Roberts and Sahu (1997).

On numerical issues, the density function can fluctuate between extreme values and can cause numerical problems. To overcome this issue, we work with log densities. One justification for doing so is that for the region around the mode, the log density is approximately quadratic which seems suitable for many optimisation techniques, see Ruanaidh and Fitzgerald (1996).

The output from the MCMC algorithm is often thinned for storage reasons, (see Carlin, Gelman, and Neal 1998). Thinning amounts to basing inference on every k'th iteration, where k is chosen such that successive draws of θ are approximately independent. Since the algorithms are the Markov chains, one

can use the draws from the sampler as a basis for inference by appealing to suitable ergodic theorems for Markov chains. An extensive treatment of geometric convergence and central limit theorems for Markov chains can be found in Meyn and Tweedie (1993) and Geyer (1992). See Chan (1993), Schervish and Carlin (1992) and Roberts and Polson (1994) for the Gibbs sampler specifically and Roberts and Tweedie (1994) for the M-H sampler. Tierney (1995) extends these results.

For estimation which does not involve any simulation, one can identify sampling variability which depends on the sample size, T. The variance of an estimate can be computed to make an inferential statement on θ, the parameter of interest. To check the quality of the estimate of θ and gauge the accuracy in the estimate of its variance, Monte Carlo simulation can be conducted, using simulated data. The estimate of the variance for these simulations can be compared to the mean of the estimated variance. One can even take this a step further and compare this to the properties of some other estimator, e.g. the maximum likelihood estimate, when it is available, (e.g. Ornstein-Uhlenbeck process). Simulating from the posterior adds an additional source of sampling variability and the numerical efficiency can be assessed through Monte Carlo standard errors, correlation of the sampler and the covariance structure between the parameters. It will be necessary to minimise the Monte Carlo variance so that the main source of variability comes from the original sample of observations. This will be tested in Section 6.2.3 when we perform MCMC's on the CIR process using 10 simulated datasets.

To monitor the convergence of each parameter, we can additionally compute the Gelman-Rubin r statistic: see Gelman and Rubin 1992a and Gelman, Carlin, Stern, and Rubin, (1995, pp. 331-332). Suppose we denote one of the parameters of θ as ϕ and run J parallel sequences of the M-H algorithm with overdispersed starting values. Each sequence is run for N iterations, to obtain draws ϕ_{ij}, $(i = 1, \ldots, N; j = 1, \ldots, J)$. The *between*, B and *inbetween*, W, sequence variances can be computed as

$$B = \frac{N}{J-1}\sum_{j=1}^{J}(\bar{\phi}_{.j} - \bar{\phi}_{..})^2, \quad \text{where } \bar{\phi}_{.j} = \frac{1}{N}\sum_{i=1}^{N}\phi_{ij}, \quad \bar{\phi}_{..} = \frac{1}{J}\sum_{j=1}^{J}\bar{\phi}_{.j},$$

$$W = \frac{1}{J}\sum_{j=1}^{J}s_j^2, \quad \text{where } s_j^2 = \frac{1}{N}\sum_{i=1}^{N}(\phi_{ij} - \bar{\phi}_{.j})^2.$$

The estimate $\text{var}(\phi|Y)$ is then given by a weighted average of B and W as

$$\text{var}^+(\phi|Y) = \frac{N-1}{N}W + \frac{1}{N}B,$$

which is in fact an overestimate of the variance of the marginal posterior density. The convergence can be assessed through $\sqrt{\hat{R}}$, where

$$\sqrt{\hat{R}} = \sqrt{\frac{\hat{\text{var}}^+(\phi|Y)}{W}} \overset{N\to\infty}{\to} 1.$$

3.4 Using residuals to assess model fit

The model is formulated based on the assumptions and prior information. Estimation and fitting of the model is then carried out to obtain estimates and carry out inference. Diagnostic tests on the residuals are then used to assess the overall fit of the model to the data. References on residuals include Box (1979), Box (1980) and Cook and Weisberg (1982). It should be noted that most residual based tests for specific alternatives for the errors are sensitive to several alternatives, for example, outliers

affect formal procedures that use outliers. In the absence of specific alternative models or special design considerations, residuals or their transformations provide a useful tool.

Assessing whether the model provides a good fit to the data can be carried out using the one-step-ahead prediction distribution functions. Throughout we will assume that the SDE is univariate, although extensions to the multivariate case are possible.

Consider the predictive distribution function for the discretely observed diffusion

$$G(y_{t+1}|y_t;\theta) = \Pr(Y_{t+1} \leq y_{t+1}|y_t;\theta),$$

evaluated at the observed value y_{t+1}. We now show that this is uniform $UID(0,1)$ under the correctness of the model. Clearly $u_{t+1} = G(y_{t+1}|y_t;\theta)$ lies in the interval $(0,1)$. Next,

$$\begin{aligned}\Pr(u_{t+1} &\leq a) = \Pr\{G(y_{t+1}|y_t;\theta) \leq a\} \\ &= \Pr\left\{y_{t+1} \leq G^{-1}(a)|y_t;\theta\right\} \\ &= a,\end{aligned}$$

which shows that u_{t+1} is uniform. Finally, since $u_{t+1}|u_t$ only depends on u_t, $u_{t+1}|u_t$ is an independent and identically distributed sequence. This result, in a different context, goes back at least to Rosenblatt (1952) and has been used by Pedersen (1994b).

We approximate $G(y_{t+1}|y_t,\theta)$ by

$$F^M(y_{t+1}|y_t,\theta) = \int \Pr(Y_{t+1} \leq y_{t+1}|y^*_{t,M},\theta) f^M(y^*_t|y_t,\theta)dy^*_t,$$

where $\Pr(Y_{t+1} \leq y_{t+1}|y^*_{t,M},\theta)$ is easily computed due to the fact that $Y_{t+1}|y^*_{t,M},\theta$ is Gaussian. Pedersen suggested estimating this integral using Monte Carlo methods, by drawing $y^{*(j)}_t \sim f^M(y^*_t|y_t,\theta)$ a large number of times and forming the average:

$$\hat{u}_{t+1} = \widehat{F}^M(y_{t+1}|y_t,\theta) = \frac{1}{R}\sum_{j=1}^{R} \Pr\left(Y_{t+1} \leq y_{t+1}|y^{*(j)}_{t,M},\theta\right).$$

The direct Monte Carlo estimate can be improved by importance sampling as was suggested for the estimation of the likelihood ordinate above. Again the results of Talay and Tubaro (1990) imply that

$$E\left|\hat{u}_{t+1} - G(y_{t+1}|y_t,\theta)\right| \leq C\left(\frac{1}{\sqrt{R}} + \frac{\Delta^\dagger}{M+1}\right).$$

Adequacy of the model is judged by the serial correlation in the $\{\hat{u}_{t+1}\}$ and by the nature of the distributional shape of both the $\{\hat{u}_{t+1}\}$ and its reflected version $\left\{2\left|\hat{u}_{t+1} - \frac{1}{2}\right|\right\}$, the latter providing information on dispersion. These statistics are also used in Rosenblatt (1952), Pedersen (1994a), Dawid (1982), Smith (1985b) and Shephard (1994). The idea of focusing on $2|\hat{u}_t - 0.5|$ appears in Kim, Shephard, and Chib (1998). In Section 7, where we consider the application of these ideas to assess model fit for the Eurodollar deposit spot rate data, we replace θ by some estimator of the parameter – usually the posterior mean. This leads to another layer of approximation. An alternative would have been to sequentially integrated out the effect of θ from these calculations, computing $G(y_{t+1}|y_t)$. A discussion of alternative ways of carrying out the implied computations for time series models is given in Gerlach, Carter, and Kohn (1999).

As a by-product, we compute the standardised forecast errors (see Pedersen 1994a) as

$$\frac{y_{t+1} - \widehat{\mathrm{E}}(y_{t+1}|y_t, \theta)}{\sqrt{\widehat{\mathrm{var}}(y_{t+1}|y_t, \theta)}},$$

where $\widehat{\mathrm{E}}(y_{t+1}|y_t, \theta) = \frac{1}{R}\sum_{j=1}^{R} Y_{t+1}^{(j)}$ and $\widehat{\mathrm{var}}(y_{t+1}|y_t, \theta) = \frac{1}{R-1}\sum_{j=1}^{R}\{Y_{t+1}^{(j)} - \widehat{\mathrm{E}}(y_{t+1}|y_t, \theta)\}^2$. Properties of these statistics are harder to evaluate then for the $\hat{u}_{t+1}$, but provide a graphical aid to help explain the inadequacies of the fitted model.

3.5 Comparison of non-nested models

We now very briefly consider the question of comparing alternative, potentially non-nested, diffusion models that have been fit to a given data set. One useful technique for making this comparison is through the calculation of the model marginal likelihood, $m(Y|\mathcal{M})$ under some model $\mathcal{M}$, which is defined as the integral of the likelihood function with respect to the prior density. This quantity can be computed from the MCMC output by a method developed in Chib (1995). Ratios of marginal likelihood give rise to the Bayes factor, given for example by $m(Y|\mathcal{M}_1)/m(Y|\mathcal{M}_2)$, for two models $\mathcal{M}_1$ and $\mathcal{M}_2$.[1] This statistic shows the evidence present in the data, which is in favour of the numerator model, relative to the model in the denominator.

Let θ denote the parameters of a given diffusion model $\mathcal{M}$, with likelihood function $f^M(Y|y_1, \theta, \mathcal{M})$ and prior density $\pi(\theta|\mathcal{M})$, where the likelihood function is computed using the efficient method just outlined. Then, the marginal likelihood of model $\mathcal{M}$ can be written as

$$m(Y|\mathcal{M}) = \frac{f^M(Y|y_1, \theta, \mathcal{M})\pi(\theta|\mathcal{M})}{\pi(\theta|Y, \mathcal{M})}.$$

The key point is that this expression, which is a consequence of Bayes theorem, is an identity in θ and can therefore be evaluated at any appropriately selected point θ^* (say). If θ^* denotes a high density point and $\hat{\pi}(\theta^*|Y, \mathcal{M})$ the estimate of the posterior ordinate at θ^*, then the marginal likelihood on the log scale is estimated as

$$\log m(Y|\mathcal{M}) = \log f^M(Y|y_1, \theta^*, \mathcal{M}) + \log \pi(\theta^*|\mathcal{M}) - \log \hat{\pi}(\theta^*|Y, \mathcal{M}), \tag{3.6}$$

where the first term is the value of the log likelihood function at θ^*, found using (8.13) and the second term is the log of the prior density which is available directly. The third term is estimated from the MCMC output by either kernel smoothing (if the dimension of θ is small) or by a marginal/conditional decomposition of the posterior ordinate followed by reduced MCMC runs to generate the draws necessary to estimate each of the marginal/conditional ordinates (see Chib 1995 for further details). It should be noted that this approach for estimating the marginal likelihood only requires a single evaluation of the likelihood function and the posterior ordinate and is thus computationally inexpensive.

[1] In Bayes factors we need to compute the relative likelihood functions of two models.

3.6 Conclusion

MCMC methods have been outlined in this Chapter, which together with the theory on stochastic differential equations outlined in Chapter 2, provide the background to the remaining parts of the thesis. Part II will now develop the method for sampling the auxiliary states, given the observed data points and the parameter of interest.

Part II

Contribution to the Literature

Contribution to the Literature

4 Single-Move Metropolis-Hastings Algorithms for the States **55**

4.1 Proposal distributions 55

4.2 Discussion and comparison 60

4.3 Analytic convergence rates for the OU process 64

4.4 Conclusion 66

4.5 Appendix 67

5 Block Sampling for the Auxiliary States **70**

5.1 A block Metropolis–Hastings sampler 71

5.2 Examples 74

5.3 How to ensure a positive definite proposal variance matrix 87

5.4 Assessing the efficiency of the sampler 90

5.5 Conclusion 94

5.6 Appendix 94

6 Parameter Estimation **97**

6.1 Sampling the parameters 97

6.2 Simulation considerations and experiments 107

6.3 Updating all M states in one block 113

6.4 Additional models 119

6.5 Hyperbolic diffusions 125

6.6 Conclusion 135

7 Application to Interest-rate data **136**

7.1 The data 136

7.2 Parametric specifications of the data to be considered 137

7.3 Results 138

7.4 MCMC estimation: updating the whole block in one sweep 142

7.5 Extensions 145

7.6 Conclusion 147

8 Estimating the likelihood for stochastic differential equations 149

8.1 Monte Carlo techniques for estimating integrals 149

8.2 Multivariate proposal densities 151

8.3 Pedersen's method 152

8.4 Geyer's approach 157

8.5 Importance sampling approach 158

8.6 Comparisons 159

8.7 Conclusion 162

8.8 Appendix 163

Part II is concerned with using MCMC methods to sample from the joint distribution of the parameter of interest θ and the auxiliary states Y^*, given the observed states Y. This is acheived by considering each conditional distribution in turn, $Y^*|Y;\theta$ and $\theta|Y,Y^*$. The next Chapter will present a single-move Metropolis-Hastings algorithm with applications to the Ornstein-Uhlenbeck process and the quadratic drift process. The drift term is first approximated by a zero-order Taylor expansion around the neighbours of the point to be updated and then a first-order Taylor expansion. The latter is shown to be more efficient in estimating the process. This Chapter also demonstrates the advantages in using a Taylor approximation to deal with any non-Gaussianity present in the likelihood of the process. Numerical convergence rates for the samplers and theoretical estimates of the efficiency are also considered. Chapter 5 then shows the increase in efficiency from block sampling. Additional convergence rates and efficiency issues are examined and applications to various models, including the Vasicek model, are presented. Chapter 6 outlines the procedure for obtaining the posterior estimates of the parameters.

Based on the MCMC output, techniques for computing the likelihood function, the marginal likelihood and diagnostic measures are covered in the remainder of Part II. The algorithm can be applied to simulated data, in order to assess the performance of the method. Various parametric models for the 7-day Eurodollar deposit spot-rate are considered in Chapter 7 and Chapter 8 compares three different methods for obtaining an expression for the likelihood.

CHAPTER 4

SINGLE-MOVE METROPOLIS-HASTINGS ALGORITHMS FOR THE STATES

In this Chapter we consider the M-H algorithm which updates $y_t^* = (y_{t,1}^*, \ldots, y_{t,M}^*)'$ one scalar component at a time. The parameter θ is assumed known for the moment and considered in Section 6. Section 4.1 considers use of a zero-order and first-order Taylor approximation of the states. This is then extended to considering a Taylor approximation of the likelihood. A discussion and comparison is provided in Section 4.2. For some processes, we can obtain closed form expressions for the convergence rate of the MCMC algorithm, which is the subject of Section 4.3. The Chapter concludes with Section 4.4.

Let $y_{t,\backslash k}^*$ (for $k = 1, \ldots M$) denote all elements of y_t^* excluding $y_{t,k}^*$. An approximate conditional density of $y_{t,k}^*$ given $y_{t,\backslash k}^*$, $y_{t,0}^* = y_t$, and $y_{t,M+1}^* = y_{t+1}$, is

$$\begin{aligned} f(y_{t,k}^*|y_{t,\backslash k}^*, y_{t,0}^*, y_{t,M+1}^*) &= f(y_{t,k}^*|y_{t,k-1}^*, y_{t,k+1}^*) \\ &\propto f(y_{t,k}^*|y_{t,k-1}^*) f(y_{t,k+1}^*|y_{t,k}^*) \end{aligned} \tag{4.1}$$

where $f(y_{t,k}^*|\cdot)$ and $f(y_{t,k+1}^*|\cdot)$ are normal densities, given by the Euler approximation (2.12). Because the process is Markovian, only the two neighbouring points need be considered at each point.

4.1 PROPOSAL DISTRIBUTIONS

We consider the zero-order and first-order Taylor approximation of the states and extend the discussion to evaluate a Taylor approximation of the likelihood. In each case, the proposal density to approximate the target density in the M-H algorithm, is normal with a mean and variance that depends only on the fixed endpoints.

The M-H algorithm is initialised by computing $y_{t,k}^*$, $k = 1, ..., M$ from

$$y_{t,k}^* = y_{t,0}^* + (y_{t,M+1}^* - y_{t,0}^*)\{k/(M+1)\}$$

for $k = 1, ..., M$, which is simply the straight line between $y_{t,0}^* = y_t$ and $y_{t,M+1}^* = y_{t+1}$. For each k and iteration j, we compute the parameters of the proposal distribution, $\tilde{\mu}_t$ and $\tilde{\sigma}_t^2$ from one of the approximations, i.e. $\tilde{\mu}_{t,i}$ and $\tilde{\sigma}_{t,i}^2$ for $i = 0$ or $i = 1$ using $y_{t,k-1}^{*(j+1)}$ and $y_{t,k+1}^{*(j)}$ and generate a candidate value z from the proposal density $q : y_{t,k}^*|y_{t,k-1}^{*(j+1)}, y_{t,k+1}^{*(j)} \sim N(\tilde{\mu}_{t,i}, \tilde{\sigma}_{t,i}^2)$. The candidate value is accepted

with probability

$$\alpha\{y_{t,k}^{*(j)}, z\} = \min\left[1, \frac{f\left\{z|y_{t,k-1}^{*(j+1)}\right\} f\left\{y_{t,k-1}^{*(j+1)}|y_{t,k+1}^{*(j)}\right\} q\left\{y_{t,k}^{*(j)}\right\}}{f\left\{y_{t,k}^{*(j)}|y_{t,k-1}^{*(j+1)}\right\} f\left\{y_{t,k-1}^{*(j+1)}|y_{t,k+1}^{*(j)}\right\} q(z)}\right]$$

where $y_{t,k}^{*(j)}$ is the value of $y_{t,k}^{*}$ from the previous iteration.

4.1.1 Zero order Taylor approximation of the states

The first term in (4.1) is quadratic in $y_{t,k}^{*}$, but the second one will in general not be. By approximating $f(y_{t,k+1}^{*}|y_{t,k}^{*})$ by a Gaussian density $q(y_{t,k+1}^{*}|y_{t,k}^{*})$ which is at most quadratic in $y_{t,k}^{*}$, we can approximate (4.1) by a Gaussian density $q(y_{t,k}^{*}|y_{t,k-1}^{*}, y_{t,k+1}^{*})$. As a first approach, replace $a(y_{t,k}^{*})$ by $a(\hat{y}_{t,k}^{*})$ and $b(y_{t,k}^{*})$ by $b(\hat{y}_{t,k}^{*})$, where $\hat{y}_{t,k}^{*}$ is some fixed value. Here we will use $\hat{y}_{t,k}^{*} = (y_{t,k-1}^{*} + y_{t,k+1}^{*})/2$. This represents a zero order Taylor approximation of $a(y_{t,k}^{*})$ and $b(y_{t,k}^{*})$. Denote $b^2(y_{t,k-1}^{*}) = \sigma_{t,k-1}^{2*}$ and $b^2(\hat{y}_{t,k-1}^{*}) = \hat{\sigma}_{t,k-1}^{2*}$. Let $h(y_{t,k+1}^{*}|y_{t,k}^{*})$ be the density of a normal distribution with mean $y_{t,k}^{*} + a(\hat{y}_{t,k}^{*})\Delta$ and a variance of $\hat{\sigma}_{t,k}^{2}\Delta$. Then

$$\begin{aligned}\log f(y_{t,k}^{*}|y_{t,k-1}^{*}, y_{t,k+1}^{*}) &\approx c + \log h(y_{t,k+1}^{*}|y_{t,k}^{*}) + \log f(y_{t,k}^{*}|y_{t,k-1}^{*}) \\ &= c' - \frac{1}{2\tilde{\sigma}_{t,0}^{2}}(y_{t,k}^{*} - \tilde{\mu}_{t,0})^2\end{aligned}$$

where

$$\tilde{\sigma}_{t,0}^{2} = \left(\frac{1}{\hat{\sigma}_{t,k}^{2}} + \frac{1}{\sigma_{t,k-1}^{2}}\right)^{-1}\Delta = \left(\frac{\hat{\sigma}_{t,k}^{2}\sigma_{t,k-1}^{2}}{\hat{\sigma}_{t,k}^{2} + \sigma_{t,k-1}^{2}}\right)\Delta \tag{4.2}$$

and

$$\tilde{\mu}_{t,0} = \frac{\tilde{\sigma}_{t,0}^{2}}{\Delta}\left(\frac{y_{t,k+1}^{*} - a(\hat{y}_{t,k}^{*})\Delta}{\hat{\sigma}_{t,k}^{2}} + \frac{y_{t,k-1}^{*} + a(y_{t,k-1}^{*})\Delta}{\sigma_{t,k-1}^{2}}.\right) \tag{4.3}$$

The proposal density for $y_{t,k}^{*}$ is thus normal with mean $\tilde{\mu}_{t,0}$ and variance $\tilde{\sigma}_{t,0}^{2}$.

4.1.2 First order Taylor approximation of the states

The second method of approximation uses the first-order Taylor expansion of $a(y_{t,k}^{*})$ around $\hat{y}_{t,k}^{*}$, given by

$$\begin{aligned}a(y_{t,k}^{*}) &\approx a(\hat{y}_{t,k}^{*}) + a'(\hat{y}_{t,k}^{*})(y_{t,k}^{*} - \hat{y}_{t,k}^{*}) \\ &= c_{t,0} + c_{t,1}y_{t,k}^{*}\end{aligned}$$

where $c_{t,0} = a(\hat{y}_{t,k}^{*}) - a'(\hat{y}_{t,k}^{*})\hat{y}_{t,k}^{*}$ and $c_{t,1} = a'(\hat{y}_{t,k}^{*})$. The volatility term $b(y_{t,k}^{*})$ is again approximated by a zero-order approximation $b(\hat{y}_{t,k}^{*})$. Let $h(y_{t,k+1}^{*}|y_{t,k}^{*})$ be the density of $N(y_{t,k}^{*} + c_{t,0}\Delta + c_{t,1}\Delta y_{t,k}^{*}, \hat{\sigma}_{t,k}^{2}\Delta)$,

where $\hat{\sigma}^2_{t,k} = b^2(\hat{y}^*_{t,k})$. Then

$$\begin{aligned}\log f(y^*_{t,k}|y^*_{t,k-1}, y^*_{t,k+1}) &\approx c + \log h(y^*_{t,k+1}|y^*_{t,k}) + \log f(y^*_{t,k}|y^*_{t,k-1}) \\ &= c - \frac{1}{2\Delta}\left[\frac{\{y^*_{t,k+1} - c_{t,0}\Delta - (1+c_{t,1}\Delta)y^*_{t,k}\}^2}{\hat{\sigma}^2_{t,k}}\right. \\ &\quad \left. + \frac{\{y^*_{t,k} - y^*_{t,k-1} - a(y^*_{t,k-1})\Delta\}^2}{\sigma^2_{t,k-1}}\right] \\ &= c' - \frac{1}{2\tilde{\sigma}^2_{t,1}}(y^*_{t,k} - \tilde{\mu}_{t,1})^2.\end{aligned}$$

So $y^*_{t,k}|y^*_{t,k-1}, y^*_{t,k+1}$ is approximately $N(\tilde{\mu}_{t,1}, \tilde{\sigma}^2_{t,1})$ with

$$\tilde{\sigma}^2_{t,1} = \left\{\frac{1}{\sigma^2_{t,k-1}} + \frac{(1+c_{t,1}\Delta)^2}{\hat{\sigma}^2_{t,k}}\right\}^{-1}\Delta = \left(\frac{1}{\sigma^2_{t,k-1}} + \frac{(1+a'(\hat{y}^*_{t,k})\Delta)^2}{\hat{\sigma}^2_{t,k}}\right)^{-1}\Delta \tag{4.4}$$

$$\begin{aligned}\tilde{\mu}_{t,1} &= \frac{\tilde{\sigma}^2_{t,1}}{\Delta}\left\{\frac{y^*_{t,k-1} + a(y^*_{t,k-1})\Delta}{\sigma^2_{t,k-1}} + \frac{y^*_{t,k+1} + c_{t,1}\Delta y^*_{t,k+1} - c_{t,0}(1+c_{t,1}\Delta)\Delta}{\hat{\sigma}^2_{t,k}}\right\} \\ &= \frac{\tilde{\sigma}^2_{t,1}}{\Delta}\left[\frac{y^*_{t,k-1} + a(y^*_{t,k-1})\Delta}{\sigma^2_{t,k-1}} + \frac{y^*_{t,k+1} + a'(\hat{y}^*_{t,k})\Delta y^*_{t,k+1} - c_{t,0}\{1+a'(\hat{y}^*_{t,k})\Delta\}\Delta}{\hat{\sigma}^2_{t,k}}\right]\end{aligned} \tag{4.5}$$

and $c_{t,0} = a(\hat{y}^*_{t,k}) - a'(\hat{y}^*_{t,k})\hat{y}^*_{t,k}$.

4.1.3 The Taylor approximation of the log-likelihood

As in the previous Section, consider the target function

$$\begin{aligned}l(y^*_{t,k}) &= \log f(y^*_{t,k}|y^*_{t,k-1}, y^*_{t,k+1}) \\ &= c + \log f(y^*_{t,k+1}|y^*_{t,k}) + \log f(y^*_{t,k}|y^*_{t,k-1}) = c + l_{t,1} + l_{t,0}\end{aligned}$$

where $l_{t,j} = -(1/2)\log\{b^2(y^*_{t,k+j-1})\Delta\} - \{2b^2(y^*_{t,k+j-1})\Delta\}^{-1}\{y^*_{t,k+j} - (y^*_{t,k+j-1} + a_{t,j-1}\Delta)\}^2$ and $a_{t,j-1} = a(y^*_{t,k+j-1})$ for $j = 0, 1$. In the previous Section, Gaussian approximations for $l(y^*_{t,k})$ were obtained by approximating $l_{t,1}$ by functions which were quadratic in $y^*_{t,k}$. An alternative approach is to use the second-order Taylor expansion of $l(y^*_{t,k})$ around some point $\hat{y}^*_{t,k}$, here $\hat{y}^*_{t,k} = (y^*_{t,k-1} + y^*_{t,k+1})/2$. This gives a quadratic expression in $y^*_{t,k}$ and thus a Gaussian proposal density.

The derivation of the proposal density is given in Appendix 5.6.1 for a general case, discussed in Section 5, where the sampling block may be multivariate. Applying the formulas (B.5.1)–(B.5.9) given there for univariate blocks ($m = 1$) gives a Gaussian proposal for $y^*_{t,k}$ given $y^*_{t,k-1}$ and $y^*_{t,k+1}$ with variance

$$\tilde{\sigma}^2 = \left\{\frac{1}{\Delta}\left(g_{t,-1} + g_{t,0}c^2_{t,0} - g_{t,0}d_{t,0}a''_{t,0}\Delta - 2g'_{t,0}d_{t,0}c_{t,0} + \frac{1}{2}g''_{t,0}d^2_{t,0} - \frac{\{g_{t,0}g''_{t,0} - (g'_{t,0})^2\}\Delta}{2g^2_{t,0}}\right)\right\}^{-1} \tag{4.6}$$

and mean

$$\tilde{\mu} = \tilde{\sigma}^2 u_t + \hat{y}^*_{t,k} \tag{4.7}$$

where

$$u_t = -\frac{1}{\Delta}\left(g_{t,-1}d_{t,-1} - g_{t,0}d_{t,0}c_{t,0} + \frac{1}{2}g'_{t,0}d^2_{t,0} - \frac{g'_{t,0}\Delta}{2g_{t,0}}\right) \tag{4.8}$$

and the other quantities are defined in equations (B.5.1)–(B.5.4) of Appendix 5.6.1. All the quantities are evaluated at $y^*_{t,k} = \hat{y}^*_{t,k}$.

4.1.4 Expressions for the OU and quadratic drift process

For the OU process, we have

$$\begin{aligned} a_{t,j} &= \mu y^*_{t,k+j},\ a'_{t,j} = \mu,\ a''_{t,j} = 0 \\ g_{t,j} &= \sigma^{-2},\ g'_{t,j} = g''_{t,j} = 0 \\ c_{t,j} &= 1 + \mu\Delta = \rho \\ d_{t,j} &= y^*_{t,k+j+1} - \rho y^*_{t,k+j} \end{aligned}$$

for $j = -1, 0$. Then we obtain

$$\begin{aligned} u_t &= -\frac{1}{\sigma^2\Delta}\{-\rho(y^*_{t,k-1} + y^*_{t,k+1}) + (1+\rho^2)\hat{y}^*_{t,k}\} \\ \tilde{\sigma}^2 &= (1+\rho^2)^{-1}\sigma^2\Delta \\ \tilde{\mu} &= \frac{\rho}{1+\rho^2}(y^*_{t,k-1} + y^*_{t,k+1}) \end{aligned}$$

which is in fact exactly what we obtain in the first-order Taylor approximation of $a(\cdot)$ and for the true conditional distribution of $y^*_{t,k}$ given $y^*_{t,k-1}$, $y^*_{t,k+1}$

For the quadratic drift process, we have

$$\begin{aligned} &a_{t,j} = \mu y^{2*}_{t,k+j},\ a'_{t,j} = 2\mu y^*_{t,k+j},\ a''_{t,j} = 0 \\ &g_{t,j} = \sigma^{-2},\ g'_{t,j} = 0,\ g''_{t,j} = 0 \\ &d_{t,j} = y^*_{t,k+j+1} - (1 + \mu\Delta y^*_{t,k+j})y^*_{t,k+j} \\ &c_{t,j} = -(1 + 2\mu\Delta y^*_{t,k+j}). \end{aligned}$$

Letting $\rho_{t,k+j} = 1 + \mu\Delta y^*_{t,k+j}$ for $j = 0, \ldots n-1$, and applying the formulas of Appendix 5.6.1, we obtain

$$u_t = \frac{1}{\sigma^2\Delta}\hat{y}^*_{t,k} - \rho_{t,k-1}y^*_{t,k-1} + y^*_{t,k+1} + 2\mu\Delta y^*_{t,k+1}\hat{y}^*_{t,k} - (1 + 2\mu\Delta\hat{y}^*_{t,k})\hat{y}^*_{t,k}(1 + \mu\Delta\hat{y}^*_{t,k})$$

and

$$\tilde{\sigma}^2 = \frac{1}{\sigma^2\Delta}(1 + \hat{\rho}^2_{t,k})$$

where $\hat{\rho}_{t,j} = -(1 + 2\mu\Delta\hat{y}^*_{t,k+j})$ The proposal distribution is again Gaussian with mean $\tilde{\mu} = \tilde{\sigma}^2 u_t + \hat{y^*_t}$ and variance matrix $\tilde{\sigma}$.

4.1.5 Examples: the OU process

For the OU process, $a(y^*_{t,k}) = \mu y^*_{t,k}$ and $b(y^*_{t,k}) = \sigma$. The Euler approximation gives

$$y^*_{t,k+1}|y^*_{t,k} \sim N(\rho y^*_{t,k}, \sigma^2\Delta)$$

where $\rho = 1 + \mu\Delta$. Since $a(y^*_{t,k})$ is linear in $y^*_{t,k}$ and $b(y^*_{t,k})$ is a constant, $f(y^*_{t,k+1}|y^*_{t,k})$ is quadratic in $y^*_{t,k}$ and $f(y^*_{t,k}|y^*_{t,k-1}, y^*_{t,k+1})$ is Gaussian. To obtain its parameters, note that

$$\begin{aligned} y^*_{t,k}|y^*_{t,k-1} &\sim N(\rho y^*_{t,k-1}, \sigma^2\Delta) \\ y^*_{t,k+1}|y^*_{t,k} &\sim N(\rho y^*_{t,k}, \sigma^2\Delta). \end{aligned}$$

Conditioning on $y^*_{t,k-1}$, we obtain the mean and variance respectively as,

$$\begin{aligned} \mathrm{E}(y^*_{t,k+1}|y^*_{t,k-1}) &= \rho^2 y^*_{t,k-1} \\ \mathrm{var}(y^*_{t,k+1}|y^*_{t,k-1}) &= (1+\rho^2)\sigma^2\Delta \\ \mathrm{cov}(y^*_{t,k}, y^*_{t,k+1}|y^*_{t,k-1}) &= \rho \mathrm{var}(y^*_{t,k}|y^*_{t,k-1}) = \rho\sigma^2\Delta. \end{aligned}$$

This implies that

$$(y^*_{t,k+1}, y^*_{t,k})|y^*_{t,k-1} \sim N\left[\begin{pmatrix}\rho^2 y^*_{t,k-1} \\ \rho y^*_{t,k-1}\end{pmatrix}, \begin{pmatrix}(1+\rho^2)\sigma^2\Delta & \rho\sigma^2\Delta \\ \rho\sigma^2\Delta & \sigma^2\Delta\end{pmatrix}\right]$$

which in turn implies that

$$y^*_{t,k}|y^*_{t,k-1}, y^*_{t,k+1} \sim N\left\{\frac{\rho}{1+\rho^2}(y^*_{t,k+1} + y^*_{t,k-1}), \frac{\sigma^2\Delta}{1+\rho^2}\right\}.$$

For the OU process we can thus perform Gibbs sampling because $f(y^*_{t,k}|y^*_{t,k-1}, y^*_{t,k+1})$ can be sampled from directly. For the M-H algorithm, the parameters of the proposal distribution based on the zero-order approximation of $a(y^*_{t,k})$ and $b(y^*_{t,k})$ are given by $\tilde{\sigma}^2_{t,0} = (\sigma^2\Delta)/2$ and

$$\tilde{\mu}_{t,0} = \frac{1}{2}\left\{\left(1 - \frac{1}{2}\Delta\mu\right) y^*_{t,k+1} + \left(1 + \frac{1}{2}\Delta\mu\right) y^*_{t,k-1}\right\}.$$

This differs from the true conditional density. We have in effect underapproximated $a(y^*_{t,k})$ by using only the first term of its Taylor expansion. The first-order approximation uses the first two terms of the expansion to obtain another Gaussian proposal density. For the OU process this gives $c_{t,0} = \mu\hat{y}^*_{t,k} - \mu\hat{y}^*_{t,k} = 0$, $c_{t,1} = \mu$ and

$$\begin{aligned} \tilde{\sigma}^2_{t,1} &= \frac{\sigma^2\Delta}{1+\rho^2} \\ \tilde{\mu}_{t,1} &= \frac{\rho}{1+\rho^2}(y^*_{t,k-1} + y^*_{t,k+1}) \end{aligned}$$

which are the parameters of the true conditional distribution. An M-H algorithm with this proposal distribution is thus equivalent to the Gibbs sampler and will always have acceptance rates which are exactly one.

If the strong solution of the process is available, we can use that to obtain a benchmark for the results

from the M-H algorithm. When the strong solution is available we can perform simulation smoothing. We work with the efficient simulation smoother of De Jong and Shephard (1995). It is essentially an MCMC method for Gaussian models. It works in the context of a non-Gaussian state space form model, see Shephard (1994) and Shephard and Pitt (1997). Decomposition of the likelihood is performed in such a way that the resulting expression invokes a Gaussian density as all the nonlinearity has been conditioned out.

For the OU process the strong solution is given by

$$y^*_{t,k} = \exp(\mu\Delta_\tau)y^*_{t,k-1} + \sigma\int_{\tau_{t,k-1}}^{\tau_{t,k}} \exp\{-\mu(s-\tau_{t,k})\}dW_s$$

so that,

$$\begin{aligned} E\{y(\tau_{t+1})|y(\tau_t)\} &= \exp(\mu\Delta_t)y(\tau_t) \\ \text{var}\{y(\tau_{t+1})|y(\tau_t)\} &= \frac{\sigma^2}{2\mu}\{\exp(2\mu\Delta_t)-1\} \\ y(\tau_{t+1})|y(\tau_t) &\sim N[\exp(\mu\Delta_t)y(\tau_t), \frac{\sigma^2}{2\mu}\{\exp(2\mu\Delta_t)-1\}]. \end{aligned}$$

where Δ_τ represents the difference $\tau_{t,k} - \tau_{t,k-1}$ (see appendix 2.7.2).

4.1.6 Examples: the Quadratic drift process

For this process we have, $a(y^*_{t,k}) = \mu y^{2*}_{t,k}$ and $b(y^*_{t,k}) = \sigma$. The Euler approximation gives

$$y^*_{t,k+1}|y^*_{t,k} \sim N(\rho_{t,k}y^*_{t,k}, \sigma^2\Delta)$$

where $\rho_{t,k} = (1+\mu\Delta y^*_{t,k})$. The parameters of the M-H proposal distribution based on zero-order approximation of $a(y^*_{t,k})$ and $b(y^*_{t,k})$ are given by $\tilde{\sigma}^2_{t,0} = (\sigma^2\Delta)/2$ and

$$\tilde{\mu}_{t,0} = \frac{\Delta}{2}\left[\frac{\mu\Delta}{4}(3y^{*2}_{t,k-1} - y^*_{t,k+1}2) - \frac{\mu\Delta}{2}y^*_{t,k+1}y^*_{t,k-1} + y^*_{t,k-1} + y^*_{t,k+1}\right].$$

For the first-order approximation, $c_{t,0} = -\mu\hat{y}^{*2}_{t,k}$ and $c_{t,1} = 2\mu\hat{y}^*_{t,k}$, so

$$\begin{aligned} \tilde{\sigma}^2_{t,1} &= \sigma^2\Delta\{1+(1+\mu\Delta\hat{y}^*_{t,k})^2\}^{-1} \\ \tilde{\mu}_{t,1} &= \frac{\tilde{\sigma}^2_{t,1}}{\sigma^2\Delta}\left[y^*_{t,k-1} + \mu y^{*2}_{t,k-1}\Delta + y^*_{t,k+1} + \mu\Delta(y^*_{t,k-1}+y^*_{t,k+1})y^*_{t,k+1} + \right. \\ &\quad \left. \mu\hat{y}^{*2}_{t,k}\{1+\mu(y^*_{t,k-1}+y^*_{t,k+1})\Delta\}\Delta\right] \end{aligned}$$

4.2 Discussion and comparison

We consider the effect of M, the number of auxiliary states, on the mean of the simulations, the correlograms and efficiency estimates. We then compare the two ad-hoc Taylor approximations of the states and the first-order Taylor approximation of the likelihood for the OU and the quadratic drift process. For the former, the strong solution can be computed and used as a benchmark for comparison.

4.2.1 Effect of M

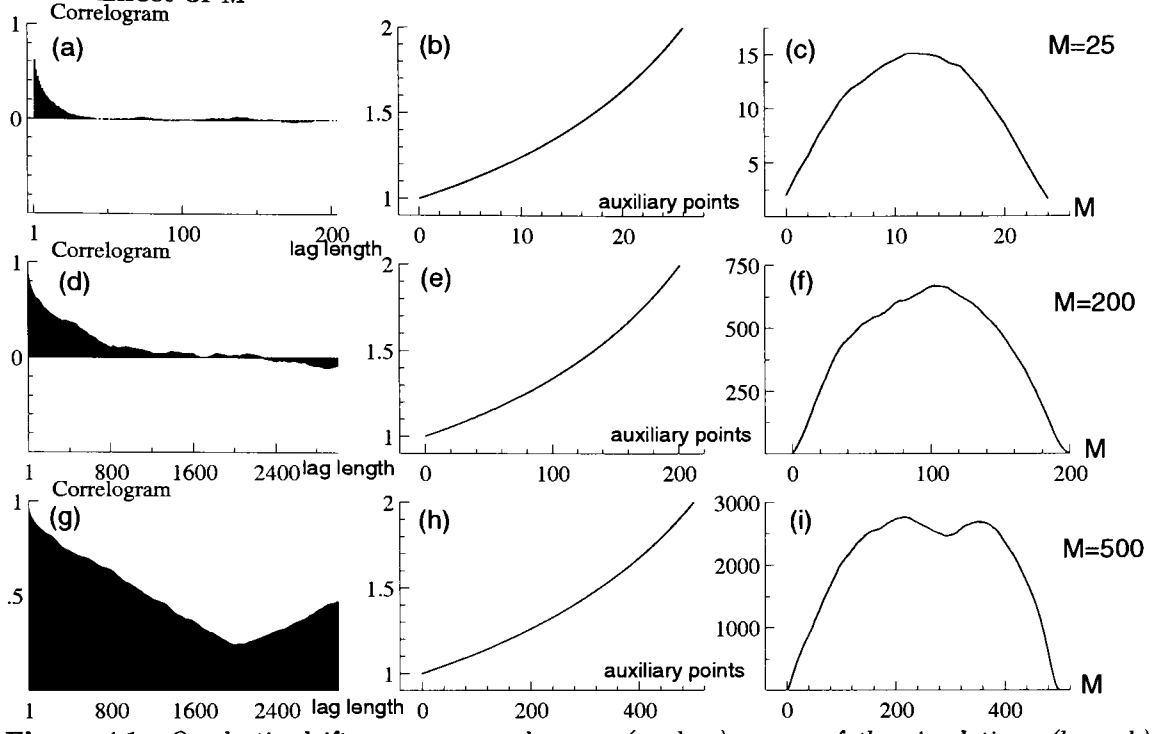


Figure 4.1: *Quadratic drift process: correlograms (a, d, g), mean of the simulations (b, e, h) and inefficiency factors (c, f, i) for the mid-state for* $M = 25$ *(top),* $M = 200$*, (middle) and* $M = 500$ *(bottom), with* $N = 200,000$ *and bandwidth* $B_N = 200$ *for* $M = 25$ *and* $B_N = 3000$ *for* $M = 200$ *and* 500. *A burn-in of* 2000 *iterations is used and every tenth simulation is recorded.*

We first consider the effect of M, the number of simulated missing values, on efficiency of sampling. We perform the M-H algorithm for the quadratic drift process using a first-order Taylor approximation of the states to generate the proposal distribution. The observed values, $y^*_{t,0} = y_t$ and $y^*_{t,M+1} = y_{t+1}$, are set to 1 and 2, respectively, and the parameters of the process are $\sigma^2 = 0.3$ and $\mu = -0.8$. Initial values of the states are drawn from an i.i.d Uniform distribution between $y^*_{t,0}$ and $y^*_{t,M+1}$, that is, they can be expressed as $y^*_{t,0} + (y^*_{t,M+1} - y^*_{t,0})U_j$, where U_j is a Uniform(0, 1) random variable. The acceptance rate is recorded and a plot of the correlogram is obtained. The key here is to observe at which point the autocorrelations die down. One can observe that as M, the number of points between the observations, increases, the correlogram takes longer to die down, see Figure 4.1, which shows correlograms (left hand side) for the state $M/2$ for $M = 25$, 200, and 500.

The mean of the simulations from the M-H algorithm is recorded for each state. The averages are plotted in Figure 4.1, (middle graphs). These were checked against the strong solution of the process, for the OU process. One can observe that as the number of points between the observations increases, the inefficiency factor plots become increasingly less smooth and an increased number of iterations is needed to obtain precise estimates of the mean path. This illustrates the inefficiency in both the Euler scheme and the single-state sampling method. The effect from the initial values is still visible for large M, for $N = 10,000$.

The autocorrelation affects the variance of the sample mean. To assess the inefficiency of the M-H scheme, we estimate the ratio of the variance of the sample mean from the M-H scheme, to that from hypothetical independent samples computed using standard time series methods. These account for the correlation between successive iterates by including (in the estimated variance) a weighted sum of the

estimated autocorrelations up to some pre-specified lag B_N. Here the Parzen window function is used for the weighting, (see Priestley 1981, Chapter 6 and Shephard and Pitt 1997). To get reliable inefficiency factor estimates for cases larger than $M = 200$, the process would have to be run using a larger number of sweeps, implying increased computer time.

The ratio is calculated as follows,

$$\begin{aligned}\hat{R}_N &= 1 + \frac{2N}{N-1}\sum_{i=1}^{B_N} K\left(\frac{i}{B_N}\right)\hat{\rho}(i) \text{ where } \hat{\rho}(i) = \frac{\hat{\Gamma}(i)}{\hat{\Gamma}(0)} \\ \hat{\Gamma}(i) &= \frac{1}{N}\sum_{j=i+1}^{N}(y^j - \bar{y})(y^{j-i} - \bar{y}), \quad \bar{y} = \frac{1}{N}\sum_{j=1}^{N} y^j,\end{aligned}$$

where the Parzen kernel is

$$\begin{aligned}K(x) &= 1 - 6x^2 + 6x^3, && x \in [0, \tfrac{1}{2}], \\ &= 2(1-x)^3, && x \in [\tfrac{1}{2}, 1], \\ &= 0, && \text{elsewhere},\end{aligned}$$

B_N represents band length, chosen as 400 in our example, $\hat{\rho}(i)$ is an estimate of the autocorrelation at lag i, and y represents the states. Plots of the relative efficiency are shown in Figure 4.1 (right hand side), for different values of M. For $M = 25$, the single-move M-H algorithm is up to 50 times less efficient, while for $M = 200$ it is more than 250 times less efficient than a hypothetical sampler which generates independent samples from the process. The statistic $\hat{R}_N$ is informative for MCMC methods; if there exists high correlation, then $\hat{R}_N$ will be large, implying that the information contribution of each additional iteration will be small. This in turn implies that the simulations will have to be run for a longer period of time. In addition, the precision of $\hat{R}_N$ reduces as the correlation increases and so one needs to design low correlation MCMC methods. The inefficiency is more pronounced for auxiliary points far from the endpoints.

4.2.2 Comparing alternative proposals

Figures 4.2 and 4.3 show simulation results for the quadratic process, with $M = 200$ and $N = 20000$. The left hand side of the first graph shows the zero-order approximation of the drift coefficient, while the right hand side depicts the first-order approximation. Acceptance rates, correlograms for the mid-state and inefficiency factor estimates are shown. It can be seen that acceptance rates are higher for the first-order approximation, but the difference is negligible for the quadratic drift function. The second graph shows the average of the simulations for both the zero and first-order approximations, using the same random seed for both. Notice that they are nearly identical.

For the OU process we can actually compare the efficiency of our method to results obtained through smoothing using the strong solution of the process. The smoother has been implemented with the State Space Form Pack by Koopman, Shephard, and Doornik (1999). Figure 4.4 shows the simulation results for the OU process, with $M = 200$ and $N = 20000$ using the M-H algorithm. The first-order approximation is shown to be equivalent to the Gibbs sampler, with acceptance rates of one. The correlograms and inefficiency factor estimates remain similar as does the mean of the simulations.

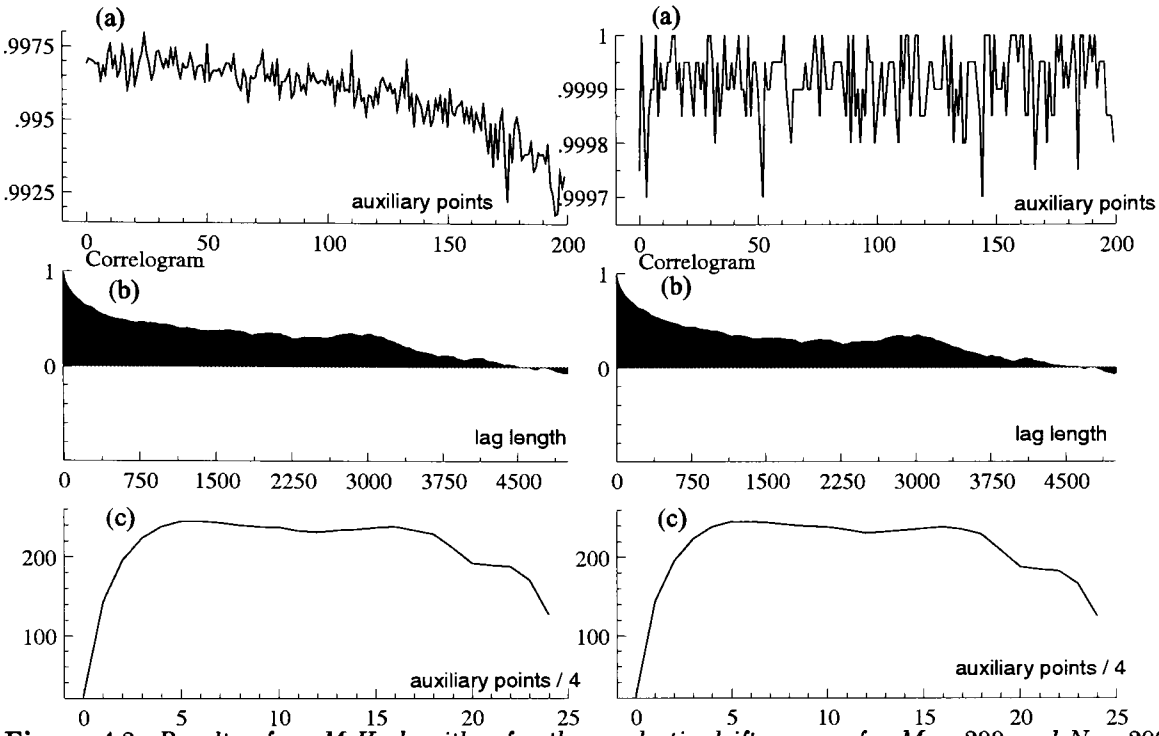


Figure 4.2: *Results of an M-H algorithm for the quadratic drift process for* $M = 200$ *and* $N = 20000$. *The left hand side shows the zero-order approximation of the drift coefficient, while the right hand side depicts the first-order approximation. Acceptance rates (a), correlograms for the mid-state (b) and inefficiency factor estimates are shown (c).*

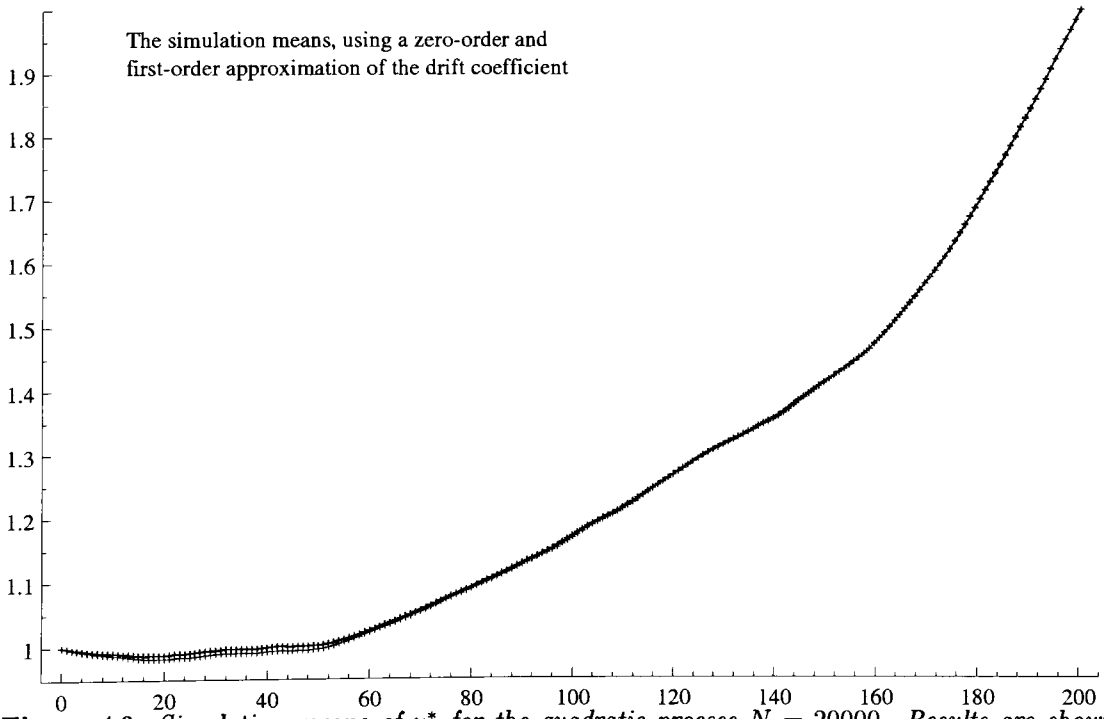


Figure 4.3: *Simulation means of* y^* *for the quadratic process* $N = 20000$. *Results are shown for the zero-order and first-order approximations of the drift coefficient.*

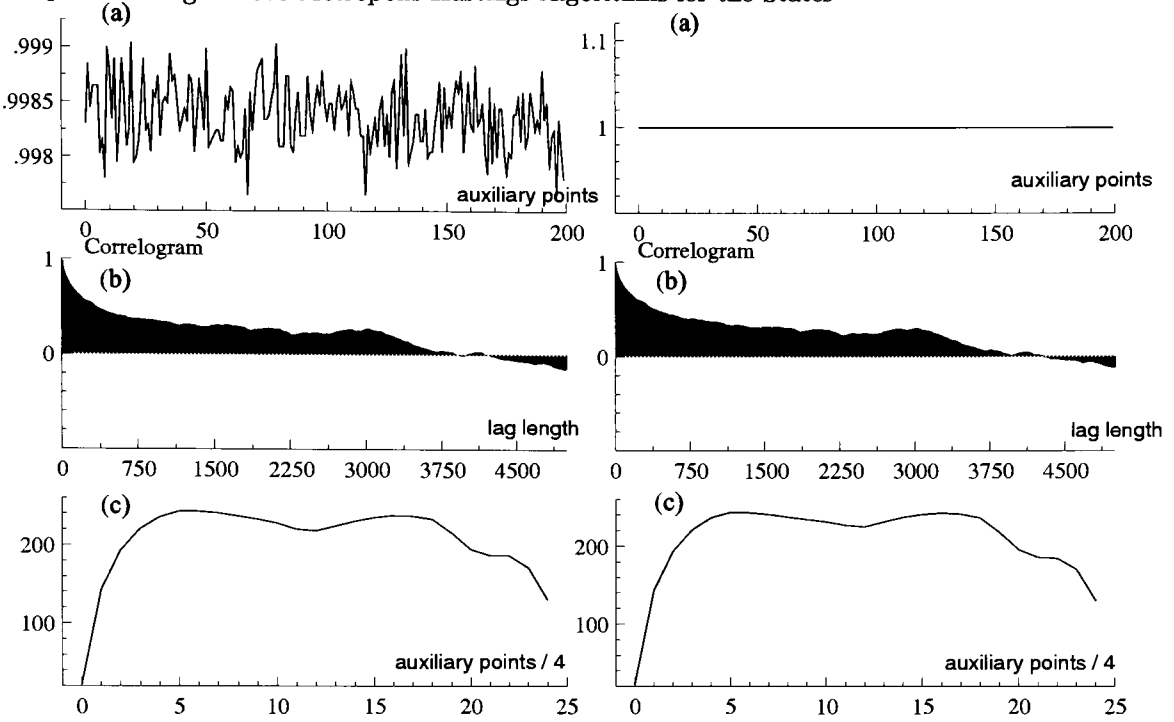


Figure 4.4: *Graphs for the OU process of the acceptance rates (a), correlograms for the mid-state (b) and inefficiency factor estimates (c) are shown for* $M = 200$ *and* $N = 20000$ *for the zero-order approximations (left panel) and the first-order approximation (right panel).*

4.2.3 Assessing the effects of using the Taylor approximation

There is an increase in both efficiency and convergence when a Taylor expansion of the log-likelihood is taken around some $\hat{y}_t^*$ as compared to the case when we employ the zero-order, or first-order approximation with single state sampling, described in Section 4. The difference in means of y can be seen in Figure 4.5, whereas Figure 4.6 shows the acceptance rates, inefficiency factor estimates and correlograms for $M = 25$ with 10000 iterations and $M = 200$ with 20000 iterations. Note that the Taylor approximation seems more efficient in Figure 4.6.

4.3 Analytic convergence rates for the OU process

For some processes, we can obtain closed form expressions for the convergence rate of an MCMC algorithm. This also provides insight into the convergence rates in cases where such closed form results are not available. For the OU case, we can obtain convergence rates for the Gibbs sampler. Essentially, we require an analytic expression for the relationship between the convergence rate of the Gibbs sampler and the parameters of the OU process. For both the M-H and the Gibbs sampler, we can compare the second largest eigenvalue in absolute value of the transition matrix, as it describes how close the sample is to the equilibrium distribution. Note however, that the transition matrix for the M-H is more complex. Frigessi, Di-Stefano, Hwang, and Sheu (1993) describe one updating dynamics model to be "better than another if the second eigenvalue in absolute value of its transition matrix is smaller than that of the other model".

To analyse the convergence rate of the single-move Gibbs sampler, the methodology of Roberts and Sahu (1997) is followed. A general framework is also outlined in Pitt and Shephard (1999), where strategies

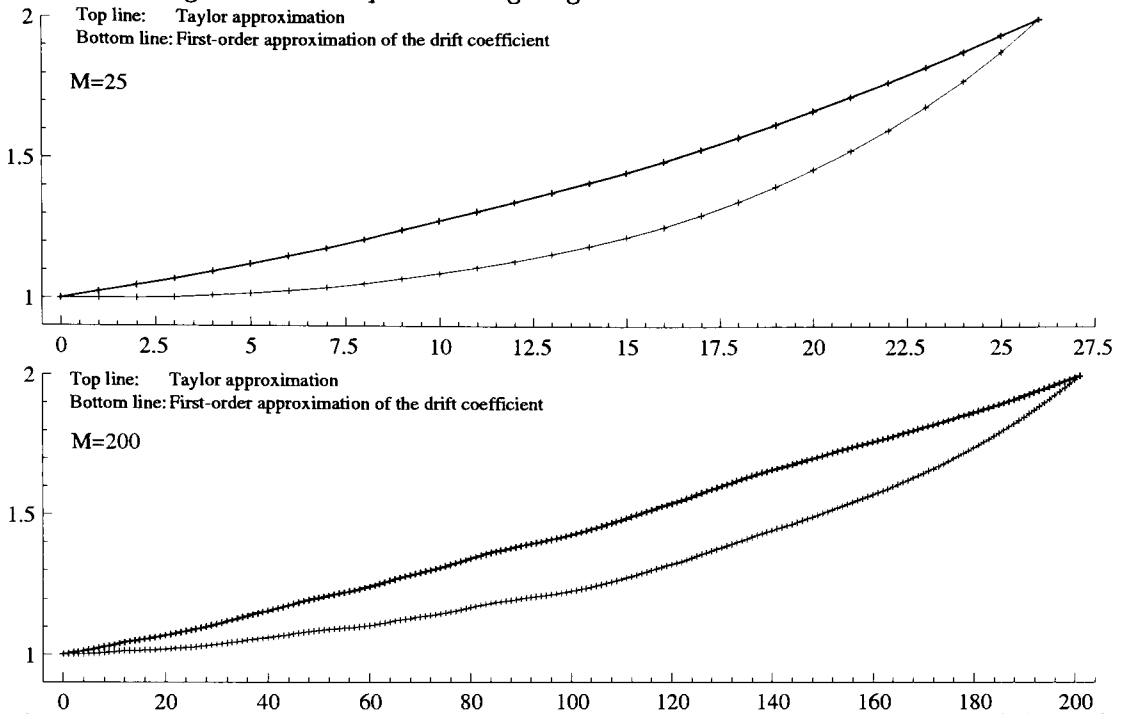


Figure 4.5: *The mean of the simulations are shown using a Taylor approximation of the log-likelihood (top lines) vs. a first-order ('ad-hoc') approximation of the drift coefficient (bottom lines) for the quadratic process for a block size of 1, for* $M = 25$ *and* $M = 200$.

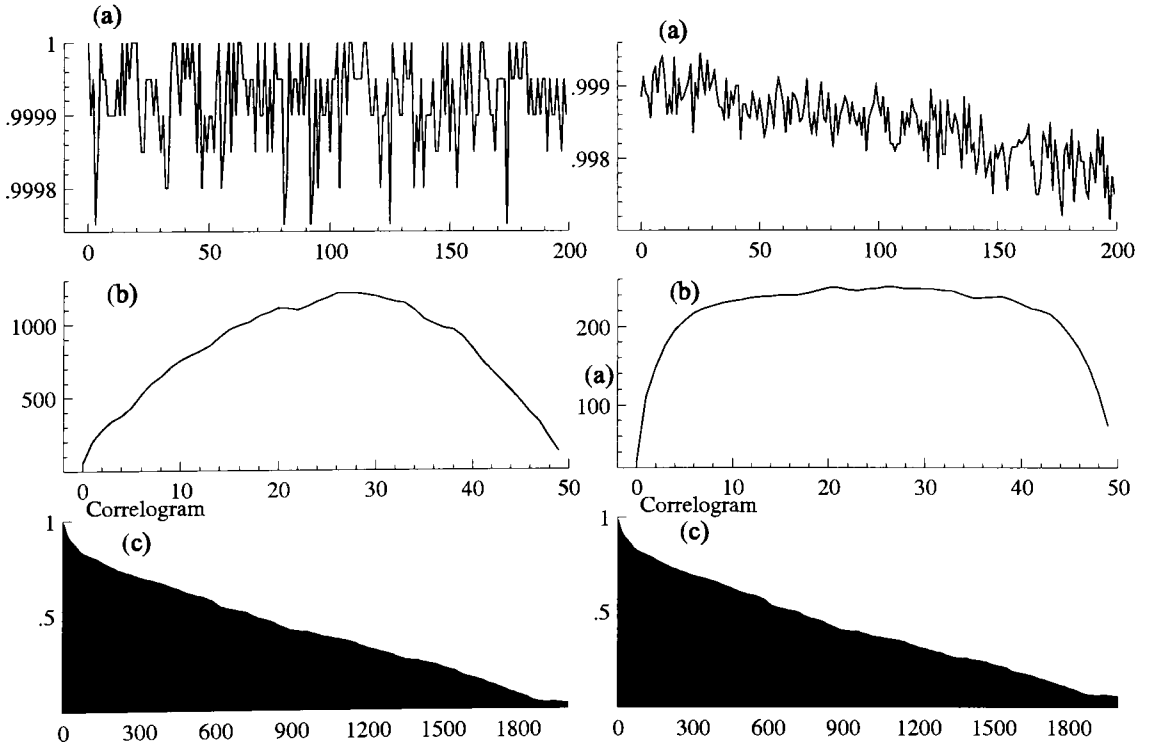


Figure 4.6: *The average acceptance rates (a), inefficiency factor estimates (b) and correlograms (c) are shown, using a first-order ('ad-hoc') approximation of the drift coefficient (left panel) vs. a Taylor approximation of the log-likelihood (right panel) for the OU process. The block size used is 1 and the number of imputed points are given by* $M = 200$.

of centring the parameterization are also explored. A derivation of the results is presented in Appendix 4.5.1. It is essentially the same as that presented in Pitt and Shephard (1999).

Now, employing the Euler scheme, the OU process can be expressed as,

$$\begin{aligned} y_{t+1} &= y_t + \mu\Delta y_t + \sigma dW_t \\ &= \rho y_t + \sigma dW_t \end{aligned}$$

where $\rho = 1 + \mu\Delta$. Hence the log density for the auxiliary points is

$$\begin{aligned} \log f(y^*_{t,0},...,y^*_{t,M+1}) &= c - \frac{1}{2\sigma^2\Delta}(y^*_{t,1} - \rho y^*_{t,0})^2 - \frac{1}{2\sigma^2\Delta}(y^*_{t,2} - \rho y^*_{t,1})^2 - \cdots - \frac{1}{2\sigma^2\Delta}(y^*_{t,M+1} - \rho y^*_{t,M})^2 \\ &= d - \frac{1}{2}y_t^{*\prime}V_t^{-1}y_t^*, \end{aligned}$$

where c, d are constants, and $y_t^* = (y^*_{t,1},\ldots,y^*_{t,M})$. Since the endpoints are fixed, $y^{*2}_{t,0}, y^{*2}_{t,M+1}$ are constants and $2\rho y^*_{t,0}y^*_{t,1}$, $2\rho y^*_{t,M}y^*_{t,M+1}$ only affect the mean and can therefore be ignored when analysing the convergence rate. Suppose we let $U_M(x) = \frac{sin[(M+1)\cos^{-1}(x)]}{\sin[\cos^{-1}(x)]}$ as in Appendix 4.5.1. We obtain the following solution for the eigenvalues, λ, where $a = \rho/(1+\rho^2)$,

$$\lambda = 2a\cos(\frac{k\pi}{M+1}) < 4a^2, \quad \text{where } k = 1,...,M.$$

The equation has been solved via Chebychev polynomials - for a more comprehensive explanation, see Appendix 4.5.1.

The convergence rate τ is defined through

$$4\cos^2(\frac{\pi}{M+1})a^2 < \tau < 4a^2.$$

Hence, for this example, we can obtain a closed form expression for the convergence rate

$$\lim_{M\to\infty}\tau = \lim_{M\to\infty}\frac{4\rho^2}{(1+\rho^2)^2} = 1.$$

Now as $\rho \to 1, \tau \to 1$, and the inefficiency factor $\frac{1}{1-\tau} \to \infty$. Hence, for a persistent parameterization, a value of $\tau = 1$ indicates slow convergence. A plot of the upper-bounds $4a^2$ against $\log(M)$ is shown in Figure 4.7. The Integrated Autocorrelation Time, (IACT), is $\frac{1+\tau}{1-\tau}$, where,

$$\begin{aligned} \frac{1+\tau}{1-\tau} &= \frac{(\rho^2+1)^2 + 4\rho^2}{(\rho^2+1)^2 - 4\rho^2} \\ &= \frac{(\rho^2+1)^2 + 4\rho^2}{(\rho^2-1)^2}. \end{aligned}$$

In the Gaussian case, τ doesn't depend on $y^*_{t,0}, y^*_{t,M+1}$ and σ^2; only the covariance structure matters, that is, $\rho = \rho(\mu, M)$.

4.4 Conclusion

The single-move M-H algorithm has been discussed in this Chapter. Section 4.1 considered use of a zero-order and first-order Taylor approximation of the states, which was compared to use of a Taylor

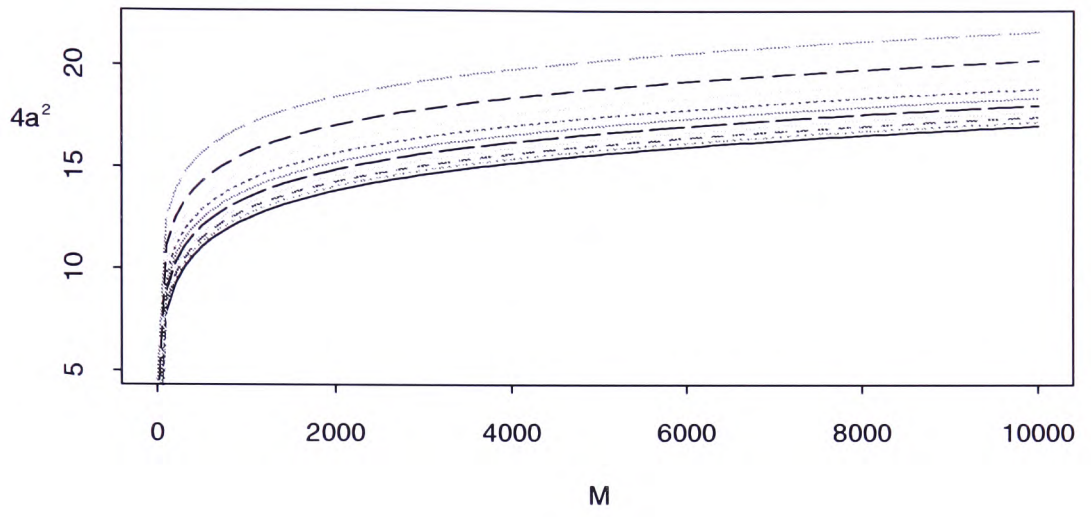


Figure 4.7: *A logit transformation of the upperbound of the convergence rate, given by* $\tau = 4a^2$, *is graphed for different values of* μ *and discretisation points,* M.

approximation of the likelihood in Section 4.2. An expression for the convergence rate of the MCMC algorithm was shown to be available for the OU process in Section 4.3. The transition matrix for the M-H algorithm is more complex, but an approximate convergence rate can be inferred using the method discussed above.

Problems with the single-move algorithm are slow convergence, particularly when M is large, which results in inefficiency in estimating the posterior moments when the chain has reached equilibrium. Chapter 5 describes a more efficient M-H sampler, which updates a block of elements of y_t^* at once.

4.5 Appendix

4.5.1 The convergence rate

Letting $a = \frac{\rho}{1+\rho^2}$, we have

$$A = I - diag(V_{t,11}^{-1}, ..., V_{t,MM}^{-1})V_t = \begin{pmatrix} 0 & a & 0 & \cdots & 0 \\ a & 0 & a & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & a & 0 & a \\ 0 & \cdots & 0 & a & 0 \end{pmatrix},$$

where the dimension of A is $M \times M$. Suppose that we let

$$D_M \equiv \det(A - \lambda I) = \begin{vmatrix} -\lambda & a & 0 & \cdots & 0 \\ a & -\lambda & a & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & a & -\lambda & a \\ 0 & \cdots & 0 & a & -\lambda \end{vmatrix},$$

so that in general we have,

$$D_{M+2} + \lambda D_{M+1} + a^2 D_M = 0$$

subject to the initial conditions $D_0 = 1$, $D_1 = -\lambda$ and $D_2 = \lambda^2 - a^2$. This equation can be solved via Chebychev polynomials. Let us define $A(t) = \sum_{M=0}^{\infty} D_M t^M$ and sum the recursion equation as follows

$$\sum_{M=0}^{\infty} t^{M+2} D_{M+2} + \lambda \sum_{M=0}^{\infty} t^{M+2} D_{M+1} + a^2 \sum_{M=0}^{\infty} t^{M+2} D_M = 0.$$

Then we obtain

$$A(t)(1 + \lambda t + a^2 t^2) = D_0 + t D_1 + \lambda t D_0 = 1,$$

implying

$$A(t) = \frac{1}{(1 + \lambda t + a^2 t^2)}.$$

Now we want to find values of T and x such that $1 + \lambda t + a^2 t^2 = 1 - 2Tx + T^2$. The values which satisfy this are $T = -ta$, and $x = \lambda/2a$. Now

$$A(t) = \frac{1}{(1 - 2Tx + T^2)} = \sum_{M=0}^{\infty} T^M U_M(x)$$

where $U_s(x)$ is the sth Chebyshev polynomial of the second kind defined as

$$U_s(x) = \frac{\sin\{(s+1)\cos^{-1}(x)\}}{\sin\{\cos^{-1}(x)\}}.$$

Hence we obtain

$$A(t) = \sum_{M=0}^{\infty} (-1)^M a^M t^M U_M\left(\frac{\lambda}{2a}\right).$$

By comparing coefficients of t^M to those of $A(t)$, we obtain

$$D_M = (-1)^M a^M U_M\left(\frac{\lambda}{2a}\right).$$

The M solutions of this characteristic equation satisfy $\sin\left\{(M+1)\cos^{-1}(\frac{\lambda}{2a})\right\} = 0$. Hence the M distinct eigenvalues are given by

$$\lambda = 2a\cos\left(\frac{k\pi}{M+1}\right), \text{ where } k = 1, 2, ..., M.$$

Clearly the largest eigenvalue is $\lambda_1 = 2a\cos\left(\frac{\pi}{M+1}\right)$. Since $D_M = -\lambda\Lambda_{M-1} - ab\Lambda_{M-2} = \Lambda_M - a^2\varepsilon\Lambda_{M-2}$, where $b = (1+\varepsilon)a$, $\varepsilon = \rho^2$ we have that the solution of $D_M(\lambda) = 0$ satisfies

$$f(\theta) = \sin(M+1)\theta - \varepsilon\sin(M-1)\theta = 0 \tag{B.4.1}$$

where $\theta = \cos^{-1}(\lambda/2a)$ and $0 < \theta < \pi$. It can be seen that the largest eigenvalue λ_1, say, corresponds to the smallest solution θ_1 of $f(\theta)$. We now wish to show that this root lies in the interval $[0, \frac{\pi}{M+1}]$. When $\theta = 0$, $f(0) = 0$ and

$$\begin{aligned} f'(\theta) &= (M+1)\cos(M+1)\theta - \varepsilon(M-1)\cos(M-1)\theta \\ \therefore f'(0) &> 0 \text{ if } \varepsilon < \frac{M+2}{M}. \end{aligned}$$

Hence $\forall\ \delta > 0$, $\exists\ 0 < \epsilon < \delta$, such that $f(\epsilon) > 0$. This demonstrates that $f(\theta)$ is positive immediately after $\theta = 0$. Now at $\theta = \frac{\pi}{M+1}$,

$$f(\theta) = -\varepsilon\sin\left\{\frac{(M-1)\pi}{M+1}\right\} < 0 \qquad \text{since } \varepsilon > 0.$$

So the smallest root θ_1, say, of $f(\theta)$ satisfies

$$0 < \theta_1 < \frac{\pi}{M+1}.$$

Since the largest root of D_M is $\lambda_1 = 2a\cos\theta_1$ we obtain

$$2a\cos\left(\frac{\pi}{M+1}\right) < \lambda_1 < 2a.$$

Hence $\tau(B) = \lambda_1^2$ satisfies

$$4a^2\cos^2\left(\frac{\pi}{M+1}\right) < \tau(B) < 4a^2.$$

Hence as $M \longrightarrow \infty$,

$$\tau(B) \longrightarrow 4a^2 = 4\frac{\rho^2}{(1+\rho^2)}.$$

Also as $\rho \longrightarrow 1$, $M \longrightarrow \infty$, then $\tau(B) \longrightarrow 1$.

CHAPTER 5

BLOCK SAMPLING FOR THE AUXILIARY STATES

The main problems with the single-move algorithm are slow convergence, particularly when M is large, and inefficiency in estimating the posterior moments when the chain has reached equilibrium. This can be explained by the fact that the auxiliary states in y_t^* are close to each other, which inevitably means that they are strongly correlated.[1] In Section 5.1, we describe a M-H sampler which updates a block of m elements of y_t^* at once, where $1 \leq m \leq M$. The proposal density is constructed in Section 5.1.1 and can either be updated in fixed or random block sizes within each sweep. The sampling method for the blocks is still based around the fixed neighbours / end points and initial values. For example, suppose the first block is of size 2 and the second of size 3. The states $y_{t,1}^{*(j+1)}$ and $y_{t,2}^{*(j+1)}$ are sampled conditioning on $y_{t,0}^*$ and $y_{t,3}^{*(j)}$; the states $y_{t,3}^{*(j+1)}, y_{t,4}^{*(j+1)}$ and $y_{t,5}^{*(j+1)}$ are then sampled conditioning on $y_{t,2}^{*(j+1)}$ and $y_{t,6}^{*(j)}$, and so on. Examples and a comparison between blocking strategies is presented in Section 5.2. For tractable models such as the OU process, it is possible to conduct a theoretical analysis of the efficiency of the sampler. This is addressed in Section 5.4. For most models, (examples of which are considered in Section 5.2), this is not possible and a method for computing numerical convergence rates is suggested in Section 5.4.2. The Chapter concludes with Section 5.5.

Previous work in this area has included a paper by Liu, Wong, and Kong (1994), where comparisons of various augmentation schemes are made by studying the covariance structure of the samples generated by the Gibbs algorithm. For dependent samples, they show that collapsing and grouping random components in the Gibbs sampler, with two or three components, usually results in a more efficient sampling scheme and produces estimates which have smaller variance. Applications include Shephard (1994) and Carter and Kohn (1994), where all states are generated at once using an errors mixture of normals. Knorr-Held (1996), Gamerman and Smith (1996) and Shephard and Pitt (1997) give examples of blocking for time series problems. Work by Shephard and Pitt (1997) indicates that the chain converges faster if the block size is selected randomly at each updating stage. The Poisson distribution is used as the randomisation device in their work, but alternatives can be used; these are discussed by Wong (1999) in the context of stochastic volatility models.

[1] Other work in this area has included a paper by Eraker (1998), where single-move sampling in the context of Brownian motion is used within a continuous time stochastic volatility model. He has found, using a discretised approach, that the chain converges slowly as has been shown above, in Figures 4.2 and 4.4.

5.1 A BLOCK METROPOLIS–HASTINGS SAMPLER

The estimation method proposed relies on Markov chain Monte Carlo methods (Tanner and Wong 1987, Geweke 1996, Gelfand and Smith 1990, Chib and Greenberg 1995 and Chib and Greenberg 1996), an outline of which was provided in Section 3.2.

In brief, the aim is to sample the Bayesian posterior density $\pi(\theta, Y^*|Y)$, where $Y = (y_1, \ldots, y_T)'$, by sampling $y_t^*|y_t, y_{t+1}; \theta$, for $t = 1, ..., T$, then sampling $\theta|Y, Y^*$ and repeating this process until the chain has converged to its stationary distribution. In particular, the method is based on the M-H algorithm, using independence samplers, (see for example Chib and Greenberg 1995 and Tierney 1994, and the outline in Section 3.3), where the proposal does not depend on the current value of the block to be sampled. However, because the mean of the proposal depends on the immediate neighbours of the block and because the underlying process for the observations is continuous, the mean of the proposal is typically close to the current value of the block. This has the important consequence that the algorithms proposed here will normally have high acceptance probabilities and are very unlikely to get stuck in the tails of the target distribution.

As outlined in Section 3.2, given the dimensionality of the distribution $y_t^*|y_t, y_{t+1}; \theta$, this sampling must be conducted with care, in order to produce a Markov chain in which the serial correlations vanish rapidly. It is important to construct algorithms that minimise the serial dependence between the draws, so that the mixing properties of the chain are maximised and all parts of the support of the target density can be reached.

5.1.1 The proposal density

The target density of interest is given by,

$$f(y_t^*|y_t, y_{t+1}) \propto \prod_{k=0}^{M} f(y_{t,k+1}^*|y_{t,k}^*), \tag{5.1}$$

where each conditional density in this expression can be derived from the discrete-time Euler approximation, given in equation (2.1). We have seen in Section 4, that in general, $y_{t,k}^*$ appears nonlinearly in the drift and volatility functions so that it is not possible to sample $y_{t,k}^*$ directly from the target density. The target density can, however, be sampled by the M-H algorithm with proposal density $q(y_t^*|y_t, y_{t+1}; \theta)$. The proposal density to approximate $f(y_t^*|y_t, y_{t+1}; \theta)$ is chosen so that it is easy to sample from.

Consider a block of size $m\,(1 \leq m \leq M)$ starting with observation $y_{t,k}^*$ $(1 \leq k \leq M)$. So we have a typical block $y_{t,(k,m)}^* = (y_{t,k}^*, y_{t,k+1}^*, \cdots, y_{t,k+m-1}^*)$ with neighbours $y_{t,k-1}^*$ and $y_{t,k+m}^*$. Denote this block as $y_{t,(k,m)}^* = (w_1, \ldots, w_m) = w$, with $w_0 = y_{t,k-1}^*$ and $w_{m+1} = y_{t,k+m}^*$. Using an appropriate discretisation of the SDE, the approximate log likelihood function for w given its neighbours is

$$\begin{aligned} l(w) &= \log f(w|w_0, w_{m+1}; \theta) && (5.2) \\ &= c + \sum_{j=0}^{m} \log f(w_{j+1}|w_j; \theta) = c + \sum_{j=0}^{m} l_j && (5.3) \end{aligned}$$

for $j = 0, \ldots, m$. For the Euler approximation

$$l_j = \log f(w_{j+1}|w_j) = -(1/2)\log\{b^2(w_j)\Delta\} - [2b^2(w_j)\Delta]^{-1}[w_{j+1} - \{w_j + a(w_j)\Delta\}]^2.$$

The target density can then be approximated by a multivariate$-t$ distribution or a multivariate-normal distribution, based on a second-order Taylor expansion of $l(w)$, performed around $\hat{w}$. Applications can be found in Chib and Greenberg (1994), Chib, Greenberg, and Winkelmann (1998) and Shephard and Pitt (1997).

Details of this expansion for the Euler discretisation are given in Appendix 5.6.1. An initial $\hat{w}$ is chosen as $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_m)$, where the $(j+1)$st element is given by $\hat{w}_{j+1} = w_0 + [(j+1)/(m+1)](w_{m+1} - w_0)$, (for $j = -1, 0, \ldots, M$). These are then iterated a few steps towards the mode of the density using a Newton-Raphson scheme, as outlined in Section 5.1.2. Convergence to the mode typically occurs quickly for these models and particularly in the applications considered in the thesis. For many models, a proposal with good mixing properties is given by considering a Gaussian approximation for $l(w)$. The second-order Taylor expansion gives a multivariate-Gaussian proposal with mean $\tilde{\mu}$ and variance matrix $\tilde{\Sigma}$, which are defined by equations (B.5.1)–(B.5.9) of Appendix 5.6.1.

A candidate value z is then sampled from

$$q(w) = q(w|w_0, w_{m+1}; \theta) \sim N(\tilde{\mu}, \tilde{\Sigma}),$$

where $\tilde{\mu}$ and $\tilde{\Sigma}$ are functions, (through $\hat{w}$), of w_0, w_{m+1} and θ, but not of $w^{(j)}$, the old value of w. The acceptance probability is now

$$\alpha\left\{z, w^{(j)}\right\} = \min\left[1, \frac{f\left\{z|w_0^{(j+1)}, w_{m+1}^{(j)}; \theta\right\} q\left\{w^{(j)}\right\}}{f\left\{w^{(j)}|w_0^{(j+1)}, w_{m+1}^{(j)}; \theta\right\} q(z)}\right].$$

For other models, it may be advisable to replace the $N(\tilde{\mu}, \tilde{\Sigma})$ by a heavier tailed distribution, such as the multivariate$-t$ density.

5.1.2 Iterating the expansion point to the mode of the density

We propose to use a Newton-Raphson scheme based on the approximate log-likelihood as a general method. We know that

$$\begin{aligned} l(\tilde{w}) &= c - \frac{1}{2}(w - \hat{w} - V_t^{-1}u_t)'V_t(w - \hat{w} - V_t^{-1}u_t) \\ \left.\frac{\partial l(\tilde{w})}{\partial w}\right|_{\hat{w}} &= -V_t(w - \hat{w} - V_t^{-1}u_t) \\ \left.\frac{\partial l(\tilde{w})}{\partial w \partial w'}\right|_{\hat{w}} &= -V_t \end{aligned}$$

and that $-V_t < 0$ for a maximum. Hence we could use a scheme of the following form

$$\hat{w}^\dagger = \hat{w} + \left(\left.\frac{\partial^2 l(\tilde{w})}{\partial w \partial w'}\right|_{\hat{w}}\right)^{-1} \left.\frac{\partial l(\hat{w})}{\partial w}\right|_{\hat{w}}$$

set $\hat{w} = \hat{w}^\dagger$ and iterate the algorithm until convergence, that is

$$|\hat{w}^\dagger - \hat{w}| < \delta$$

for small positive δ, or until V_t is positive definite. We can perform the algorithm for each block as we proceed through the iterations.

We could also obtain the overall mode of $l(\tilde{w})$ as an initial starting value for w in the M-H iterations. For SSF models for example, an efficient Newton-Raphson procedure would be to first simulate the states given the neighbours and the $\hat{w}$'s, using the simulation smoother of De Jong and Shephard (1995), through use of the matrices stored from the Kalman Filter. The following step would be to expand $l(\tilde{w})$ around $\hat{w}$, set $\hat{w}$ to the mean of the resulting expectation smoother and expand $\hat{w}$ and continuing in this fashion. The expansion is performed for each iteration of the proposed MCMC sampler. In our case, the transition and measurement equations are equivalent and we already possess analytic expressions for the first and second derivatives, allowing the Newton-Raphson scheme described above to be computed easily. It is important to note that the algorithm should check for V_t being positive definite in the process of obtaining the expansion point $\hat{w}$. Otherwise, the scheme may progress towards a minimum instead of the maximum; this issue is addressed in Section 5.3.

5.1.3 Selecting the block size

Initially, the algorithm was run over various fixed block sizes. The improvement was considerable as compared to the single-state sampling scheme. The results become even more striking for random block sizes. The selection of block sizes to sample as the MCMC method runs through one sweep is carried out randomly and independently of the process using a Poisson random variable. The choice of the specific distribution is ad-hoc and serves simply as a randomisation device.

If m is too large, the sampler may be slow, as generating candidate values from high dimensional multivariate normals is difficult and could lead to rejection. Updating all M augmented variables at once is not generally advisable, as the sampler may then run into problems when the diffusion is very non-linear or when M is large. In such cases, the proposal density may not be able to accurately approximate $f(y_t^*|y_t, y_{t+1}, \theta)$. On the other hand, if m is chosen too small, correlation will be high for the reasons explained above. A more effective general approach is to use a M-H step within a M-H algorithm by updating in sequence contiguous subsets, or blocks, increasing the block size as M increases.

Suppose we choose the block size to be selected by a Poisson distribution where λ denotes the average block size, or *tuning parameter*. That is, the block size is greater than or equal to 1 and has mean λ. The block sizes, m_i are then generated so that all M points in the interval are covered, (the last block size has to be chosen appropriately). Hence each sample point is updated in every sweep. The performance of the sampler is affected by the choice of the block size, m. The algorithm for selecting the block sizes could take the following form,

1. select λ and M,
2. choose $m_1 \sim 1+$ Poisson$(\lambda - 1)$: if $m_1 \geq M$ set $m_1 = M$ and stop, otherwise set $i = 2$ and
3. choose $m_i \sim 1+$ Poisson$(\lambda - 1)$, $\begin{array}{l} \text{if } \sum_{j=1}^{i} m_j \geq M \text{ set } m_i = M - \sum_{j=1}^{i-1} m_j \text{ and stop,} \\ \text{if } \sum_{j=1}^{i} m_j < M \text{ set } i = i + 1 \text{ and repeat.} \end{array}$

To summarise and in the context of the whole procedure, the following algorithm, indexed by a tuning parameter $\lambda \geq 1$ is advocated:

1. Initialise Y^*, θ.

2. Set $k = 0$.
3. Draw $m \sim \text{Poisson}(\lambda) - 1$; set $k = \min(m, M)$.
4. Update $y^*_{t(k,m)} | y^*_{t,k-1}, y^*_{t,k+m}, \theta$.
   (a) If $k < M$ goto 3.
   (b) Else update θ from $\theta | Y, Y^*$ and then goto 2.

5.2 EXAMPLES

The OU, quadratic drift process and Vasicek model are considered in Sections 5.2.1–5.2.3. Summary plots for each process including acceptance rates, estimates of the efficiency of the sampler and correlograms are shown for various fixed and random block sizes. The algorithm is shown to be more efficient for the random block case, especially when M is large.

5.2.1 The OU process

For the OU process

$$\begin{aligned}
&a_{t,j} = \mu y^*_{t,k+j}, \;\; a'_{t,j} = \mu, \;\; a''_{t,j} = 0 \\
&g_{t,j} = \sigma^{-2}, \;\; g'_{t,j} = 0, \;\; g''_{t,j} = 0 \\
&d_{t,j} = y^*_{t,k+j+1} - (1 + \mu\Delta) y^*_{t,k+j} = y^*_{t,k+j+1} - \rho y^*_{t,k+j} \\
&c_j = \rho
\end{aligned}$$

and the formulas of Appendix 5.6.1 give

$$u_t = -\frac{1}{\sigma^2 \Delta} \begin{pmatrix} \hat{y}^*_{t,k} - \rho y^*_{t,k-1} - (\hat{y}^*_{t,k+1} - \rho \hat{y}^*_{t,k})\rho \\ \vdots \\ \hat{y}^*_{t,k+m-1} - \rho \hat{y}^*_{t,k+m-2} - (y_{t,k+m} - \rho \hat{y}^*_{t,k+m-1})\rho \end{pmatrix}$$

and

$$V_t = \frac{1}{\sigma^2 \Delta} \begin{pmatrix} 1+\rho^2 & -\rho & & \\ -\rho & 1+\rho^2 & \ddots & \\ & \ddots & \ddots & \ddots \\ & & & -\rho \\ & & -\rho & 1+\rho^2 \end{pmatrix}. \tag{5.4}$$

The proposal distribution is again Gaussian with mean $\tilde{\mu} = \tilde{\Sigma} u_t + \hat{y}^*_t$ and variance matrix $\tilde{\Sigma} = V_t^{-1}$. This is in fact the exact conditional density of y^*_t given $y^*_{t,k-1}$ and $y^*_{t,k+m}$. This can be seen by writing

the log-likelihood as

$$\begin{aligned}
\log f(y_t^*|y_{t,k-1}^*, y_{t,k+m}^*) &= c + \sum_{j=0}^{m} \log f(y_{t,k+j}^*|y_{t,k+j-1}^*) \\
&= c - \frac{1}{2}\sum_{j=0}^{m} \frac{1}{\sigma^2\Delta}(y_{t,k+j}^* - \rho y_{t,k+j-1})^{*2} \\
&= c' - \frac{1}{2}(y_t^{*\prime} V y_t^* - 2y_t^{*\prime} u_*) \\
&= c'' - \frac{1}{2}(y_t^* - V^{-1}u_*)' V (y_t^* - V^{-1}u_*)
\end{aligned}$$

where V_t is given by (5.4) and

$$u_{t*} = (\rho y_{t,k-1}^*, 0, \ldots, 0, \rho y_{t,k+m}^*)'.$$

The conditional density is thus

$$y_t^*|y_{t,k-1}^*, y_{t,k+m}^* \sim N(\tilde{\mu}_*, \tilde{\Sigma})$$

where $\tilde{\Sigma} = V_t^{-1}$ and $\tilde{\mu}_* = \tilde{\Sigma} u_{t*}$. It is easy to show that $\tilde{\mu}_* = \tilde{\mu}$.

5.2.2 The quadratic drift process

For the quadratic drift process, we have

$$\begin{aligned}
&a_{t,j} = \mu y_{t,k+j}^{*2}, \quad a'_{t,j} = 2\mu y_{t,k+j}^*, \quad a''_{t,j} = 0 \\
&g_{t,j} = \sigma^{-2}, \quad g'_{t,j} = 0, \quad g''_{t,j} = 0 \\
&d_{t,j} = y_{t,k+j+1}^* - (1 + \mu\Delta y_{t,k+j}^*) y_{t,k+j}^* \\
&c_{t,j} = -(1 + 2\mu\Delta y_{t,k+j}^*).
\end{aligned}$$

Letting $\rho_{t,j} = 1 + \mu\Delta y_{t,k+j}^*$ for $j = 0, \ldots m-1$, and applying the formulas of Appendix 5.6.1, we obtain

$$u_t = \frac{1}{\sigma^2\Delta}\begin{Bmatrix} \hat{y}_{t,k}^* - \rho_{t,-1} y_{t,k-1}^* + \hat{y}_{t,k+1}^* + 2\mu\Delta\hat{y}_{t,k+1}\hat{y}_{t,k}^* - (1 + 2\mu\Delta\hat{y}_{t,k}^*)\hat{y}_{t,k}^*(1 + \mu\Delta\hat{y}_{t,k}^*) \\ \vdots \\ \hat{y}_{t,k+m-1}^* - \rho_{t,m-2}\hat{y}_{t,k+m-2}^* + y_{t,k+m}^* + 2\mu\Delta\hat{y}_{t,k+m-1}^* y_{t,k+m}^* \\ -(1 + 2\mu\Delta\hat{y}_{t,k+m-1}^*)\hat{y}_{t,k+m-1}^*(1 + \mu\Delta\hat{y}_{t,k+m-1}^*) \end{Bmatrix}$$

and

$$V_t = \frac{1}{\sigma^2\Delta}\begin{pmatrix} 1 + \hat{\rho}_{t,0}^2 & -\hat{c}_{t,0} & & & \\ -\hat{c}_{t,0} & 1 + \hat{\rho}_{t,1}^2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & & & \hat{c}_{t,m-2} \\ & & & -\hat{c}_{t,m-2} & 1 + \hat{\rho}_{t,m-1}^2 \end{pmatrix}, \tag{5.5}$$

where $\hat{c}_{t,j} = -(1 + 2\mu\Delta\hat{y}_{t,k+j}^*)$. The proposal distribution is again Gaussian with mean $\tilde{\mu} = \tilde{\Sigma} u_t + \hat{y}_t^*$ and variance matrix $\tilde{\Sigma} = V_t^{-1}$. Figures 5.1–5.3 show some results of general simulated paths between

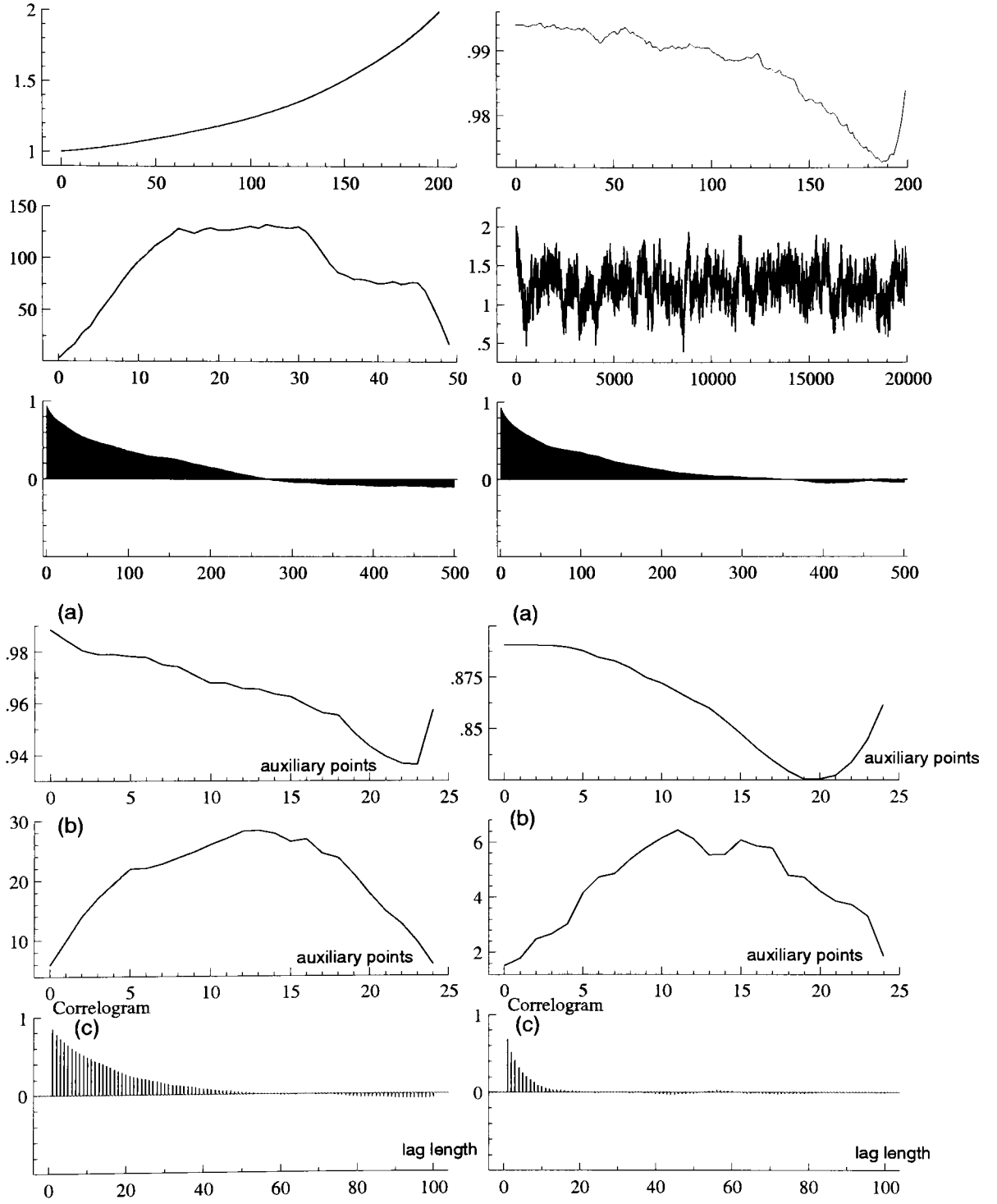


Figure 5.1: *Top: summary plots for the quadratic drift process, using an average block size of* 10, *with* $M = 200$ *and* $N = 20,000$. *From top left: mean of the simulations (a), average acceptance rates (b), inefficiency factor estimates (c), simulated values of the mid-auxiliary state (d) and the correlograms for the mid-auxiliary state (e) and quarter-auxiliary state (f). Figures (e) and (f) show the correlations over the iterations, with the auxiliary state fixed. Bottom: acceptance rates (a), inefficiency factor estimates (b) and correlograms (c) for average block sizes of* $\lambda = 2$ *(left panels) and* $\lambda = 10$ *(right panels),* $M = 25$ *and* $N = 10,000$, *for the quadratic drift process.*

two fixed points for this process. The summary graphs in the top plot of Figure 5.1 depict the average of the simulated paths between two fixed points y_t and y_{t+1} for the case for $M = 200$, $N = 20,000$ and random block size with $\lambda = 10$. The first shows the mean value of y_t^*, averaged over the sweeps, the second shows the acceptance rates and the third shows the inefficiency factor estimate. The fourth graph shows simulated values of the mid-state, while the fifth and sixth depict the correlograms for the mid and quarter states, respectively. The bottom plot in Figure 5.1 shows the case when we have $M = 25$ auxiliary points between the observed values and we compare the means of random block sizes 2 and 10. The iterations were run over $N = 10,000$ sweeps. Though the acceptance rates decline slightly, there is a marked improvement in both efficiency and convergence for larger block sizes. inefficiency factor estimates decrease from 30 to 6 at the middle states indicating that we are only six times as inefficient as if we were to draw i.i.d. samples. The correlogram is even more striking, showing heavy persistence at lags of 60 for the block size of 2 case against a correlogram that dies out by 30 lags for the block size of 10 case. On comparing these to Figure 4.1, we see that single-state sampling performance is poor.

We can also plot the autocorrelations for the mid-state for different block sizes, for $M = 25$ and $M = 200$. This is shown in the two graphs in Figure 5.2. Notice the clear decrease in autocorrelations as the block sizes increase, (which is more apparent in the random block case), and the improvement over the fixed block case. The means for the states are depicted in top plot in Figure 5.3. Notice the effects of the increase in block size in estimation of the mean and how the random block size is smoother than the fixed block case.

The bottom plot in Figure 5.3 shows the case when $M = 200$ and compares average random block sizes of 2 and 50. The iterations were run over 20,000 sweeps. Again, acceptance rates decline slightly when λ is larger, but there is an increase in both efficiency and convergence. inefficiency factor estimates decrease from 1000 to 15 at the middle states indicating that we are only 15 times as inefficient as if we were to draw from i.i.d. samples. The correlogram shows heavy persistence at lags of 500 for the block size of 2 case against a correlogram that dies out by 100 lags for the block size of 50 case.

5.2.3 Vasicek model

The Vasicek model is the mean-reverting version of the OU process and can be expressed as

$$dy(t) = \{\alpha - \beta y(t)\}dt + \sigma dW(t)$$

where the short-term interest rate $y(t)$ is defined as the strong solution to the diffusion process just defined. The parameters α, β and σ are strictly positive constants. The strong solution of the process is a Markov process with continuous sample paths and Gaussian increments. For a detailed exposition on the Vasicek model, see Vasicek (1977) and Musiela and Rutkowski (1997).

The explicit solution of the process is characterised by

$$y(\tau_t) = y(\tau_{t-1})\exp\{-\beta(\tau_t - \tau_{t-1})\} + \frac{\alpha}{\beta}[1 - \exp\{-\beta(\tau_t - \tau_{t-1})\} + \sigma\int_{\tau_{t-1}}^{\tau_t}\exp\{-\beta(\tau_t - s)\}\,dW_s.$$

Hence, the conditional distribution of $y(\tau_t)|y(\tau_{t-1})$ is normal with mean and variance given by

$$\begin{aligned} E\{y(\tau_t)|y(\tau_{t-1})\} &= y(\tau_{t-1})\exp\{-\beta(\tau_t - \tau_{t-1})\} + \frac{\alpha}{\beta}[1 - \exp\{-\beta(\tau_t - \tau_{t-1})\}] \\ var\{y(\tau_t)|y(\tau_{t-1})\} &= \frac{\sigma^2}{2\beta}[1 - \exp\{-2\beta(\tau_t - \tau_{t-1})\}]. \end{aligned}$$

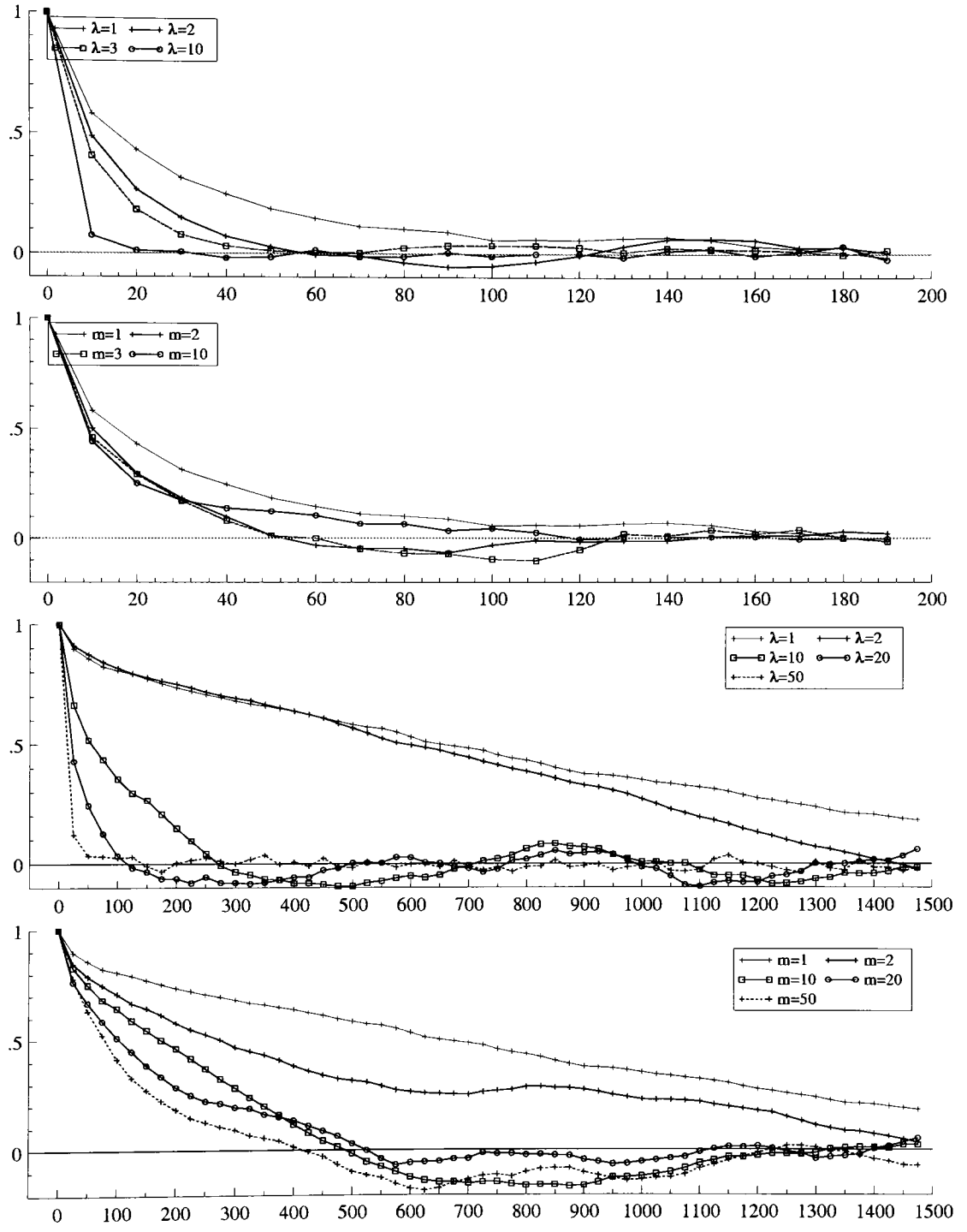


Figure 5.2: *Top: correlogram for different average blocks sizes (top graph) and fixed block sizes (bottom graph), $M = 25$ and $N = 10,000$, for the quadratic drift process. Bottom: correlogram for different average blocks sizes (top plot) and fixed block sizes (bottom plot), $M = 200$ and $N = 20,000$, for the quadratic drift process.*

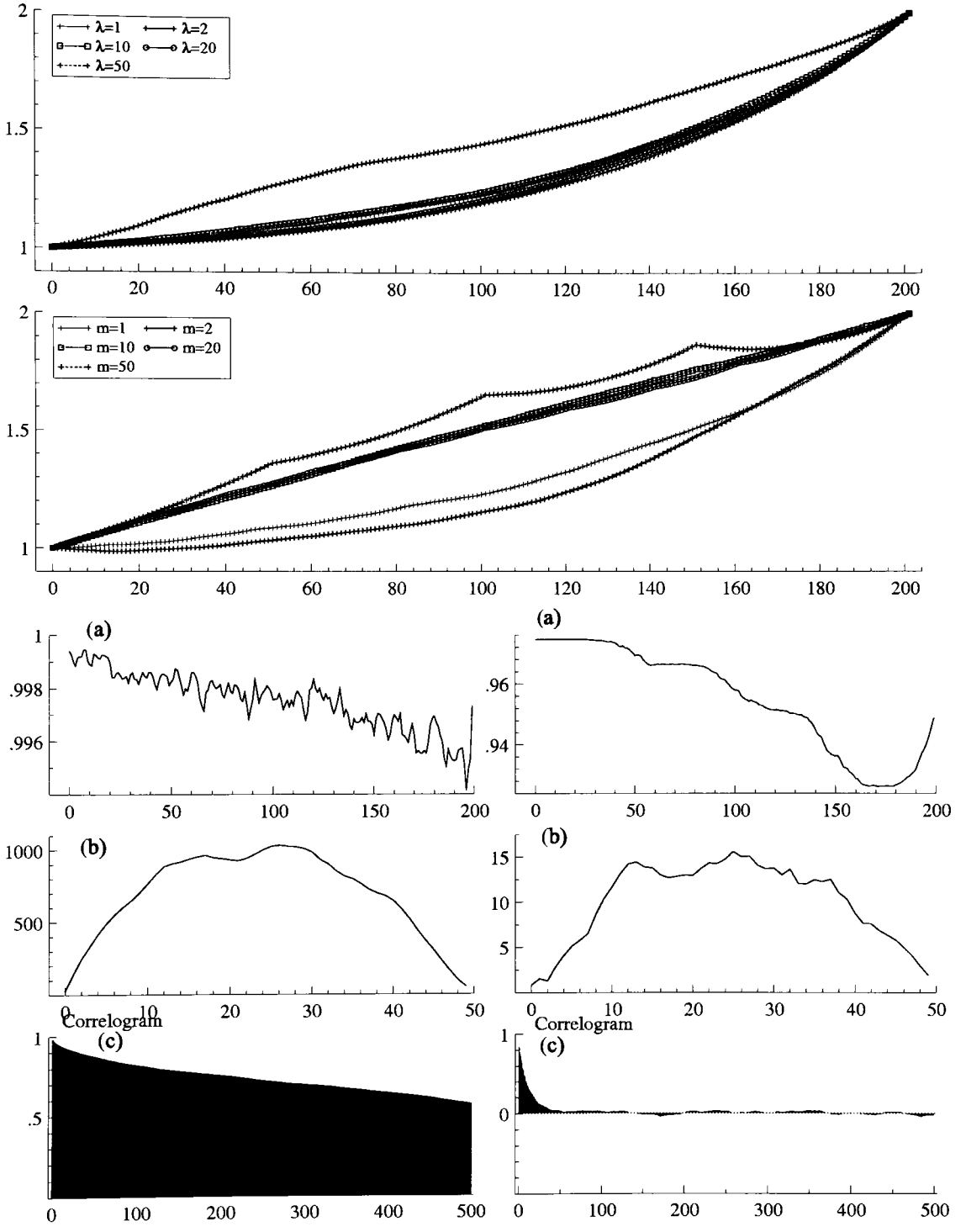


Figure 5.3: *Top plot shows the means for the states in the quadratic drift process, using* $M = 200$ *and* $N = 20,000$. *The first panel shows the smoothed mean from using random block sizes of* $\lambda = l = 1, 2, 10, 20$ *and* 50 *against the mean using equivalent fixed block sizes in the second panel. Bottom: acceptance rates (a), inefficiency factor estimates (b) and correlogram (c) for block sizes of* $\lambda = 2$ *and* 50, $M = 200$ *and* $N = 20,000$, *for the quadratic drift process.*

Turning to the Taylor expansion of the log-likelihood, we follow the procedure outlined in the appendix, section 5.6.1. Applying the Euler approximation, we obtain

$$y^*_{t,k+1} = y^*_{t,k} + \alpha\Delta - \beta y^*_{t,k}\Delta + \sigma(W_{t,k+1} - W_{t,k})$$

implying that

$$y^*_{t,k+1}|y^*_{t,k} \sim N\{\alpha\Delta + (1-\beta\Delta)y^*_{t,k}, \sigma^2\Delta\}.$$

Hence we can express the log-likelihood as

$$l(y^*_t) = c - \frac{1}{2\sigma^2\Delta}\sum_{j=0}^{m}[y^*_{t,k+j} - \{\alpha\Delta + (1-\beta\Delta)y^*_{t,k+j-1}\}]^2$$

The quantities which determine the parameters $\tilde{\mu}$ and $\tilde{\Sigma}$ of the proposal distribution are given by

$$\begin{aligned}
&a_{t,j} = \alpha - \beta y^*_{t,k+j}, \;\; a'_{t,j} = -\beta, \;\; a''_{t,j} = 0 \\
&g_{t,j} = \sigma^{-2}, \;\; g'_{t,j} = 0, \;\; g''_{t,j} = 0 \\
&d_{t,j} = y^*_{t,k+j+1} - \{\alpha\Delta + (1-\beta\Delta)y^*_{t,k+j}\} = y^*_{t,k+j+1} - (\alpha\Delta + \gamma y^*_{t,k+j}) \\
&c_{t,j} = 1 - \beta\Delta = \gamma.
\end{aligned}$$

Thus

$$\begin{aligned}
u_{t,j+1} &= -\frac{1}{\sigma^2\Delta}[\{y^*_{t,k+j} - (a\Delta + \gamma y^*_{t,k+j-1}\Delta)\} - \gamma\{y^*_{t,k+j+1} - (a\Delta + \gamma y^*_{t,k+j}\Delta)\}] \\
V^t_{j+1,j+1} &= \frac{1}{\sigma^2\Delta}(1+\gamma^2) \\
V^t_{j+1,j+2} &= -\frac{\gamma}{\sigma^2\Delta},
\end{aligned}$$

for $j = 0, \ldots, m-1$, all evaluated at $y^*_t = \hat{y}^*_t$.

The dramatic decrease in the correlograms indicates that estimates are unreliable for the single-move case, unless we run the algorithm for a longer period of time, and that larger block sizes for this case are clearly more efficient. Similar results are obtained from Figure 5.4, where the top plot shows the inefficiency factor estimates and correlograms for random block sizes of $1, 2$ and 10 for $M = 25$ and $N = 10,000$. There is a marked increase in efficiency as we increase the random block size; the top left hand side graphs shows that we are up to 80 times less efficient as compared to the case of independent samples, while the bottom left hand side graph shows that we are only 5 times as inefficient. In addition, there is a decrease in the persistence of the correlogram, indicating that larger block sizes (at least up to 10) for this case are more efficient. The bottom plot in Figure 5.4 shows the inefficiency factor estimates and correlograms for random block sizes of $1, 10$ and 50 for $M = 200$ and $N = 20,000$. Notice the decrease in inefficiency as we increase the random block size; the top left hand side graphs shows that we are up to 1200 times less efficient as compared to the case of independent samples, while the bottom left hand side graph shows that we are only 20 times as inefficient. The plots in Figure 5.5 shows the average of the simulations for random block sizes of $1, 2$ and 10 for $M = 200$ and $N = 20,000$. Again, there is little difference for the mean between the different block sizes. The same seed was used in all figures which may imply that we are underestimating the true variability.

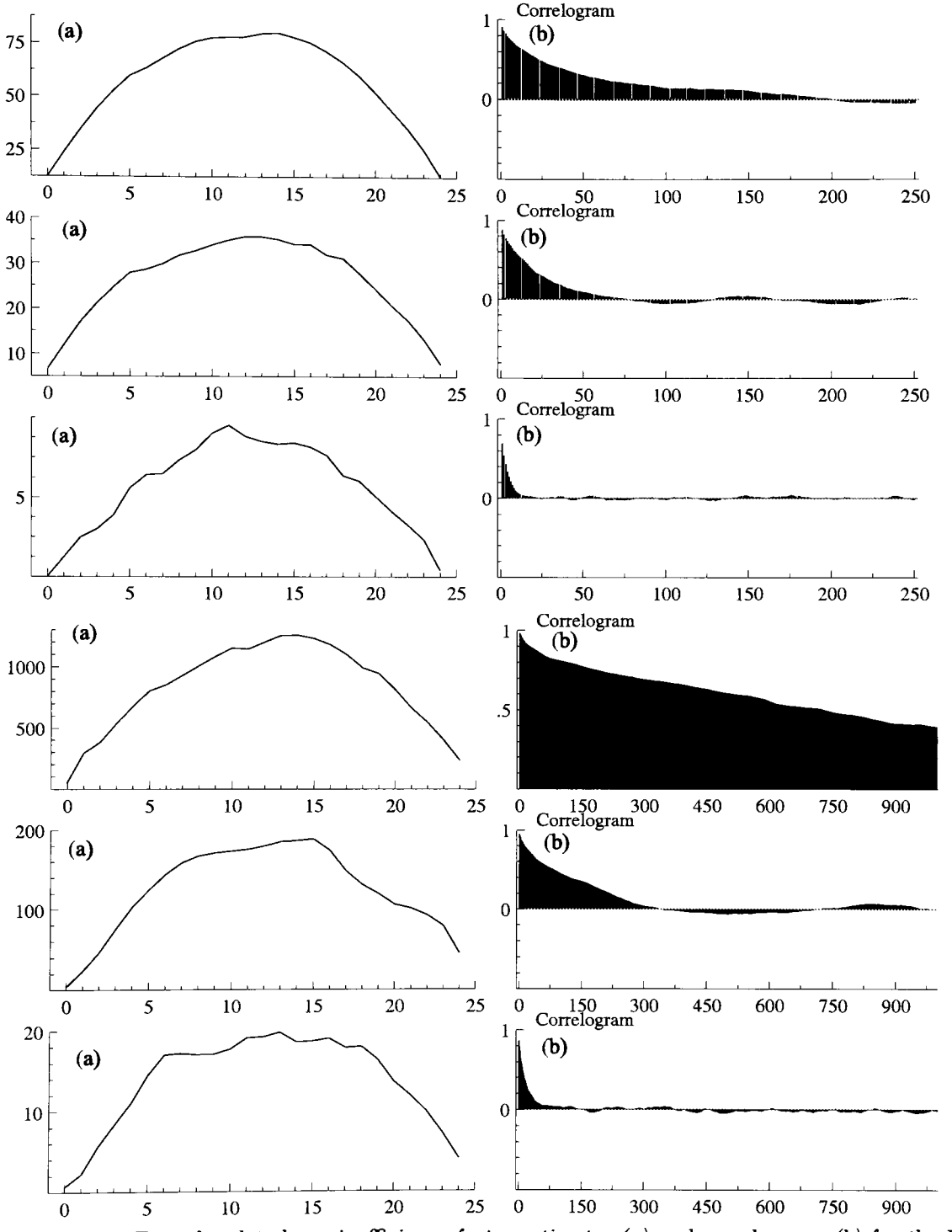


Figure 5.4: *Top: the plot shows inefficiency factor estimates (a) and correlograms (b) for the Vasicek model for random block sizes of* $1, 2$ *and* 10 *for* $M = 25$ *and* $N = 10,000$. *Bottom: inefficiency factor estimates (a) and correlograms (b) for the Vasicek model, for random block sizes of* $1, 10$ *and* 50 *for* $M = 200$ *and* $N = 20,000$ *are shown.*

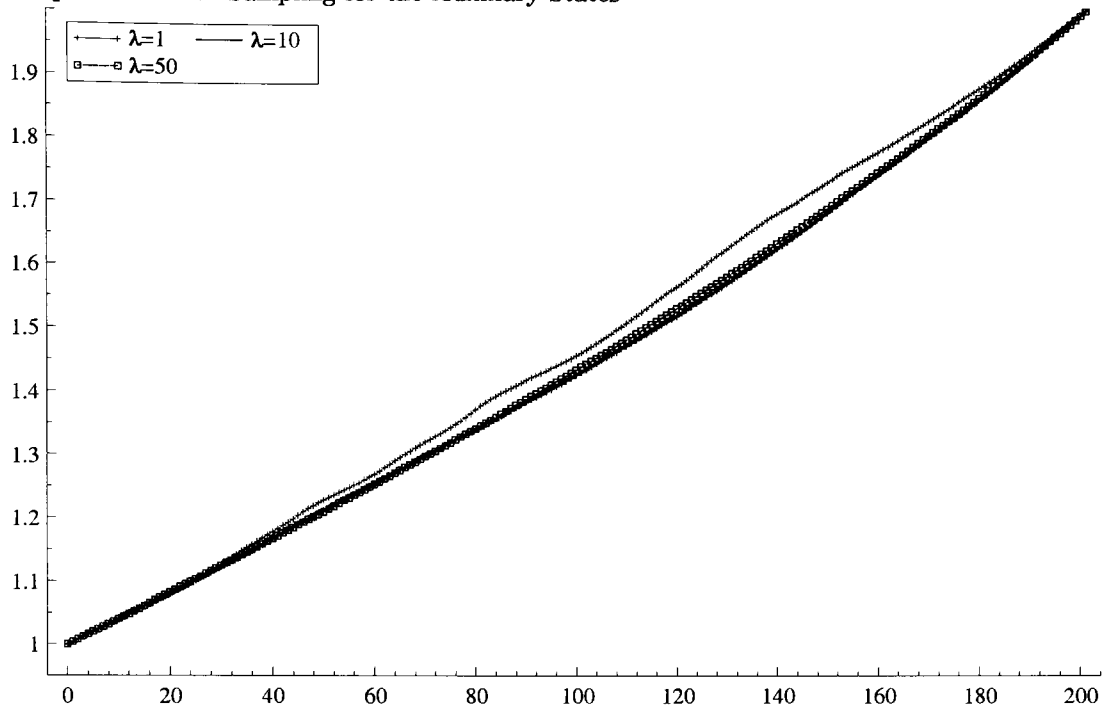


Figure 5.5: *The average of the simulations for the Vasicek model are shown for random block sizes of* $1, 10$ *and* 50 *for* $M = 200$ *and* $N = 20,000$.

5.2.4 Cox, Ingersoll, and Ross (CIR) model

The CIR model, also known as the *square-root process*, was developed by Cox, Ingersoll, and Ross (1985) for term-structure modelling, and can be expressed as

$$dy(t) = \{\alpha - \beta y(t)\}dt + \sigma\sqrt{y(t)}dW(t)$$

where $y(t)$ is the short-term interest rate. The presence of the square-root in the diffusion coefficient means that the diffusion only takes positive values, though it can reach zero.

For our purposes, we are interested in finding the Euler approximation and taking a second-order Taylor expansion of the log-likelihood, though the explicit solution is available and leads to a non-central chi-squared distribution for the conditional transition density and a Gamma distribution for the marginal density, see Feller (1951). Suppose we apply the transformation $x(t) = \log y(t)$, due to the restriction that $y(t) > 0$.applying Itô's lemma (see for example Øksendal 1995), we obtain

$$dx(t) = \left[\frac{\alpha}{\exp\{x(t)\}} - \beta - \frac{\sigma^2}{2\exp\{x(t)\}}\right]dt + \frac{\sigma}{\exp\{x(t)/2\}}dW(t)$$

Using the Euler approximation on the transformed variable

$$x^*_{t,k+1} = x^*_{t,k} + \left\{\frac{\alpha}{\exp(x^*_{t,k})} - \beta - \frac{\sigma^2}{2\exp(x^*_{t,k})}\right\}\Delta + \frac{\sigma}{\sqrt{\exp(x^*_{t,k})}}\eta_{t,k+1}$$

where $\eta_{t,k} \sim N(0, \Delta)$, implying that

$$x^*_{t,k+1}|x^*_{t,k} \sim N\left[\left\{\frac{\alpha}{\exp(x^*_{t,k})} - \beta - \frac{\sigma^2}{2\exp(x^*_{t,k})}\right\}\Delta + x^*_{t,k}, \frac{\sigma^2}{\exp(x^*_{t,k})}\Delta\right].$$

So the log-likelihood for a block of m points can be expressed as,

$$\begin{aligned} l(x^*_t) &= c - \frac{1}{2}\sum_{j=0}^{m} \log\left\{\frac{\sigma^2\Delta}{\exp(x^*_{t,k+j-1})}\right\} \\ &\quad -\frac{1}{2\sigma^2\Delta}\sum_{j=0}^{m}\exp(x^*_{t,k+j-1})\left[x^*_{t,k+j} - \left\{\frac{\alpha}{\exp(x^*_{t,k+j-1})} - \beta - \frac{\sigma^2}{2\exp(x^*_{t,k+j-1})}\right\}\Delta - x^*_{t,k+j-1}\right]^2. \end{aligned}$$

This is just the standard Euler for $x^*_{t,k+j} = \log(y^*_{t,k+j})$.[2] Using the notation in appendix 5.6.1, we can define the following

$$\begin{aligned} a_{t,j} &= \frac{\alpha}{\exp(x^*_{t,k+j})} - \beta - \frac{\sigma^2}{2\exp(x^*_{t,k+j})} \\ g_{t,j} &= \frac{\exp(x^*_{t,k+j})}{\sigma^2} \\ d_{t,j} &= x^*_{t,k+j+1} - x^*_{t,k+j} - a_{t,j}\Delta \\ c_{t,j} &= 1 + a'_{t,j}\Delta. \end{aligned}$$

We thus obtain,

$$u_{t,j+1} = -\frac{1}{\Delta}\left(g_{t,j-1}d_{t,j-1} - g_{t,j}d_{t,j}c_{t,j} + \frac{1}{2}g'_{t,j}d^2_{t,j} - \frac{g'_{t,j}\Delta}{2g_{t,j}}\right) \tag{5.6}$$

and

$$\begin{aligned} V^t_{j+1,j+1} &= \frac{1}{\Delta}\left[g_{t,j-1} + g_{t,j}c^2_{t,j} - g_{t,j}d_{t,j}a''_{t,j}\Delta - 2g'_{t,j}d_{t,j}c_{t,j} + \frac{1}{2}g''_{t,j}d^2_{t,j} - \right. \\ &\qquad \left. \frac{\{g_{t,j}g''_{t,j} - (g'_{t,j})^2\}\Delta}{2g^2_{t,j}}\right] \end{aligned} \tag{5.7}$$

$$V^t_{j+1,j+2} = V^t_{j+2,j+1} = \frac{1}{\Delta}(-g_{t,j}c_{t,j} + g'_{t,j}d_{t,j}) \tag{5.8}$$

[2] In terms of $y^*_{t,k}$, this is simply,

$$\log y^*_{t,k+j} = \log x^*_{t,k+j-1} + \left(\frac{\alpha}{y^*_{t,k+j-1}} - \beta - \frac{\sigma^2}{2y^*_{t,k+j-1}}\right)\Delta + \left(\frac{\sigma^2}{y^*_{t,k+j-1}}\right)^{\frac{1}{2}}\eta_{t,k+j}.$$

for $j = 0, \ldots, m-1$, where

$$\begin{aligned} a'_{t,j} &= -\frac{\alpha}{\exp(x^*_{t,k+j})} + \frac{\sigma^2}{2\exp(x^*_{t,k+j})} \\ a''_{t,j} &= \frac{\alpha}{\exp(x^*_{t,k+j})} - \frac{\sigma^2}{2\exp(x^*_{t,k+j})} \\ g'_{t,j} &= g''_{t,j} = \frac{\exp(x^*_{t,k+j})}{\sigma^2} \end{aligned}$$

Some results from implementing the M-H algorithm for the CIR model are shown in Figures 5.6–5.7. The first simulation (top plot in Figure 5.6) was performed with $M = 25$, $N = 10,000$, using a random block size with mean 2. This shows that we are up to 75 times less efficient than an i.i.d. sampler. Correlograms for both the mid-state and quarter state die down after 100 lags. For the second simulation (bottom plot in Figure 5.6), the number of simulations was $N = 20,000$ with $M = 200$ auxiliary points and average random block sizes of 2, 5 and 20 were used. inefficiency factor estimates show that for $\lambda = 2$ and 20, the sampler can be up to 1500 times less efficient than a sampler drawing i.i.d. random variables. $\lambda = 5$ is up to 1000 times less efficient, but this suggests that there is an optimal block size for each value of M. This is reinforced by Figure 5.7, which shows the autocorrelations of simulated values for the mid-state, with $M = 25$ and $N = 10,000$, using average random block sizes of 2 and 10 (top plot), and $M = 200$ with $N = 20,000$, using average random block sizes of 2, 5 and 20 (lower plot). The parameters were $\alpha = 0.5$, $\beta = 0.8$ and $\sigma^2 = 0.03$ for all the simulations.

5.2.5 The Hull-White process

This model can again be used for term-structure modelling, and can be expressed as

$$dy(t) = \{\alpha - \beta y(t)\}dt + \sigma y(t)^{\gamma} dW(t)$$

where $y(t)$ is the short-term interest rate. Suppose again that we take the transformation $x(t) = \log y(t)$, due to the restriction that $y(t) > 0$. Then on applying Itô's lemma we obtain

$$dx(t) = \left[\frac{\alpha}{\exp\{x(t)\}} - \beta - \frac{\sigma^2}{2}\exp\{2(\gamma-1)x(t)\}\right] dt + \sigma \exp\{(\gamma-1)x(t)\} dW(t).$$

The Euler approximation takes the form,

$$x^*_{t,k+1} = x^*_{t,k} + \left[\frac{\alpha}{\exp(x^*_{t,k})} - \beta - \frac{\sigma^2}{2}\exp\{2(\gamma-1)x^*_{t,k}\}\right]\Delta + \sigma\exp\{(\gamma-1)x^*_{t,k}\}(W_{t,k+1} - W_{t,k})$$

implying that

$$x^*_{t,k+1}|x^*_{t,k} \sim N\left[x^*_{t,k} + \frac{\alpha\Delta}{\exp(x^*_{t,k})} - \beta\Delta - \frac{\sigma^2\Delta}{2}\exp\{2(\gamma-1)x^*_{t,k}\}, \sigma^2\exp\{2(\gamma-1)x^*_{t,k}\}\Delta\right].$$

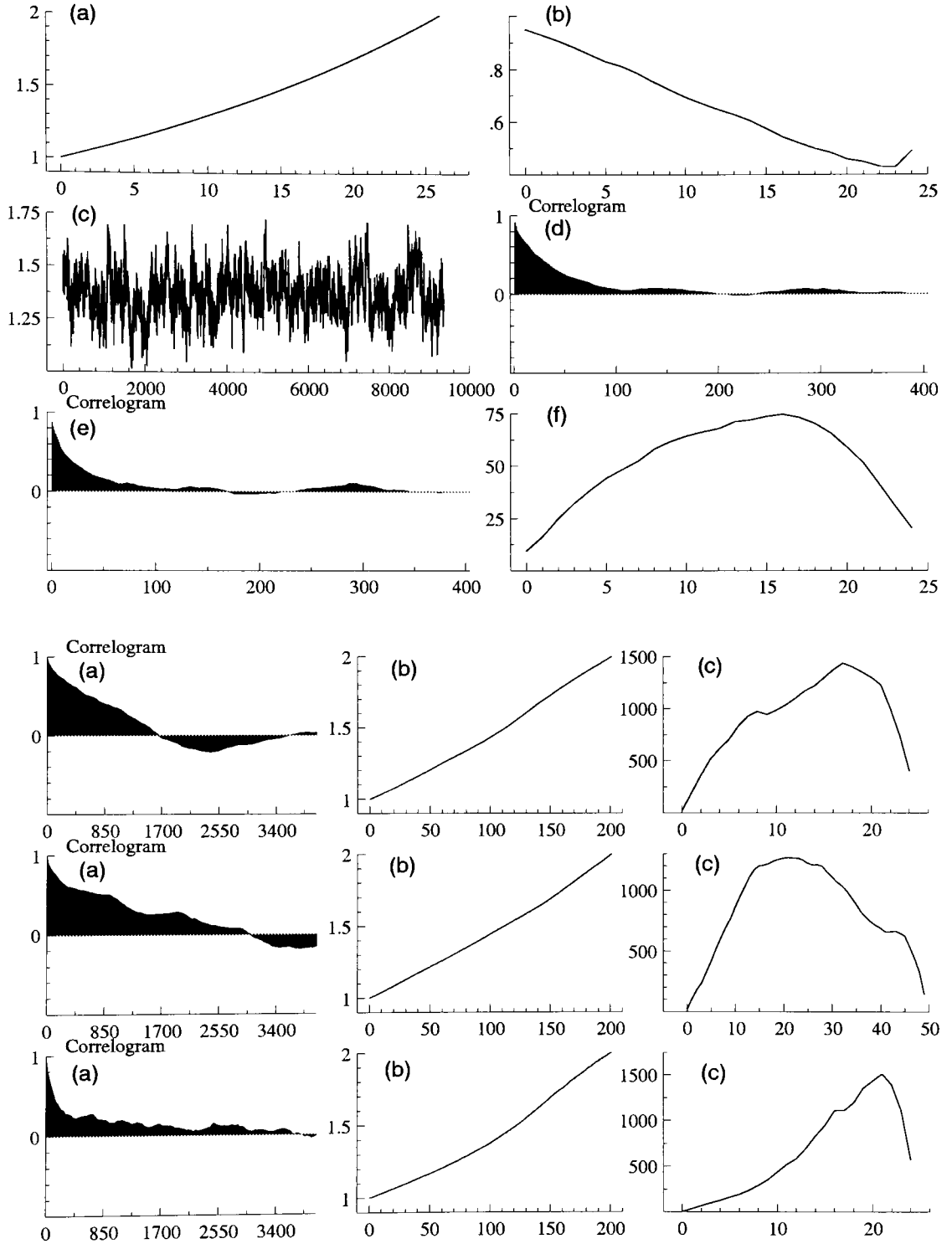


Figure 5.6: *Top: summary plots for the CIR model, with $M = 25$ and $\lambda = 2$. The plots are, clockwise from top left, simulation mean of y (a), the acceptance rates (b), the simulated values of the mid-state (c), correlograms for the mid-auxiliary (d) and quarter-auxiliary (e) states, and the inefficiency factor estimates (f). Bottom: summary plots for the CIR model, with $M = 200$ and $N = 20,000$. Graphs show correlograms for the mid-state (a), the means of the simulations (b) and the inefficiency factor estimates (c).*

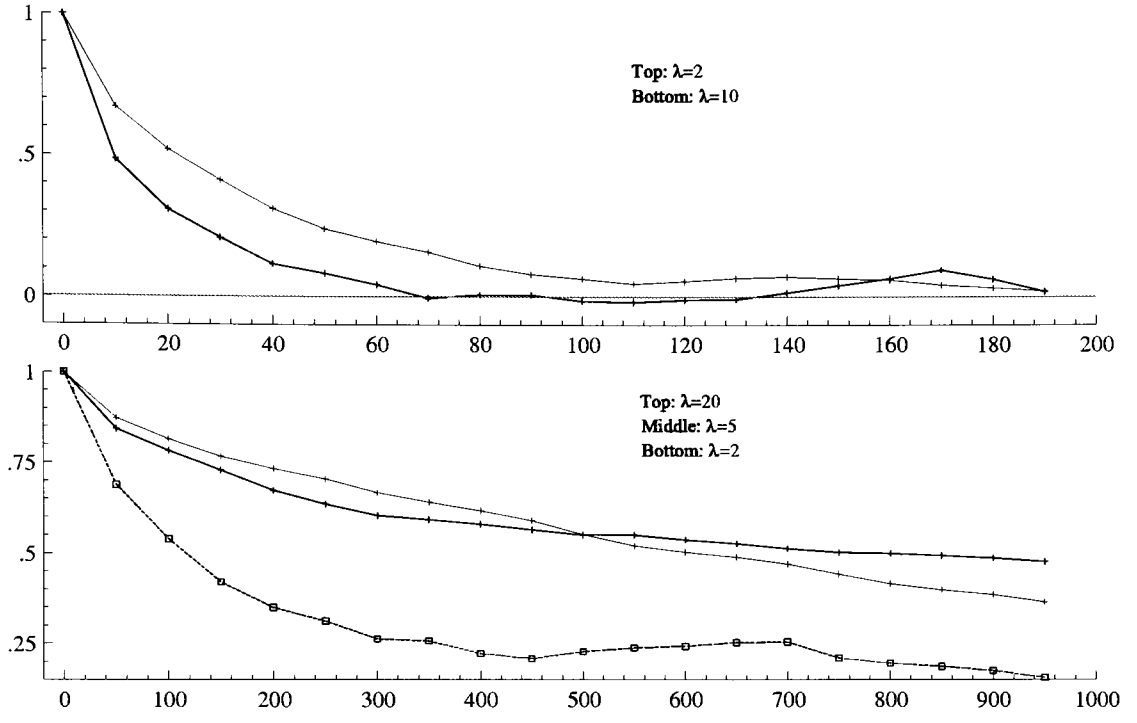


Figure 5.7: *Autocorrelations of simulated values for the mid-state are shown for the CIR model, with* $M = 25$ *and* $N = 10,000$*, using average random block sizes of 2 and 10 (top plot). The lower plot uses* $M = 200$ *discretisation points, with* $N = 20,000$ *and average random block sizes of 2, 5 and 20.*

So the log-likelihood for a block of m points can be expressed as

$$l(x_t^*) = c - \frac{1}{2}\sum_{j=0}^{m} \log\left[\sigma^2 \exp\{2(\gamma-1)x^*_{t,k+j-1}\}\Delta\right] - \frac{1}{2\sigma^2\Delta}\sum_{j=0}^{m} \frac{\left[x^*_{t,k+j} - x^*_{t,k+j-1} - \frac{\alpha\Delta}{\exp\{x^*_{t,k+j-1}\}} + \beta\Delta + \frac{\sigma^2\Delta}{2}\exp\{2(\gamma-1)x^*_{t,k+j-1}\}\right]^2}{\exp\{2(\gamma-1)x^*_{k+j-1}\}}.$$

Again by defining $x^*_{t,k+j} = \log(y^*_{t,k+j})$we can see that the expression just the standard Euler considered for some of the earlier examples, except that it is defined in terms of logarithms. We can then define

$$\begin{aligned}
a_{t,j} &= \frac{\alpha}{\exp(x^*_{t,k+j})} - \beta - \frac{\sigma^2}{2}\exp\{2(\gamma-1)x^*_{t,k+j}\} \\
a'_{t,j} &= -\frac{\alpha}{\exp(x^*_{t,k+j})} - \sigma^2(\gamma-1)\exp\{2(\gamma-1)x^*_{t,k+j}\} \\
a''_{t,j} &= \frac{\alpha}{\exp(x^*_{t,k+j})} - 2\sigma^2(\gamma-1)^2\exp\{2\gamma-1)x^*_{t,k+j}\} \\
g_{t,j} &= \frac{1}{\sigma^2\exp\{2(\gamma-1)x^*_{t,k+j}\}} \\
g'_{t,j} &= \frac{1}{\sigma^2} - \frac{2(\gamma-1)}{\exp\{2(\gamma-1)x^*_{t,k+j}\}} \\
g''_{t,j} &= \frac{4(\gamma-1)^2}{\sigma^2\exp\{2(\gamma-1)x^*_{t,k+j}\}} \\
d_{t,j} &= x^*_{t,k+j+1} - x^*_{t,k+j} - a_{t,j}\Delta \\
c_{t,j} &= 1 + a'_{t,j}\Delta.
\end{aligned}$$

For $j = 0, \ldots m-1$, on applying the formulae in Appendix 5.6.1, we obtain equations 5.6–5.8, as for the CIR process. Some results of running the M-H algorithm for the Hull-White model are shown in the plots in Figure 5.8. For the top plot, $M = 25$, the number of simulations was $N = 10,000$, and the random block size was 2. For the bottom plot, autocorrelations of simulated values of state $M/2$ is shown, with $M = 25$, $N = 10,000$, block sizes of 2, 5 and $M = 200$, $N = 20,000$, with block sizes of 2,5, and 10 respectively. In all cases, the parameters were $\alpha = 0.8$, $\beta = 0.5$, $\gamma = 0.25$ and $\sigma^2 = 0.3$.

5.3 How to ensure a positive definite proposal variance matrix

The matrix $V_t = \Sigma^{-1}$ in the equations in 5.6.1 is not guaranteed to be positive definite. For example, consider the Hull-White model (block size of one) and $\alpha = \beta = 1$, $\gamma = 0.4$, $M = 1$, so that $\Delta = (1/2)$, $y^*_{t,k+1} = 20$, $\hat{y}^*_{t,k} = 15$ and $y^*_{t,k-1} = 10$. Then the diagonal element of V_t is

$$V_{11} = -6.28.$$

Though this example was chosen with M small, it nonetheless shows that V_t may not be positive definite. The problem is rare for most models, since the problematic terms in V_t depend on $d_{t,j}$ and/or Δ, which are normally small, when M is fairly large and $\hat{y}^*_t$ is well chosen. However, it can occur for the more complicated models such as the Hull-White model, as was shown above. The Choleski decomposition of V_t will fail if V_t is negative definite, indefinite or singular. A matrix $V_t \in \mathbb{R}^{m\times m}$ is positive definite if $x'Ax > 0$ for all non-zero $x \in \mathbb{R}^m$. If the quadratic form takes on both positive and negative values, then V_t is indefinite. Systematic 'pivot' strategies can be employed to deal with indefinite matrices, (see Golub and van Loan (1996), Section 4.4). In addition, all its principle submatrices are positive definite, in particular, all the diagonal entries are positive.

If V_t is non-positive definite, there are various techniques one may wish to employ. Using a suitable algorithm, one can check whether the determinant of V_t is positive definite. In the event that the determinant is non-positive, conditions on the elements of V_t can be imposed. This essentially implies placing restrictions on V_t, through $g_{t,j}$, $a_{t,j}$ and their respective derivatives; $c_{t,j}$ and $d_{t,j}$ must obviously

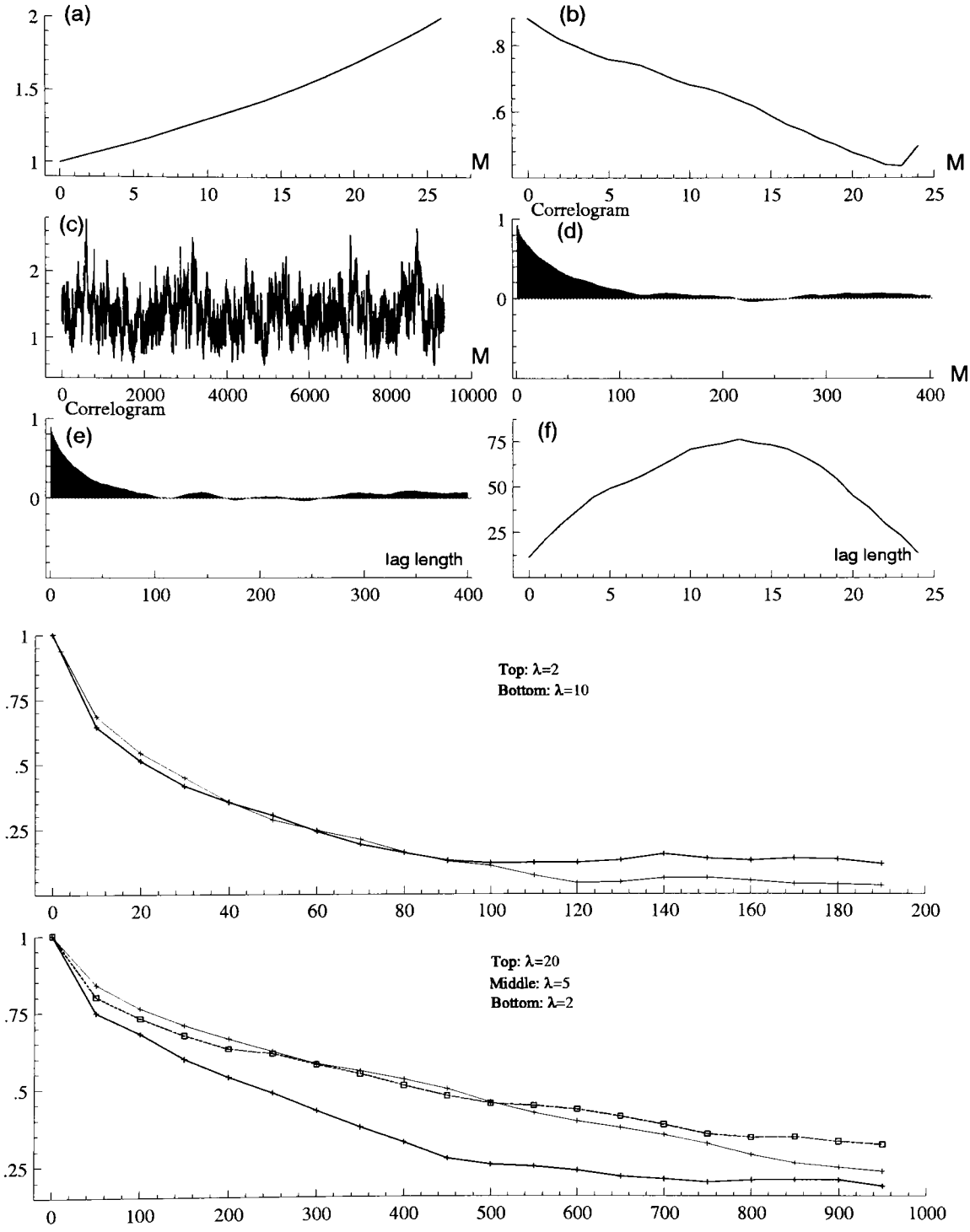


Figure 5.8: *Top: summary plots for the Hull-White model, with* $M = 25$ *and* $\lambda = 2$*. The plots are the same as in figure 5.6. Bottom: autocorrelations of simulated values for the mid-state for the Hull-White model, with* $M = 25$ *and* $N = 10,000$*, using average random block sizes of 2, 5 (top panel). Bottom panel uses* $M = 200$ *discretisation points and* $N = 20,000$ *iterations, with average random block sizes of 2,5, and 10.*

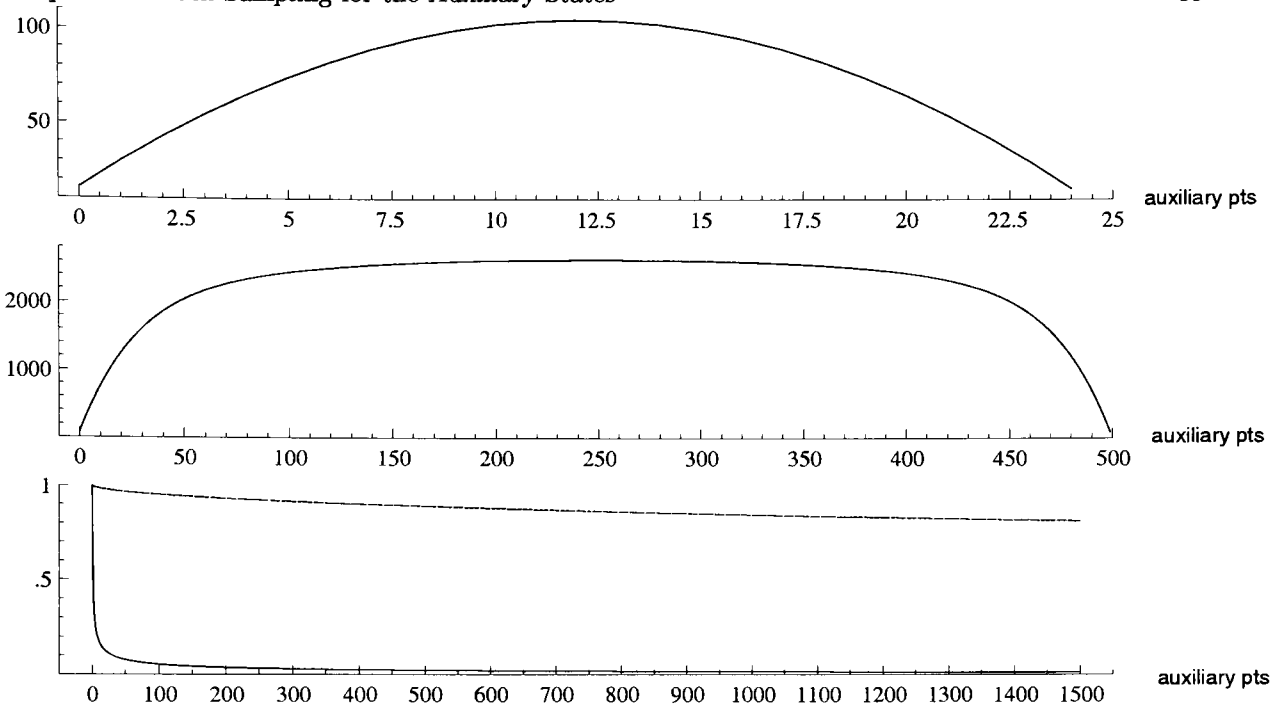


Figure 5.9: *The top and middle plots show the theoretical efficiency for the M-H sampler for an OU process with $M = 25$, 500, for the first state and nid-state respectively. The bottom plot shows the theoretical autocorrelation function for the first state (bottom line) and the mid-state (top line), with $M = 500$ and $M = 1500$.*

be left to vary freely. One attempt would be to change $\hat{y}_t^*$ and iterate towards the mode of $l(y_t^*)$ until V_t was positive definite. An alternative would be to set the appropriate derivatives of the likelihood to zero; this method will subsequently be explored in more detail. For example, if the non-linear volatility model has $\gamma \neq 0.5$, and has non-positive definite V_t, we could use a proposal based on the CIR model, which has $\gamma = 0.5$. We could also restrict the diagonal elements to

$$V_{j+1,j+1}^t = \frac{1}{\Delta}\left(g_{t,j-1} + g_{t,j}c_{t,j}^2\right).$$

In addition, we could define

$$V_{j+1,j+2}^t = \frac{1}{\Delta}(-g_{t,j}c_{t,j}),$$

that is, setting $a'' = g' = g'' = 0$ everywhere. The resulting V_t is positive definite. For $m = 2$:

$$\begin{aligned} |V_t| &= \frac{1}{\Delta}\begin{vmatrix} g_{t,-1} + g_{t,0}c_{t,0}^2 & -g_{t,0}c_{t,0} \\ -g_{t,0}c_{t,0} & g_{t,0} + g_{t,1}c_{t,1}^2 \end{vmatrix} \\ &= \frac{1}{\Delta}\{(g_{t,-1} + g_{t,0}c_{t,0}^2)(g_{t,0} + g_{t,1}c_{t,1}^2) - g_{t,0}^2c_{t,0}^2\} \\ &= \frac{1}{\Delta}(g_{t,-1}g_{t,0} + g_{t,-1}g_{t,1}c_{t,1}^2 + g_{t,0}g_{t,1}c_{t,0}^2c_{t,1}^2) > 0 \end{aligned}$$

The general result can be shown by induction. This ad-hoc matrix is actually the limit of the original V_t when M tends to infinity and $d_{t,j}$ tends to zero, for all j. However, though we are free to choose the proposal density and setting $a'' = g' = g'' = 0$ is fully general, the method is not entirely optimal.

We can identify three separate cases for which V_t is not positive definite: V_t negative definite, indefinite, or singular. In the case where V_t is negative definite or singular, we could adopt a quadratic-hill climbing approach, first proposed by Goldfeld, Quandt, and Trotter (1966), by introducing a positive definite matrix $H = (V_t + \mu I)^{-1}$, where μ varies as the process is iterated. For example, if V_t is negative definite, set μ larger than the greatest negative eigenvalue of V_t, while if V_t is singular, set μ greater than zero such that H is positive definite. As the maximum is approached, μ can be set to zero, reducing the algorithm to the Newton-Raphson method.

We adopt a method proposed by Greenstadt (1967), which can deal with all three cases simultaneously. Suppose π_i is the i'th characteristic root of V_t and c_i is its associated characteristic vector, then for a block size of m

$$V_{t,\pi} = \sum_{i=1}^{m} |\pi_i| c_i c_i'.$$

V_t is then guaranteed to be positive definite. Finally, we use $\Sigma = V_t^{-1}$ where V_t is the first positive definite V_t. For the CIR and Hull-White processes, using the logarithms of the states proves to be more stable in ensuring that V_t is positive definite, than using the non-transformed states.

5.4 Assessing the efficiency of the sampler

The efficiency of the sampler can be assessed theoretically in the case of the OU process, due to the constant volatility coefficient in the specification of the SDE, as was shown in Section 4.3 for the single-move sampler. These results are now extended to deal with blocks and they show that the sampler is typically more efficient at the end points and (not surprisingly) least efficient for the mid-state. For most models however, this is not possible and a numerical method is presented in Section 5.4.2.

5.4.1 A theoretical analysis of the efficiency of the sampler

The geometric rate of convergence of a chain was defined in Section 3.2.2. Roberts and Sahu (1997) provide exact computable rates of convergence for the Gibbs sampler, when the target is Gaussian.

Consider a process with non-zero mean,

$$y(t+1) = By(t) + b + \epsilon(t+1)$$

where assuming normality, $\text{var}\{y(t+1)\} = \tilde{\Sigma}$ implying that $\text{var}\{y(t+1)|y(t)\} = \tilde{\Sigma} - B\tilde{\Sigma}B'$ and $\text{E}\{y(t+1)|y(t)\} = By(t) + b$. We can express the autocovariance function as

$$\begin{aligned}
\Gamma(s) &= \text{E}\{y(t)y(t-s)'\} - \text{E}\{y(t)\}\text{E}\{y(t-s)'\} \\
&= B\text{E}\{y(t-1)y(t-s)'\} + b\text{E}\{y(t-s)'\} + \text{E}\{\epsilon(t)y(t-s)'\} - \text{E}\{y(t)\}\text{E}\{y(t-s)'\} \\
&= B\Gamma(s-1) + \text{E}\{\epsilon(t)y(t-s)'\} - \text{E}\{\epsilon(t)\}\text{E}\{y(t-s)'\} \\
&= B\Gamma(s-1) + \text{E}\{\epsilon(t)y(t-s)'\}.
\end{aligned}$$

For $s \geq 1$, the second term is zero, so

$$\Gamma(s) = B\Gamma(s-1)$$

while for $s = 0$, we obtain

$$\mathrm{E}\{\epsilon(t)y(t)'\} = \tilde{\Sigma} - B\tilde{\Sigma}B'$$

or more obviously, $\Gamma(0) = \mathrm{var}\{y(t)\} = \tilde{\Sigma}$ and so we obtain,

$$\Gamma(s) = B^s\tilde{\Sigma}.$$

Following the procedures of Roberts and Sahu (1997), we can express the conditional mean of y_t^* at iteration $(j+1)$ as

$$\mathrm{E}(y_t^{*(j+1)}|y_t^{*(j)}) = By_t^{*(j)} + b$$

where $B = (I-L)^{-1}U$, L represents the blocks on the lower triangle of a matrix A and $U = A - L$. For the OU process, $\tilde{\Sigma} = V_t^{-1}$ and

$$A = I - diag(\tilde{\Sigma}_{11}^{-1}, ..., \tilde{\Sigma}_{mm}^{-1})\tilde{\Sigma} = \begin{pmatrix} 0 & a & 0 & \cdots & 0 \\ a & 0 & a & & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & a & 0 & a \\ 0 & \cdots & 0 & a & 0 \end{pmatrix},$$

where the dimension of A is $m \times m$ and we denote $a = \frac{\rho}{1+\rho^2}$, where $\rho = 1 + \mu\Delta$. Then

$$B = \begin{pmatrix} 0 & a & 0 & \dots & 0 & 0 \\ 0 & a^2 & a & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & & & \vdots \\ 0 & a^{m-1} & a^{m-2} & \dots & a^2 & a \\ 0 & a^m & a^{m-1} & \dots & a^3 & a^2 \end{pmatrix}.$$

As $\tilde{\Sigma} = V_t^{-1}$. For m even, this can be expressed as

$$\tilde{\Sigma} = |V_t|^{-1} \times$$
$$\begin{pmatrix} e_0^{m-1} & \cdots & & & & & \\ o_0^{m-2} & e_0^{m-1} + e_1^{m-2} & \cdots & & & & \\ e_1^{m-2} & o_0^{m-2} + o_1^{m-3} & e_0^{m-1} + e_1^{m-2} & \cdots & & & \\ o_1^{m-3} & e_1^{m-2} + e_2^{m-3} & o_0^{m-2} + o_1^{m-3} & e_0^{m-1} + e_1^{m-2} & \cdots & & \\ \vdots & \vdots & \vdots & \vdots & \ddots & & \\ e_{(m-2)/2}^{(m-2)/2+1} & o_{(m-2)/2-1}^{(m-2)/2} + o_{(m-2)/2}^{(m-2)/2} & & & \cdots & e_0^{m-1} + e_1^{m-2} & \cdots \\ o_{(m-2)/2}^{(m-2)/2} & e_{(m-2)/2}^{(m-2)/2+1} & o_{(m-2)/2-1}^{(m-2)/2+1} & & \cdots & o_0^{m-2} & e_0^{m-1} \end{pmatrix}$$

where $e_l^u = \sum_{i=l}^{u} \rho^{2i}$, $o_l^u = \sum_{i=l}^{u} \rho^{2i+1}$ and $|V| = e_0^m = \sum_{i=0}^{m} \rho^{2i}$. Similarly, for m odd,

$$\tilde{\Sigma} = |V_t|^{-1} \times$$
$$\begin{pmatrix}
e_0^{m-1} & \cdots & & & & & \\
o_0^{m-2} & e_0^{m-1}+e_1^{m-2} & \cdots & & & & \\
e_1^{m-2} & o_0^{m-2}+o_1^{m-3} & e_0^{m-1}+e_1^{m-2} & \cdots & & & \\
o_1^{m-3} & e_1^{m-2}+e_2^{m-3} & o_0^{m-2}+o_1^{m-3} & e_0^{m-1}+e_1^{m-2} & \cdots & & \\
\vdots & \vdots & \vdots & \vdots & \ddots & & \\
o_{(m-1)/2}^{(m-1)/2+1} & e_{(m-1)/2-1}^{(m-1)/2+1}+e_{(m-1)/2}^{(m-1)/2} & & & \cdots & e_0^{m-1}+e_1^{m-2} & \cdots \\
e_{(m-1)/2}^{(m-1)/2} & o_{(m-1)/2}^{(m-1)/2+1} & e_{(m-1)/2-1}^{(m-1)/2+1} & & \cdots & o_0^{m-2} & e_0^{m-1}
\end{pmatrix}.$$

The matrix B^s has elements which consist of sums of increasing powers in s of 2 and a. A closed form expression is in general not available. Suppose the eigenvalue decomposition of B can be expressed as $L\Lambda L^{-1}$, where $\lambda = diag\{\lambda_1, \ldots, \lambda_m\}$ and L is a non-singular (mxm) matrix of eigenvectors obtained from setting $|B' - \lambda I| = 0$. Then if $|\lambda_i| < 1$ for all $i = (1, \ldots, m)$

$$B^s = L\Lambda^s L^{-1}.$$

Suppose we denote $|B - \lambda I|$ by D_m, then we can obtain the following recursion

$$D_m - (a^2 - \lambda)D_{m-1} + \sum_{i=2}^{m-1} (-1)^i a^{2i} D_{m-i} = 0$$

to obtain the eigenvalues numerically. The initial conditions are

$$\begin{aligned} D_0 &= 1 \\ D_1 &= -\lambda \\ D_2 &= \lambda^2 - a^2\lambda. \end{aligned}$$

The autocovariance matrix $\Gamma(s)$ can be standardised to obtain the autocorrelation matrix

$$P(s) = T_0^{-1}\Gamma(s)T_0^{-1}$$

where $T_0^2 = diag\{\tilde{\Sigma}\} = \{\gamma_1(0), \ldots, \gamma_m(0)\}$. Then

$$\begin{aligned} P(s) &= T_0^{-1}\Gamma(s)T_0^{-1} \\ &= [diag(\tilde{\Sigma})]^{-\frac{1}{2}} L\Lambda^s L^{-1} \tilde{\Sigma} [diag(\tilde{\Sigma})]^{-\frac{1}{2}}. \end{aligned}$$

Hence denoting each diagonal element of $P(s)$ as $\rho_{ii}(s)$ for $i = (1, \ldots, m)$ we can evaluate the theoretical estimate of the efficiency as

$$1 + 2\sum_{\tau=1}^{K} \rho_{ii}(s)$$

for some maximum lag length K. Note that the theoretical efficiency is an infinite sum, but in practice to evaluate this quantity, we can only sum to a finite index K.

The two graphs in the bottom plot of Figure 5.9 show the theoretical efficiency for the M-H sampler for an OU process, for $M = 25$, $K = 1000$ and $M = 500$ and with $K = 1500$. Notice that the efficiency decreases dramatically as the number of points M increases. The efficiency measure tends to a rectangular shape as M gets large, indicating that the inefficiency measure is the same regardless of the state; the efficiency being highest for states near the end points which are fixed. The third graph shows the theoretical autocorrelation function for the first and mid-state for $M = 500$, using a maximum lag length of $1,500$. One can notice that the autocorrelations die down reasonably quickly for the first state whereas for the mid-state (and most) other states, these remain high even after 1500 lags, as the middle graph shows.

5.4.2 Obtaining the convergence rates numerically

The inverse covariance matrix, V_t can be expressed as

$$\frac{1}{\Delta}\begin{pmatrix} \phi_{t,0} & \rho_{t,0} & 0 & \dots & 0 & 0 \\ \rho_{t,0} & \phi_{t,1} & \rho_{t,1} & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & & \\ 0 & 0 & 0 & \dots & \rho_{t,m-2} & \phi_{t,m-1} \end{pmatrix} \tag{5.9}$$

where $\phi_{t,j} = (1/\Delta)\{g_{t,j-1}+g_{t,j}c_{t,j}^2-g_{t,j}d_{t,j}a''_{t,j}\Delta-2g'_{t,j}d_{t,j}c_{t,j}+(1/2)g''_{t,j}d_{t,j}^2\}$ and $\phi_{t,j} = (1/\Delta)(-g_{t,j}c_{t,j}+g'_{t,j}d_{t,j})$. The usual procedure is to define a matrix $A = I -$ diag $(V_{t,11}^{-1},\dots,V_{t,mm}^{-1})V_t$, and $D_m = \det(A-\lambda I)$, that is

$$D_m = \begin{vmatrix} 0 & -\frac{\rho_{t,0}}{\phi_{t,0}} & 0 & \dots & 0 \\ -\frac{\rho_{t,0}}{\phi_{t,1}} & 0 & -\frac{\rho_{t,1}}{\phi_{t,1}}0 & \dots & 0 \\ \ddots & \ddots & \ddots & & \\ 0 & 0 & 0 & \dots & 0 \end{vmatrix}.$$

Suppose we let $a_{ij} = \frac{\rho_{t,i}}{\phi_{t,j}}$, we obtain

$$D_{m+2} = -\lambda D_{m+1} - a_{m,m+1}a_{m,m}D_m,$$

where $D_0 = 1$, $D_1 = -\lambda$ and $D_2 = \lambda^2 - a_{00}a_{01}$. The usual technique is to apply $A(t) = \sum_{m=0}^{\infty} D_m t^m$ so that one can use a Chebyshev polynomial of second-order. However, we now have the element $a_{m,m}a_{m,m+1}$ present in the summation instead of a constant. We are able to obtain constant expressions for each of the $a_{m,m}a_{m,m+1}$ terms as we can express each term with respects to the fixed points, or perform backwards recursion on the original equation. However, though the individual expressions can be evaluated, they still vary as m varies and we are therefore unable to take them out of the summation. It must also be noted that the a_{ij}'s are only available for the states considered and can therefore not be summed from $(m+1)$ to infinity.

An alternative method is to obtain the eigenvalues by bisection, see Golub and van Loan, (section 8.5 1996) for a comprehensive treatment. An algorithm for computing the eigenvalues is given in appendix 5.6.2.

5.5 Conclusion

The method based on block sampling was presented in Section 5.1, using either fixed or random block sizes within each sweep. Examples and a comparison between blocking strategies was presented in Section 5.2. The Chapter then considered a theoretical analysis of the efficiency of the sampler in Section 5.4. For most models, however, this was shown not to be possible so that a method for computing numerical convergence rates was required. We now move onto the question of sampling the parameters in Chapter 6.

5.6 Appendix

5.6.1 Obtaining the proposal for block sampling

Consider a block of size m $(1 \leq m \leq M)$ starting with observation $y^*_{t,k}$, $(1 \leq k \leq M)$. So we have a block $y^*_{t(k,m)} = (y^*_{t,k}, y^*_{t,k+1}, \ldots, y^*_{t,k+m-1}) = (w_1, \ldots, w_m) = w$ with neighbours $y^*_{t,k-1} = w_0$ and $y^*_{t,k+m} = w_{m+1}$. To deal with the non-Gaussian element of the likelihood, we propose to perform a second order Taylor expansion of $l(w)$ around a fixed value of w, denoted by $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_m)$. This gives

$$l(w) \approx c + x'u_t - \frac{1}{2}x'V_t x = c' - \frac{1}{2}(x - V_t^{-1}u_t)'V_t(x - V_t^{-1}u_t)$$

where $x = (w - \hat{w})$ and

$$\begin{aligned} u_t &= \frac{\partial l(\hat{w})}{\partial w} = \{u_{t,j}\} \quad \text{for } j = 1, \ldots m \\ V_t &= -\frac{\partial^2 l(\hat{w})}{\partial w \partial w'} = \{V^t_{ij}\} \quad \text{for } i, j = 1, \ldots, m. \end{aligned}$$

This shows that $l(w)$ is approximately the log of a multivariate-Gaussian density with mean $\tilde{\mu} = \hat{w} + V_t^{-1}u_t$ and variance matrix $\tilde{\Sigma} = V_t^{-1}$. To obtain expressions for these parameters, define

$$a_{t,j} = a(y^*_{t,k+j}) = a(w_{j+1}) \tag{B.5.1}$$

$$g_{t,j} = \{b^2(y^*_{t,k+j})\}^{-1} = \{b^2(w_{j+1})\}^{-1} \tag{B.5.2}$$

$$d_{t,j} = y^*_{t,k+j+1} - (y^*_{t,k+j} + a_{t,j}\Delta) = w_{j+2} - (w_{j+1} + a_{t,j}\Delta) \tag{B.5.3}$$

$$c_{t,j} = 1 + a'_{t,j}\Delta \tag{B.5.4}$$

for $j = -1, 0, \ldots, m-1$. Let $a'_{t,j}$ and $a''_{t,j}$ denote the first and second derivatives of $a(w_{j+1})$ with respect to w_{j+1}, and define $g'_{t,j}$ and $g''_{t,j}$ similarly. Consider a given element, w_{j+1} $(j = 0, \ldots, m-1)$ of the block. The only terms in $l(w)$ involving w_{j+1} are

$$\begin{aligned} l_{t,j} &= \frac{1}{2}\log\left(\frac{g_{t,j-1}}{\Delta}\right) - \frac{g_{t,j-1}}{2\Delta}[w_{j+1} - (w_j + a_{t,j-1}\Delta)]^2 \\ l_{t,j+1} &= \frac{1}{2}\log\left(\frac{g_{t,j}}{\Delta}\right) - \frac{g_{t,j}}{2\Delta}[w_{j+2} - (w_{j+1} + a_{t,j}\Delta)]^2 \end{aligned}$$

which implies that

$$\frac{\partial l(w)}{\partial w_{j+1}} = \frac{\partial l_{t,j}}{\partial w_{j+1}} + \frac{\partial l_{t,j+1}}{\partial w_{j+1}} = -\frac{1}{\Delta}\left(g_{t,j-1}d_{t,j-1} - g_{t,j}d_{t,j}c_{t,j} + \frac{1}{2}g'_{t,j}d^2_{t,j} - \frac{\Delta g'_{t,j}}{2g_{t,j}}\right).$$

For the second derivatives,

$$\frac{\partial^2 l}{\partial w_{j+1}^2} = -\frac{1}{\Delta}\left[g_{t,j-1} + g_{t,j}c_{t,j}^2 - g_{t,j}d_{t,j}a''_{t,j}\Delta - 2g'_{t,j}d_{t,j}c_{t,j} + \frac{1}{2}g''_{t,j}d_{t,j}^2 - \frac{\{g_{t,j}g''_{t,j} - (g'_{t,j})^2\}\Delta}{2g_{t,j}^2}\right].$$

The only term involving w_{j+1} and w_{j+2} is $l_{t,j+1}$, implying that

$$\frac{\partial^2 l(w)}{\partial w_{j+1}\partial w_{j+2}} = \frac{\partial^2 l_{t,j+1}}{\partial w_{j+1}\partial w_{j+2}} = -\frac{1}{\Delta}(g'_{t,j}d_{t,j} - g_{t,j}c_{t,j})$$

In summary, the proposal density $q(w|w_0, w_{m+1})$ obtained from the second-order Taylor expansion of $l(w)$ is Gaussian with mean

$$\tilde{\mu} = \hat{w} + \tilde{\Sigma}u_t \tag{B.5.5}$$

and variance matrix

$$\tilde{\Sigma} = V_t^{-1} \tag{B.5.6}$$

where u_t is an $m \times 1$ vector with elements

$$u_{t,j+1} = -\frac{1}{\Delta}\left(g_{t,j-1}d_{t,j-1} - g_{t,j}d_{t,j}c_{t,j} + \frac{1}{2}g'_{t,j}d_{t,j}^2 - \frac{g'_{t,j}\Delta}{2g_{t,j}}\right) \tag{B.5.7}$$

for $j = 0, \ldots, m-1$, and V_t is an $m \times m$ matrix with elements

$$\begin{aligned} V_{j+1,j+1}^t &= \frac{1}{\Delta}\left[g_{t,j-1} + g_{t,j}c_{t,j}^2 - g_{t,j}d_{t,j}a''_{t,j}\Delta - 2g'_{t,j}d_{t,j}c_{t,j} + \frac{1}{2}g''_{t,j}d_{t,j}^2\right. \\ &\qquad \left. - \frac{\{g_{t,j}g''_{t,j} - (g'_{t,j})^2\}\Delta}{2g_{t,j}^2}\right] \end{aligned} \tag{B.5.8}$$

$$V_{j+1,j+2}^t = V_{j+2,j+1}^t = \frac{1}{\Delta}(-g_{t,j}c_{t,j} + g'_{t,j}d_{t,j}) \tag{B.5.9}$$

and $V_{j+1,j+s}^t = 0$ for all j when $s > 2$. All quantities in (B.5.7)–(B.5.9) are evaluated at $w = \hat{w}$. It may be advisable to replace the $N(\tilde{\mu}, \tilde{\Sigma})$ distribution by a multivariate$-t$ density, or some other heavy tailed multivariate distribution in some applications.

5.6.2 An algorithm for computing the numerical convergence rates

Suppose $V_{t,r}$ represents the leading r-by-r principal sub-matrix V_t in 5.9. Define the polynomials $p_r(\lambda) = \det(V_{t,r} - \lambda I)$, for $r = 0, \ldots, m-1$. A determinantal expansion shows that

$$\begin{aligned} p_0(\lambda) &= 1 \\ p_1(\lambda) &= \phi_1 - \lambda \\ p_2(\lambda) &= (\phi_1 - \lambda)(\phi_2 - \lambda) - \rho_1^2 \\ &\vdots \\ p_r(\lambda) &= (\phi_{r-1} - \lambda)p_{r-1}(\lambda) - \rho_{r-1}^2 p_{r-2}(\lambda) \\ &\vdots \\ p_{m-1}(\lambda) &= p(\lambda). \end{aligned}$$

Now, $p_{m-1}(\lambda)$ can be evaluated in $O(m)$ calculations. Suppose $p_{m-1}(a)p_{m-1}(b) < 0$ and $a < b$. We can use the following iteration:

While $|a - b| > \epsilon(|a| + |b|)$, $c = \frac{a+b}{2}$

If $p_{m-1}(c)p_{m-1}(a) < 0$, $b = c$

Else $a = c$.

This algorithm is guaranteed to terminate with $(a+b)/2$ an approximate eigenvalue of V_t. The iteration converges linearly in that the error is approximately halved at each step. We can also use the Sturm sequence property to compute the k'th largest eigenvalue of V_t. If V_t has no zero sub-diagonal entries, then the eigenvalues of $V_{t,r-1}$ strictly separate the eigenvalues of $V_{t,r}$, that is

$$\lambda_r(V_{t,r}) < \lambda_{r-1}(V_{t,-1}) < \lambda_{r-1}(V_{t,r}) < \ldots < \lambda_2(V_{t,r}) < \lambda_1(V_{t,r-1}) < \lambda_1(V_{t,r}).$$

In addition, if we denote the number of sign changes in the sequence

$$\{p_0(\lambda), p_1(\lambda), \ldots, p_{m-1}(\lambda)\}$$

by $a(\lambda)$, then $a(\lambda)$ equals the number of eigenvalues of V_t that are less than λ. Note that if $p_r(\lambda) = 0$, then $p_r(\lambda)$ has the opposite sign of $p_{r-1}(\lambda)$.

CHAPTER 6

PARAMETER ESTIMATION

Having imputed values for the auxiliary states, the next step in the algorithm is to obtain an estimate of the model parameters, by sampling from the posterior distribution, $f(\theta|Y, Y^*)$. A generic procedure is outlined in Section 6.1. In cases where, for example, a non-conjugate prior is being considered, or when the posterior cannot be written down explicitly, other algorithms, (including Adaptive Rejection sampling or the Accept/Reject algorithm), can be used. Issues concerning the simulation of typical financial data and choosing initial parameter values are covered in Section 6.2. An illustration using simulated data for the CIR model, in order to assess the performance of our method, is presented. Updating the entire block of M latent states within each pair of observations, (instead of sampling in smaller block sizes), is discussed in Section 6.3, before considering additional models such as the Affine CIR and general parametric model in Section 6.4. The hyperbolic diffusion, which has a heavier tailed distribution than those already considered and has been shown to provide a good fit to many financial time series, is considered in 6.5 and the Chapter concludes with Section 6.6.

6.1 SAMPLING THE PARAMETERS

Given the observed and simulated states, a Bayesian analysis of the parameters is straightforward. Let $z = (y_1, y^*_{1,1}, \ldots, y^*_{1,M}, y_2, y^*_{2,1}, \ldots, y^*_{T-1,M}, y_T)' = (z_1, \ldots z_n)'$ where $n = (T-1)(M+1)+1$ is the length of z. Here both observed states Y and simulated auxiliary states Y^* in z are treated as the observed data.

To establish a general procedure for these models, consider the following two cases

$$f(z, y_1; \theta) = \prod_{t=2}^{n} f(z_t|z_{t-1}; \theta)\, f(y_1; \theta) \tag{6.1}$$

$$f(z|y_1; \theta) = f(z_2, \ldots z_n|y_1; \theta) = \prod_{t=2}^{n} f(z_t|z_{t-1}; \theta). \tag{6.2}$$

In (6.1), the unconditional density of y_1 is tractable, but an assumption of stationarity must be imposed to obtain the unconditional distribution. If we model y_1 such that $f(y_1; \theta)$ does not depend on θ, then we obtain no information on θ and so whether we analyse equation (6.1) or (6.2) on making inferences about θ, both are equivalent. An example of this is where y_1 is given a diffuse distribution; this is often done in non-stationary time series analysis, see for example, Hendry, (1995, Section 5.3).

In general, $f(y_1; \theta)$ is intractable and so we will typically work with the conditional likelihood (6.2) to analyse the properties of θ. The impact of the initial distribution is one term relative to the $n-1$

conditional likelihoods and so it can be argued that though $f(y_1;\theta)$ contains information on θ, the effect from omitting the term is negligible. This can be easily formalised if z_t is stationary or mixing. Hence for the remaining models, we work with

$$f(z|y_1;\theta) \propto \prod_{t=2}^{n} f(z_t|z_{t-1};\theta),$$

where each $f(z_t|z_{t-1};\theta)$ is a Gaussian density, with mean $z_{t-1} + a(z_{t-1};\theta)\Delta$ and variance $b^2(z_{t-1};\theta)\Delta$. Note that these models are Markov chains, so that equation (6.2) is particularly simple. Given this specification, we can estimate the parameters within a Bayesian framework.

6.1.1 Generic algorithms for models with a multiplicative volatility structure and constant coefficients

Suppose that for the homoscedastic cases, we assume an inverse gamma $(\frac{p}{2}, \frac{1}{2\,S_0\,p})$ prior for σ^2, where $S_0 = c\,\Delta$, p and c representing constants, (typically 10 and 0.001 respectively). Let $\theta_{\backslash\sigma^2}$ represent the remaining parameters, excluding σ^2. Then we know that

$$f(\sigma^2|z;\theta_{\backslash\sigma^2}) \propto f(z|\sigma^2,\theta_{\backslash\sigma^2})\, f(\sigma^2|\theta_{\backslash\sigma^2}).$$

If we let $b^2(z_{t-1}) = \sigma^2 c(z_{t-1})$, then

$$\log f(z|\sigma^2,\theta_{\backslash\sigma^2}) = -\frac{(n-2)}{2}\log 2\pi\Delta - \frac{(n-2)}{2}\log\sigma^2 - \frac{1}{2\sigma^2\Delta}\sum_{t=2}^{n}\frac{\{z_t - z_{t-1} - a(z_{t-1})\Delta\}^2}{c(z_{t-1})}$$

and

$$f(\sigma^2|\theta_{\backslash\sigma^2}) \propto (\sigma^2)^{-(\frac{p}{2}+1)}\exp(-\frac{S_0\,p}{2\sigma^2})\,(S_0)^{\frac{p}{2}}\,.$$

So,

$$\log f(\sigma^2|z;\theta_{\backslash\sigma^2}) \propto \text{ constant } -\left(\frac{n+p}{2}\right)\log\sigma^2 - \frac{1}{2\sigma^2}\left[\Delta^{-1}\sum_{t=2}^{n}\frac{\{z_t - z_{t-1} - a(z_{t-1})\Delta\}^2}{c(z_{t-1})} + \frac{S_0 p}{\Delta}\right]$$

implying that

$$f(\sigma^2|z;\theta_{\backslash\sigma^2}) \equiv \mathcal{X}^{-2}_{n+p}\Delta^{-1}\left[\sum_{t=2}^{n}\frac{\{z_t - z_{t-1} - a(z_{t-1})\Delta\}^2}{c(z_{t-1})} + S_0 p\right]. \tag{6.3}$$

To obtain the posterior distribution for the drift coefficient there are several procedures we could follow, see for example Gelman, Carlin, Stern, and Rubin (1995). In the exposition which follows, each model will be treated separately, specifying a prior for the parameters $\theta_{\backslash\sigma^2}$. We first analyse the case where we have normal data with a conjugate prior distribution. This will provide a good framework from which we can build for the more complicated models where conjugacy is typically impossible or where a simple conjugate form is not realistic.

To obtain the posterior for the drift coefficient, suppose we look at the univariate case, where the

parameter of interest is $\theta_{\backslash\sigma^2}$. The log-likelihood for the process is given by

$$\log f(z|\sigma^2, \theta_{\backslash\sigma^2}) \propto c - \frac{1}{2\Delta}\sum_{t=2}^{n}\left\{\frac{z_t - z_{t-1} - a(z_{t-1})\Delta}{b^2(z_{t-1})}\right\}.$$

Now suppose we consider the likelihood as an exponential of a quadratic form in μ; then the family of conjugate prior densities is

$$f(\theta_{\backslash\sigma^2}) \propto \exp\left\{-\frac{1}{2\tau_0^2}(\theta_{\backslash\sigma^2} - \theta_{0\backslash\sigma^2})^2\right\} \tag{6.4}$$

that is, $\theta_{\backslash\sigma^2} \sim N(\theta_{0\backslash\sigma^2}, \tau_0^2)$. A conjugate prior density implies that the posterior distribution for $\theta_{\backslash\sigma^2}$ is Normal as it is the exponential of a quadratic form.

For the multivariate case, suppose we consider the conjugate prior density for say, $\theta_{\backslash\sigma^2} = (\theta'_{1\backslash\sigma^2}, \theta'_{2\backslash\sigma^2})'$ by

$$f(\theta_{\backslash\sigma^2}) = \exp\left\{-\frac{1}{2}(\theta_{\backslash\sigma^2} - \theta_{0\backslash\sigma^2})'\Gamma(\theta_{\backslash\sigma^2} - \theta_{0\backslash\sigma^2})\right\},$$

implying that $(\theta'_{1\backslash\sigma^2}, \theta'_{2\backslash\sigma^2})' \sim N(\theta_{0\backslash\sigma^2}, \Gamma^{-1})$. We express the log-likelihood of the observations as

$$\log f(\theta_{\backslash\sigma^2}|z;\sigma^2) \propto c - \frac{1}{2\sigma^2\Delta}(z - X\theta_{\backslash\sigma^2}\Delta)'(z - X\theta_{\backslash\sigma^2}\Delta).$$

Then the log posterior can be expressed as

$$\begin{aligned}\log f(\theta_{\backslash\sigma^2}|z;\sigma^2) &= c - \frac{1}{2\sigma^2\Delta}(z - X\theta_{\backslash\sigma^2}\Delta)'(z - X\theta_{\backslash\sigma^2}\Delta) - \frac{1}{2}(\theta_{\backslash\sigma^2} - \theta_{0\backslash\sigma^2})'\Gamma^{-1}(\theta_{\backslash\sigma^2} - \theta_{0\backslash\sigma^2}) \\ &= c' - \frac{1}{2}(v - V^{-1}u)'V(v - V^{-1}u)\end{aligned}$$

where

$$u = \left(\frac{X'z}{\sigma^2} + \Gamma\theta_{0\backslash\sigma^2}\right) \tag{6.5}$$

$$V = \left(\frac{X'X\Delta}{\sigma^2} + \Gamma\right) \tag{6.6}$$

$$v = \theta_{\backslash\sigma^2}. \tag{6.7}$$

Hence $\log f(\theta_{\backslash\sigma^2}|z;\sigma^2) \sim N(V^{-1}u, V^{-1})$.

6.1.2 The Accept/Reject algorithm and Adaptive Rejection Sampling

The Accept/Reject algorithm will be used frequently throughout this Section. For further references on the method, see Rubinstein (1981) and Ripley, (1987, p. 60). The method, (due to von Neumann 1951) generates random variates from a target distribution which may be difficult to sample from. This is achieved by generating a candidate value Y from a sampling distribution which is easier to sample from. Values are retained with a probability depending on Y.

In practice, suppose we wish to sample from a p.d.f. $f(x) = \frac{h(x)}{\int h(x)\,dx}$, but have a way to sample from the p.d.f. $g(x)$. Suppose further that we can find a constant M, where $0 < M < \infty$ such that $h \leq Mg$

for all x. The general algorithm proceeds as follows

1. Generate Y from p.d.f. g.
2. Generate U from $\mathbb{U}(0,1)$.
3. If $U \leq \frac{h(Y)}{Mg(Y)}$, then return X=Y.
4. If rejected, return to step 1.

The probability that Y is accepted is

$$P(Y \text{ is accepted}) = \frac{\int h(x)\,dx}{M}.$$

As the number of trials before a Y is accepted has a geometric distribution with mean M, the algorithm works best when M is small.

If the posterior is not available in closed form, or one needs to estimate a model involving non-conjugacy which may be computationally expensive, we can sample from this posterior inside an M-H algorithm, by using generic methods such as *Adaptive Rejection Metropolis* sampling (ARMS), (see Gilks, Best, and Tan 1995), which involves sampling one parameter at a time.[1] This procedure will be used for the Affine CIR and general parametric models presented in Section 6.4, and will be briefly outlined here.

Adaptive Rejection Sampling, (or ARS, see Gilks and Wild 1992), is a method of rejection sampling from any univariate log-concave density $f(x)$, which does not require the mode of the density to be known. It is adaptive in the sense that the envelope and the squeezing functions converge to the density as sampling proceeds. We let, for example, $h_1(x)$ and $h_2(x)$ denote the piecewise exponential functions, constructed to sample from the log of the target density, where $h(x) = logf(x)$, so that $h_1(x) \geq h(x) \geq h_2(x)$ for all x in the domain of $f(x)$. It also does not require the bounding functions to be determined in advance. If a point is rejected, the envelope and squeezing functions are updated by introducing an additional tangent to the log density, decreasing the probability of rejection of subsequent sampled points. An advantage of this method, is that samples can be drawn with few evaluations of the density, so that the method is typically used for models which are computationally expensive to evaluate.

For densities $f(x)$ which are not log-concave, (as may be the case for many of the diffusions we consider), ARS is not appropriate as the function $h(x)$ may not be an envelope to the true log density. Gilks, Best, and Tan (1995) propose an M-H step to deal with the non-concavity in the log density. The proposed point is accepted or rejected, depending on the value of the ratio of the density and envelope, evaluated at both the previously sampled value and the new point. If a value is rejected, then the previously sampled point is retained as the new value. The algorithm is most efficient for densities which are actually log-concave. Fewer rejections in the M-H step imply fewer point-evaluations of the log density, reducing computer time. The probability of rejection can be reduced by choosing good initialisation values for the points used to construct the tangents in the domain of $f(x)$. Gilks, Best, and Tan advocate basing initialisation points on the value of the previous iteration given that it is independent of the new proposed value.

[1] since this method can only be applied univariately.

6.1.3 The posteriors for the parameters of the OU process

For the drift coefficient of the OU process, a conjugate analysis (as described above) can be carried out easily. We will present an alternative approach based on a non-conjugate prior distribution. Though choosing a non-conjugate prior distribution may make interpretations of the posterior less transparent and computation more difficult, in this case it is easy to work with and the prior can be chosen to ensure that the parameter is within its region of stationarity. The volatility coefficient is obtained using a conjugate prior distribution.

Assuming a prior of $\chi_p^{-2} S_0 p$ for σ^2, where typically $p = 10$ and $S_0 = c \times \Delta$, where c is a constant (typically 0.001) that does not depend on p, we achieve the posterior density

$$\sigma^2 | z; \mu \sim \chi^{-2}_{n+p} \Delta^{-1} \left[\sum_{t=2}^{n} \{ z_t - (1 + \mu\Delta) z_{t-1} \}^2 - z_1^2 (2\mu\Delta + \mu^2\Delta^2) + S_0 p \right]$$

where $n = (T-1) \times M + T$, the total amount of observed and simulated data and μ is the most recently updated value, over the iterations. Suppose we employ a beta prior on $(1+\phi)/2$, where $\phi = 1 + \mu\Delta$ with $\mathrm{E}(\mu) = \frac{1}{\Delta}\left(\frac{2\delta}{\delta+\gamma} - 2\right)$. Then the prior of μ is proportional to

$$f(\mu|\sigma^2) \sim \frac{\Gamma(\delta+\gamma)}{\Gamma(\delta)\Gamma(\gamma)} \left(\frac{\mu\Delta + 2}{2}\right)^{\gamma-1} \left(\frac{-\mu\Delta}{2}\right)^{\delta-1} \frac{2}{\Delta}.$$

So

$$\begin{aligned}
\log f(\mu|z;\sigma^2) &\propto \log f(z|\mu,\sigma^2) + \log f(\mu|\sigma^2) \\
&= \text{constant} - \frac{1}{2\sigma^2\Delta} \sum_{t=2}^{n} \{ z_t - (1+\mu\Delta) z_{t-1} \}^2 + \frac{z_1^2}{2\sigma^2\Delta} (2\mu\Delta + \mu^2\Delta^2) \\
&\quad - \frac{1}{2} \log(2\mu\Delta + \mu^2\Delta^2) + (\gamma - 1)\log(\mu\Delta + 2) + (\delta - 1)\log(-\mu\Delta) \\
&= \text{constant} + l(\mu) + p(\mu),
\end{aligned}$$

where $p(z_1; \mu, \sigma^2)$ has unconditional distribution $N(0, \frac{\sigma^2}{1-\phi^2})$. The first term is Gaussian but the second term, due to the choice of the prior is non-Gaussian. Suppose we define $s_1^2 = \sum_{t=2}^{n} z_{t-1}^2$ and $s_{0,1} = \sum_{t=2}^{n} z_t z_{t-1}$, then the first term is in fact

$$l(\mu) = \text{constant} - \frac{\Delta s_1^2}{2\sigma^2} \left\{ \mu - \left(\frac{s_{0,1}}{s_1^2} - 1 \right) \right\}^2,$$

which is normal with mean $\{(s_{0,1}/s_1^2) - 1\}$ and variance $\{\sigma^2/(\Delta s_1^2)\}$. The posterior is non-conjugate and so a first-order approximation of $p(\mu)$ is taken and an Accept/Reject algorithm is performed to obtain an estimate for μ. Now since $\frac{1}{2}\log(-2\mu\Delta - \mu^2\Delta^2) = \frac{1}{2}\log(\mu\Delta + 2) + \frac{1}{2}\log(-\mu\Delta)$, redefining δ as $\delta + (1/2)$ and doing the same for γ,

$$p(\mu) = (\gamma - 1)\log(\mu\Delta + 2) + (\delta - 1)\log(-\mu\Delta).$$

Applying a first-order Taylor approximation to $p(\mu)$ we obtain

$$\begin{aligned} p(\mu) &\approx p(\hat{\mu}) + (\mu - \hat{\mu})p'(\hat{\mu}) \\ &= (\gamma - 1)\log(\hat{\mu}\Delta + 2) + (\delta - 1)\log(-\hat{\mu}\Delta) + (\mu - \hat{\mu})\left\{\frac{(\gamma - 1)\Delta}{\hat{\mu}\Delta + 2} + \frac{\Delta(\delta - 1)}{\hat{\mu}\Delta}\right\}. \end{aligned}$$

This means that we can bound the true posterior density from above by the following density

$$\log g(\mu|\sigma^2, y) = c - \frac{\Delta\sum_{t=2}^{n} z_{t-1}^2}{2\sigma^2}\left[\mu - \left\{\frac{\sum_{t=2}^{n} z_t z_{t-1}}{\sum_{t=2}^{n} z_{t-1}^2} - 1 + \frac{\sigma^2(\gamma - 1)}{\sum_{t=2}^{n} z_{t-1}^2(\hat{\mu}\Delta + 2)}\right. \tag{6.8}$$

$$\left. + \frac{\sigma^2(\delta - 1)}{\sum_{t=2}^{n} z_{t-1}^2\hat{\mu}\Delta}\right\}\right]^2. \tag{6.9}$$

Values of μ are drawn from (6.9), and we can apply the result on acceptance rates outlined in Shephard and Pitt, (1997, Section 2 and given in Theorem 6.1 below), which states that the acceptance probability for making proposals for the parameters is accepted with probability $\exp\{p(\mu) - p(\hat{\mu}) - p'(\hat{\mu})(\mu - \hat{\mu})\}$. For the Accept/Reject algorithm, if a value is rejected, another candidate value in the stationary region is proposed using (6.9) until acceptance, producing variates from $\log f(\mu|z; \sigma^2)$. As we have restricted values of μ to lie in the region of stationarity, the proposals may often be rejected. To overcome this, one could employ the probability integral transform theorem proposed by Feller (1971) which enables sampling from a truncated distribution, which for the normal is denoted by $TN(\cdot)$, without rejection. A proposal μ^* of μ is drawn from

$$TN_{[-(2/\Delta),0]}\{\tilde{\mu},\ \mathrm{var}(\tilde{\mu})\},$$

where

$$\begin{aligned} \tilde{\mu} &= \frac{s_{0,1}}{s_1^2} - 1 + \frac{\sigma^2(\gamma - 1)}{s_1^2(\hat{\mu}\Delta + 2)} + \frac{\sigma^2(\delta - 1)}{s_1^2\hat{\mu}\Delta} \text{ and} \\ \mathrm{var}(\tilde{\mu}) &= \frac{\Delta s_1^2}{\sigma^2}. \end{aligned}$$

6.1.4 The posterior in the quadratic drift model

The posterior of the volatility coefficient σ^2 can be obtained using the results in Section 6.1.1. The log-likelihood for the observations is given by

$$\log f(z|\sigma^2, \mu) = -\frac{(n-2)}{2}\log 2\pi\Delta - \frac{(n-2)}{2}\log\sigma^2 - \frac{1}{2\sigma^2\Delta}\sum_{t=2}^{n}\{z_t - z_{t-1} - \mu z_{t-1}^2\Delta\}^2.$$

Suppose we consider the likelihood as an exponential of a quadratic form in μ, as in 6.4, then

$$f(\mu) \propto \exp\left\{-\frac{1}{2\tau_0^2}(\mu - \mu_0)^2\right\},$$

that is, $\mu \sim N(\mu_0, \tau_0^2)$. Hence

$$\begin{aligned}\log f(\mu|z;\sigma^2) &= c - \frac{1}{2\sigma^2\Delta}\sum_{t=2}^{n}\{z_t - z_{t-1} - \mu z_{t-1}^2\Delta\}^2 - \frac{1}{2\tau_0^2}(\mu-\mu_0)^2 \\ &= c - \frac{1}{2}\left(\frac{\sum_{t=2}^n z_{t-1}^4\Delta}{\sigma^2} + \frac{1}{\tau_0^2}\right)\left\{\mu^2 - 2\mu\left(\frac{\frac{\sum_{t=2}^n z_t z_{t-1}^2}{\sigma^2} - \frac{\sum_{t=2}^n z_{t-1}^3}{\sigma^2} - \frac{\mu_0}{\tau_0^2}}{\frac{\sum_{t=2}^n z_{t-1}^4\Delta}{\sigma^2} + \frac{1}{\tau_0^2}}\right)\right\}.\end{aligned}$$

This implies that $f(\mu|z) \propto \exp\left\{-\frac{1}{2\tau_1^2}(\mu-\mu_1)^2\right\}$, where on defining $s_{0,1}^2 = \sum_{t=2}^n z_t z_{t-1}^2$ and $s_1^r = \sum_{t=2}^n z_{t-1}^r$,

$$\mu_1 = \frac{\frac{s_{0,1}^2}{\sigma^2} - \frac{s_1^3}{\sigma^2} - \frac{\mu_0}{\tau_0^2}}{s_1^4\Delta}\sigma^2 + \frac{1}{\tau_0^2},\ \tau_1^2 = \left(\frac{s_1^4\Delta}{\sigma^2} + \frac{1}{\tau_0^2}\right)^{-1}.$$

6.1.5 The posterior of the drift in the Vasicek model

The posterior of the volatility coefficient σ^2 can again be obtained using the results in 6.1.1. In this case, the log-likelihood for the observations is given by

$$\log f(z|\sigma^2,\alpha,\beta) = -\frac{(n-2)}{2}\log 2\pi\Delta - \frac{(n-2)}{2}\log\sigma^2 - \frac{1}{2\sigma^2\Delta}\sum_{t=2}^{n}(z_t - z_{t-1} - \alpha\Delta + \beta\Delta z_{t-1})^2.$$

We now consider the conjugate prior density for $\gamma = (\alpha', \beta')'$ as in (6.7). Suppose, for notational convenience that we define two matrices v and X by the following

$$\begin{aligned} v &= \begin{pmatrix} z_2 - z_1 \\ \vdots \\ z_n - z_{n-1} \end{pmatrix} \\ X &= \begin{pmatrix} 1 & -z_1 \\ \vdots & \vdots \\ 1 & -z_{n-1} \end{pmatrix}. \end{aligned}$$

Then the log posterior can be expressed as

$$\begin{aligned}\log f(\gamma|z;\sigma^2) &= c - \frac{1}{2\sigma^2\Delta}(v - X\gamma\Delta)'(v - X\gamma\Delta) - \frac{1}{2}(\gamma-\gamma_0)'\Gamma^{-1}(\gamma-\gamma_0) \\ &= c' - \frac{1}{2}(w - V^{-1}u)'V(w - V^{-1}u),\end{aligned}$$

where

$$\begin{aligned} u &= \left(\frac{X'v}{\sigma^2} + \Gamma\gamma_0\right) \\ V &= \left(\frac{X'X\Delta}{\sigma^2} + \Gamma\right) \text{ and} \\ w &= \gamma. \end{aligned}$$

We then obtain $\log f(\gamma|z;\sigma^2) \sim N(V^{-1}u, V^{-1})$.

6.1.6 The posterior for the drift and volatility of the CIR model

The log-likelihood for the observed and simulated values in the CIR model is given by

$$\begin{aligned}\log f(\log z|\alpha,\beta,\sigma^2) &= -\frac{(n-2)}{2}\log 2\pi\Delta - \frac{(n-2)}{2}\log\sigma^2 + \frac{1}{2}\sum_{t=2}^{n}\log z_{t-1} \\ &\quad -\frac{1}{2\sigma^2\Delta}\sum_{t=2}^{n}\left\{\sqrt{z_{t-1}}\left(\log z_t + \frac{\sigma^2\Delta}{2z_{t-1}} - \log z_{t-1} - \frac{\alpha\Delta}{z_{t-1}} + \beta\Delta\right)\right\}^2.\end{aligned}$$

Suppose we let $\psi = (\alpha', \beta')'$ and express its prior as

$$f(\psi) = \exp\left\{-\frac{1}{2}(\psi-\gamma_0)'\Gamma_0(\psi-\gamma_0)\right\}.$$

If we further denote

$$v = \begin{bmatrix} \sqrt{z_1}\left(\log z_2 + \frac{\sigma^2\Delta}{2z_1} - \log z_1\right) \\ \vdots \\ \sqrt{z_{n-1}}\left(\log z_n + \frac{\sigma^2\Delta}{2z_{n-1}} - \log z_{n-1}\right)\end{bmatrix}$$

$$\text{and } X = \begin{pmatrix} z_1^{-\frac{1}{2}} & -z_1^{\frac{1}{2}} \\ \vdots & \vdots \\ z_{n-1}^{-\frac{1}{2}} & -z_{n-1}^{\frac{1}{2}}\end{pmatrix},$$

then the log-posterior can be expressed as

$$\log f(\psi|z;\sigma^2) = c - \frac{1}{2\sigma^2\Delta}(v - X\psi\Delta)'(v - X\psi\Delta) - \frac{1}{2}(\psi-\gamma_0)'\Gamma_0(\psi-\gamma_0).$$

We thus obtain

$$\log f(\psi|z;\sigma^2) \sim N(V^{-1}u,\, V^{-1}),$$

where

$$u = \left(\frac{X'v}{\sigma^2} + \Gamma_0\gamma_0\right),\ V = \left(\frac{X'X\Delta}{\sigma^2} + \Gamma_0\right).$$

The posterior of the volatility coefficient as described in (6.3) is no longer conjugate, since the log process is being considered. We propose taking a first-order Taylor expansion of the non-conjugate part to bound $\log f(\sigma^2|z;\alpha,\beta)$ from above and hence obtain the posterior using an Accept/Reject algorithm.

Taking a prior of $S_0\, p \times$ inverse gamma $(\frac{p}{2}, \frac{1}{2})$ for σ^2, where $S_0 = c\,\Delta$ as before, the log-posterior is

$$\log f(\sigma^2|z;\alpha,\beta) = c' - \xi_1(\log\sigma^2) - \xi_2\sigma^{-2} - S_1\sigma^2,$$

where $\xi_1 = \left(\frac{n+p}{2}\right)$, $\xi_2 = \frac{1}{2\Delta}(\sum_{t=2}^{n} z_{t-1}d_{t-1}^2 + S_0\,p\,\Delta)$, $S_1 = \frac{\Delta}{8}\sum_{t=2}^{n} z_{t-1}^{-1}$ and $d_{t-1} = \log z_t - \log z_{t-1} -$

$\frac{\alpha\Delta}{z_{t-1}} + \beta\Delta$. Suppose we define $\lambda = \log\sigma^2$, then

$$\begin{aligned}\log f(\lambda|z;\alpha,\beta) &= \log f(\sigma^2|z;\alpha,\beta) + \log\left|\frac{\partial\sigma^2}{\partial\lambda}\right| \\ &= c' - (\xi_1 - 1)\lambda - \xi_2\exp(-\lambda) + p(\lambda), \end{aligned} \tag{6.10}$$

with $\xi_2 > 0$, and $p(\lambda)$ concave in λ, so that $p''(\lambda)$ is negative for all λ. Then the following theorem describes how we can sample λ from (6.10) using the Accept/Reject algorithm.

Theorem 6.1. *Suppose the log posterior of* $\lambda = $ *const* $+ l(\lambda) + p(\lambda)$, *where* $l(\lambda)$ *is the log of the density of* λ *when* $exp(\lambda)$ *has an inverse Gamma distribution with parameters* $(\xi_1 - 1, \xi_2)$, *where* $\xi_1 > 1$, $\xi_2 > 0$, *and* $p(\lambda)$ *is concave in* λ, *so that* $p''(\lambda)$ *is negative for all* λ. *Then we can sample from* $\lambda|z,\theta$, *where* θ *represents the remaining parameters in the model, by making suggestions*

$$exp(\lambda) \sim \Gamma^{-1}\{\xi_1 - p'(\hat{\lambda}) - 1, \xi_2\}, \tag{6.11}$$

where $\xi_1 - p'(\hat{\lambda}) > 1$, $\xi_2 > 0$ *and* $\hat{\lambda}$ *is an arbitrary fixed value of* λ. *These proposed values are accepted with probability* $\exp\{p(\lambda) - p(\hat{\lambda}) - p'(\hat{\lambda})(\lambda - \hat{\lambda})\}$.

Proof. In practice, the parameters of the distribution, $\xi_1 - p'(\hat{\lambda}) - 1$ and ξ_2 are positive. Since $p(\lambda)$ is concave in λ, bounding implies that we can use a standard Accept/Reject algorithm.

We can express the log posterior as

$$\begin{aligned}\log f(\lambda) &= c' + p(\lambda) - (\xi_1 - 1)\lambda - \xi_2\exp(-\lambda) \\ &\leq c' + p(\hat{\lambda}) + p'(\hat{\lambda})(\lambda - \hat{\lambda}) - (\xi_1 - 1)\lambda - \xi_2\exp(-\lambda) \\ &= c - \{\xi_1 - p'(\hat{\lambda}) - 1\}\lambda - \xi_2\exp(-\lambda) \\ &= \log g(\lambda) + \log M \end{aligned}$$

where $\log M = p(\hat{\lambda}) - \hat{\lambda}p'(\hat{\lambda}) - \{\xi_1 - 1 - p'(\hat{\lambda})\}\log\xi_2 + \log\Gamma\{\xi_1 - 1 - p'(\hat{\lambda})\}$ and $\log g = -\Gamma\{\xi_1 - 1 - p'(\hat{\lambda})\} + \{\xi_1 - 1 - p'(\hat{\lambda})\}\log(\xi_2) - \lambda\{\xi_1 - 1 - p'(\hat{\lambda})\} - \xi_2\exp(-\lambda)$. Here $g(\lambda)$ is the density of λ when $\exp(\lambda)$ follows an inverse Gamma distribution with parameters $\{\xi_1 - 1 - p'(\hat{\lambda}), \xi_2\}$. We can thus use the Accept/Reject algorithm (as described in Section 6.1.2), to sample σ^2. First, $\sigma^2 = \exp(\lambda)$ is drawn from the inverse Gamma distribution implied by $g(\lambda)$. This is accepted with probability $f/gM = \exp\{p(\lambda) - p(\hat{\lambda}) - p'(\hat{\lambda})(\lambda - \hat{\lambda})\}$. Here, $p(\lambda) = -S_1\exp(\lambda) = p'(\lambda)$. □

The expansion point $\hat{\lambda}$ should be chosen so that the acceptance probability is maximised, or alternatively so that M is minimised. This occurs when $\hat{\lambda}$ is the solution to the equation,

$$\frac{\partial\log M}{\partial\hat{\lambda}} = -\hat{\lambda}p''(\hat{\lambda}) + p''(\hat{\lambda})\log\xi_2 - \frac{\partial\log\Gamma\{\xi_1 - 1 - p'(\hat{\lambda})\}}{\partial\hat{\lambda}}p''(\hat{\lambda}) = 0.$$

Denoting the digamma function $\frac{\partial\log\Gamma()}{\partial\hat{\lambda}}$ by $\Psi()$, we obtain

$$\hat{\lambda}_{\text{opt}} = \log(b) - \Psi\{a - p'(\hat{\lambda}_{\text{opt}}) - 1\}.$$

Here, $\Psi\{\xi_1 - p'(\hat{\lambda}) - 1\} \approx \log\{\xi_1 - p'(\hat{\lambda})\}$, since $\xi_1 - p'(\hat{\lambda})$ is $O(n)$, see Abramowitz and Stegun, (1970,

equation 6.3.18, p. 259). Thus,

$$\hat{\lambda}_{\text{opt}} \approx \log(\xi_2) - \log\{\xi_1 - p'(\hat{\lambda}_{\text{opt}})\},$$

which is also the mode of $\log f(\sigma^2|z;\theta)$. The solution can be found iteratively by substituting the current value of $\hat{\lambda}$ until convergence is reached.

6.1.7 The posterior for the drift of the Hull-White model

For the Hull-White model, the log-likelihood for the observed and simulated values is given by

$$\begin{aligned}\log f(\log z|\alpha,\beta,\sigma^2) &= -\frac{(n-2)}{2}\log 2\pi\Delta - \frac{(n-2)}{2}\log\sigma^2 - (\gamma-1)\sum_{t=2}^{n}\log z_{t-1} \\ &\quad -\frac{1}{2\sigma^2\Delta}\sum_{t=2}^{n}\frac{\left\{\log z_t - \log z_{t-1} - \frac{\alpha\Delta}{z_{t-1}} + \beta\Delta + \frac{\sigma^2\Delta}{2}z_{t-1}^{2(\gamma-1)}\right\}^2}{z_{t-1}^{(2\gamma-1)}}.\end{aligned}$$

Suppose we let $\psi = (\alpha', \beta')'$, express the prior as

$$f(\psi) = \exp\left\{-\frac{1}{2}(\psi-\psi_0)\Psi(\psi-\psi_0)\right\},$$

and further denote

$$v = \begin{bmatrix} \log z_2 - \log z_1 + \frac{\sigma^2\Delta}{2z_1^{2(\gamma-1)}} \\ \vdots \\ \log z_n - \log z_{n-1} + \frac{\sigma^2\Delta}{2z_{n-1}^{2(\gamma-1)}} \end{bmatrix}$$

$$\text{and } X = \begin{pmatrix} -\frac{1}{z_1} & 1 \\ \vdots & \vdots \\ -\frac{1}{z_{n-1}} & 1 \end{pmatrix}.$$

Then

$$\log f(\gamma|z;\sigma^2) = c - \frac{1}{2\sigma^2\Delta}(v - X\psi\Delta)'(v - X\psi\Delta) - \frac{1}{2}(\psi-\psi_0)'\Psi(\psi-\psi_0).$$

Hence, if we use the same definitions for u and V as was defined in (6.7) with $\Psi = \Gamma$, we obtain

$$\log f(\psi|z;\sigma^2) \sim N(V^{-1}u,\, V^{-1}).$$

The posterior of the volatility coefficient as described in (6.3), is again non-conjugate due to the fact that we are considering the log of the process. We propose using the same procedure as was outlined for the CIR process, so that we take a first-order Taylor expansion of the non-conjugate part to upperbound $\log f(\sigma^2|z;\alpha,\beta,\gamma)$ and hence obtain the posterior using an Accept/Reject algorithm.

Taking the prior $p\, S_0\, \chi_p^{-2}$ for σ^2, where $S_0 = c\,\Delta$ and c is a constant; the log-posterior is

$$\log f(\sigma^2|z;\alpha,\beta,\gamma) = c' - \xi_1(\log\sigma^2) - \xi_2\sigma^{-2}g(\lambda),$$

where

$$\xi_1 = \left(\tfrac{n+p}{2}\right), \qquad \xi_2 = \tfrac{1}{2\Delta}\left(\sum_{t=2}^{n} z_{t-1}^{-2(\gamma-1)} d_{t-1}^2 + S_0\, p\, \Delta\right),$$
$$S_1 = \tfrac{\Delta}{8}\sum_{t=2}^{n} z_{t-1}^{2(\gamma-1)} \qquad d_{t-1} = \log z_t - \log z_{t-1} - \tfrac{\alpha\Delta}{z_{t-1}} + \beta\Delta + \tfrac{\sigma^2 \Delta z_{t-1}^{2(\gamma-1)}}{2}.$$

For the Hull-White process there is the additional parameter γ. Suppose we put a non-informative prior on γ, taking the form of a discrete uniform distribution between 0 and 2, denoted by $\mathbb{U}(0,2)$ with I support points. For computational simplicity, we consider only a discrete set of possible values for γ. To obtain the full conditional posterior of γ, we use

$$P(\gamma = \gamma_i | z; \alpha, \beta, \sigma^2) = \frac{f(z|\gamma = \gamma_i, \alpha, \beta, \sigma^2)\, P(\gamma = \gamma_i)}{\sum_{i=1}^{I} f(z|\gamma = \gamma_i, \alpha, \beta, \sigma^2)\, P(\gamma = \gamma_i)}. \tag{6.12}$$

The parameter I represents the number of mass points in the prior distribution of γ. The posterior mean is obtained by sampling from this discrete distribution. A similar strategy is adopted in Albert and Chib (1993). To implement the procedure, variates from a multinomial distribution need to be generated. We propose to use the *Direct* method, first proposed by Dagpunar, (1988, p. 168). It requires a set-up phase in which the cumulative probability vector π^* for the multinomial distribution is first generated.

Suppose $\pi_i = P(\gamma = \gamma_i | z; \alpha, \beta, \sigma^2)$. Consider a multinomial distribution with parameters $(1, P)$ where $\pi = (\pi_1, \ldots, \pi_k)$, $\pi_i > 0$ and $\sum_{i=1}^{k} \pi_i = 1$ for $1 \leq i \leq k$. Then the following algorithm can be used to generate random variates from the appropriate distribution.

1. Compute $\pi_i^* = \sum_{j=1}^{i} \pi_j$, for $i = 1, \ldots, k$.
2. Generate a random variate U from the uniform distribution $\mathbb{U}(0,1)$.
3. Return i if $\pi_{i-1}^* < U \leq \pi_i^*$, (where $\pi_0^* = 0$).

The direct method is most efficient if few or only one draw is needed for each π, as is the case in this application.

An additional computational issue to avoid overflow involves subtracting a constant from the log-likelihood of the states given the parameters. This additional term will cancel in equation (6.12).

6.2 Simulation considerations and experiments

6.2.1 Simulating the data

For the OU, Vasicek and CIR processes, the data generating process (DGP) is determined by the respective strong solutions. For the remaining models however, no strong solution is available. We propose to generate the observations using the Euler approximation employing a very fine discretisation. Given an initial observation y_0, we can simulate R values of the process using the following specification

$$y_t = y_{t-1} + a(y_{t-1})\Delta^\dagger + b(y_{t-1})\sqrt{\Delta^\dagger}\epsilon_t \tag{6.13}$$

where $\epsilon_t \sim N(0,1)$ and $\Delta^\dagger = \frac{1}{R}$. We then record the last value of the R observations, say y_1 and repeat the process T times, starting with y_1. We thus obtain T observations, where T is typically chosen as 500. A large value of R, (e.g 100,000) is chosen to minimise the bias induced by the Euler discretisation.

Here, R acts as a precision parameter; the larger R is, the closer the DGP used for the simulated data is to the true underlying process.

6.2.2 Initial values for the parameters

To start the M-H algorithm, initial values for the parameters are required. Various procedures can be implemented using the observed values of the process. For the OU, quadratic drift process, Vasicek and CIR models, the maximum likelihood estimates (based on $M = 0$) are used. For the Hull-White model, we fix a value of γ at 0.5 and the remaining estimates are set to their respective MLE's.

The Markov chain is initialised using a *burn-in* period, (usually set to $N = 200$ for most of the applications considered in this Section). The burn-in period may vary depending on how rapidly the chain mixes, requiring longer periods for slow mixing chains. Since the chain is assumed to be irreducible, (requiring that the chain be able to reach all parts of the state space), initial starting values will in principle, not affect the stationary distribution.

In MCMC methods, draws are used as a basis for making inference, by appealing to certain ergodic theorems for Markov chains, as discussed in Part I of the thesis. The sample average over a run of the Markov chain is used to approximate an expectation with respect to the equilibrium distribution of the chain. If the chain is *Harris recurrent*, (see Nummelin, 1984, Ch. 4), then convergence does not depend on the initial starting distribution. Geyer (1992) suggests that burn-in periods are unnecessary, though if extreme starting values are to be avoided, a burn-in length of one to two percent of the total number of iterations N, can be used. For a discussion on whether averaging over one long run of the chain allows valid inference to be made from MCMC results or whether multiple sequences should be used, see Geyer (1992) and Gelman and Rubin (1992a).

6.2.3 MCMC algorithms on the CIR process using simulated data

Data is simulated to represent typical weekly and daily financial data sets. Using an initial value of $y_0 = 1$, $T = 500$ observations are obtained using $\Delta^\dagger = 1$ for daily series and $\Delta^\dagger = 5$ for weekly data, using the strong solution of the process. Parameter values of 0.5, 0.2, and 0.05 are taken for α, β and σ^2 respectively, (which is high for financial data). Here, $\Delta^\dagger$ can be thought of as the time interval in the observed data and determines bias in the discretised time gap.

Figures 6.1–6.3 show paths, correlograms and histograms for the posteriors of α, β and σ^2, given the data Y. Top graphs have $M = 0$, middle graphs have $M = 1$ and bottom graphs have $M = 10$, with $\Delta^\dagger = 1$ and $N = 10,000$ in all simulations. In Figure 6.4, paths, correlograms and histograms are shown for the posterior of σ^2, where the data used has been generated with a discretisation step of $\Delta^\dagger = 5$. All four graphs show the clear advantage of using auxiliary variables in the estimation procedure; the most obvious changes being seen in Figure 6.4, due to the increase in the discretisation bias in generating the data. Bias decreases quickly and the autocorrelation in the sampler is low for all three parameters. The posterior means, covariances and correlations (in italics) are additionally shown in Table 6.1, using the data with $\Delta^\dagger = 5$. Here, the M-H algorithm was run with $M = 0, 1, 10$ ($\lambda = 3$), 20 ($\lambda = 5$) and 30 ($\lambda = 9$) for $N = 10,000$ iterations. The simulation inefficiency factors are very small for $M < 20$. For $M = 20$, the sampler drawing variables from the posterior is up to 14 times less efficient than if we were to draw independent random variables. This figure increases to 20 for $M = 30$. The correlograms in Figures 6.1–6.3 reinforce this point, dying down quickly when there are a small number of auxiliary

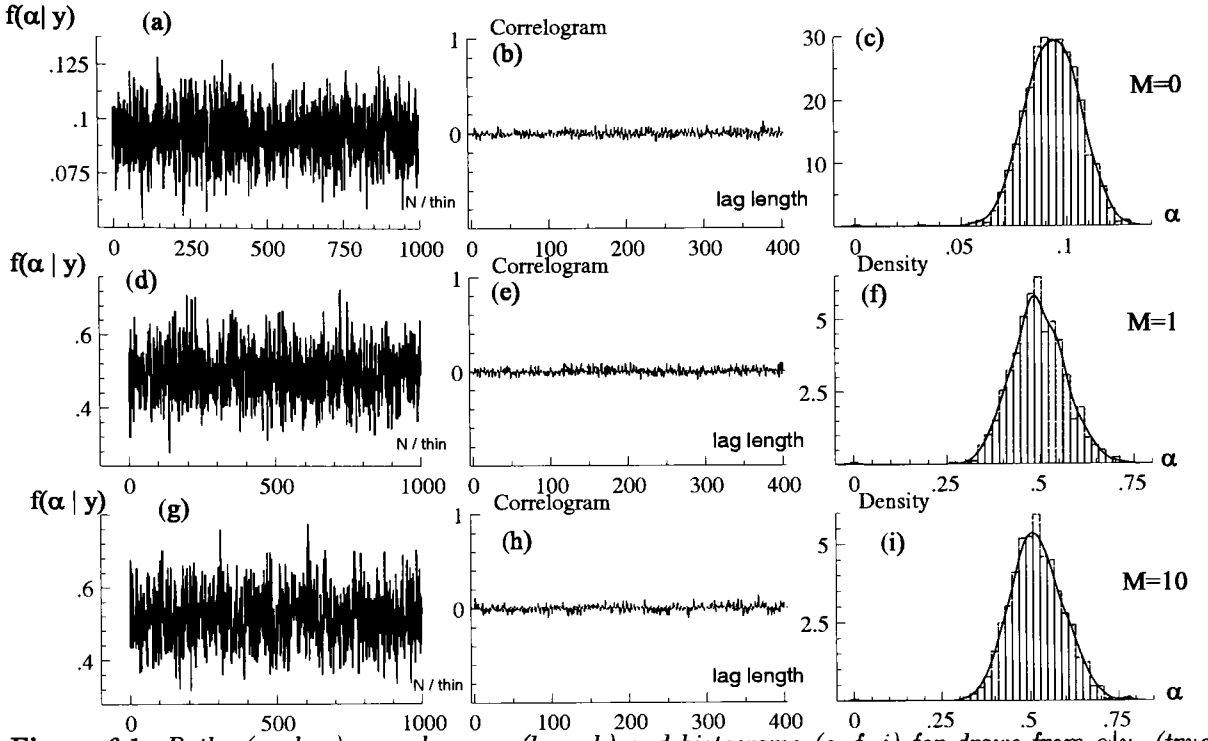


Figure 6.1: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for draws from $\alpha|y$, (true value 0.5). Top graph uses discretisation $M = 0$, middle graphs have $M = 1$ and bottom graphs have $M = 10$, $\lambda = 3$. $N = 10,000$ with every tenth simulation recorded and $\Delta^{\dagger} = 1$ in all simulations.*

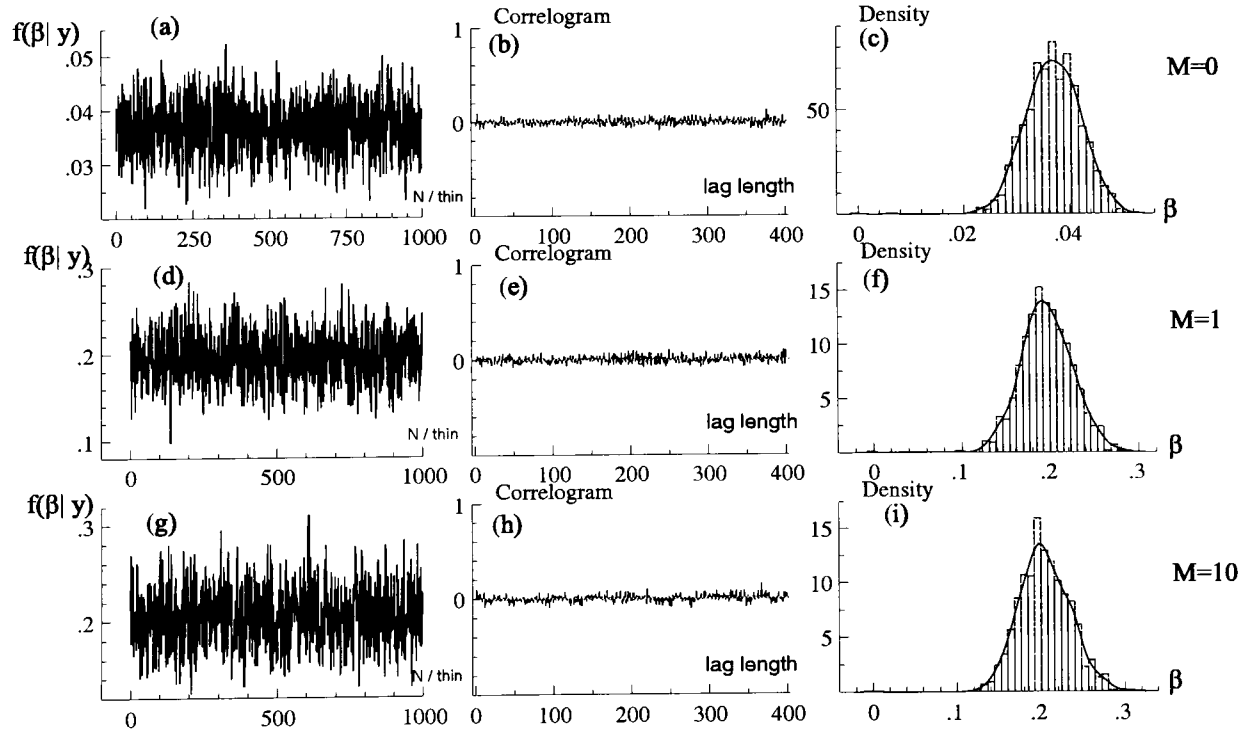


Figure 6.2: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for draws from $\beta|y$, (true value 0.2). Top graph uses discretisation $M = 0$, middle graphs have $M = 1$ and bottom graphs have $M = 10$, $\lambda = 3$. $N = 10,000$ with every tenth simulation recorded and $\Delta^{\dagger} = 1$ in all simulations.*

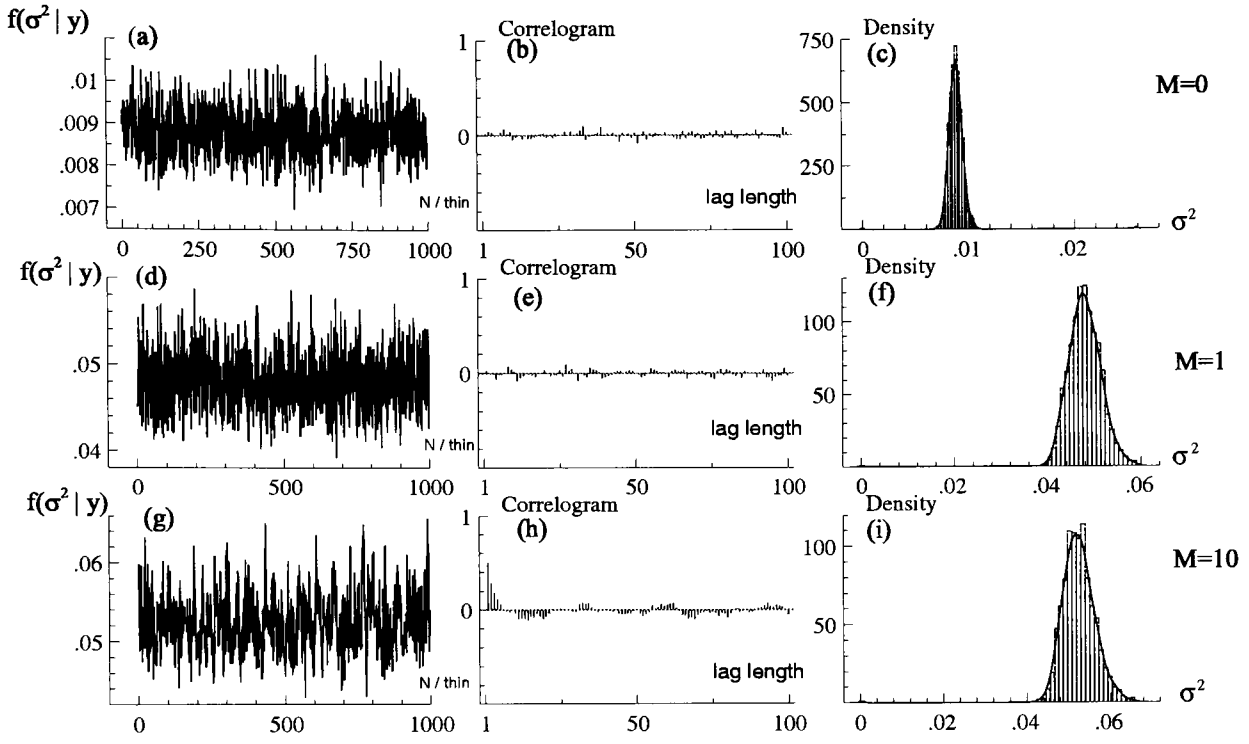


Figure 6.3: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for draws from $\sigma^2|y$, (true value 0.05). Top graph uses discretisation $M = 0$, middle graphs have $M = 1$ and bottom graphs have $M = 10$, $\lambda = 3$. $N = 10,000$ with every tenth simulation recorded and $\Delta^\dagger = 1$ in all simulations.*

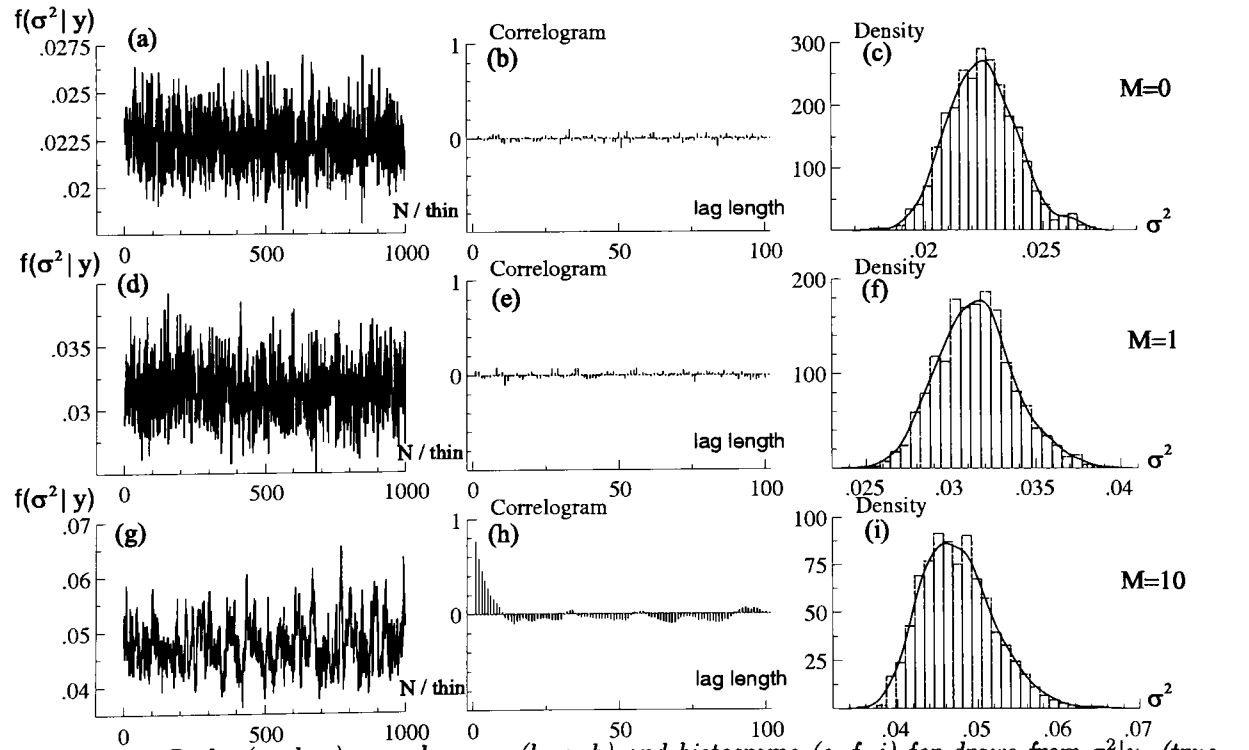


Figure 6.4: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for draws from $\sigma^2|y$, (true value 0.05). Top graph uses discretisation $M = 0$, middle graphs have $M = 1$ and bottom graphs have $M = 10$, $\lambda = 3$. $N = 10,000$ with every tenth simulation recorded and $\Delta^\dagger = 5$ in all simulations.*

variables introduced between the observed data. There is strong correlation between α and β (0.99), across M, while correlation between σ^2 and α, β initially increases with M, fluctuating around 0.75 and 0.8 for higher values of M.

Summary statistics for the parameters (CIR process)					
	Posterior Mean	Inefficiency	Covariance & *Correlation*		
$M = 0$					
$\alpha\|z$	0.31760	0.77854	0.0054081	*0.97279*	*-0.092305*
$\beta\|z$	0.12770	0.84724	0.00021815	9.2992e-005	*-0.21829*
$\sigma^2\|z$	0.022571	1.1696	-1.2008e-005	-1.1775e-005	3.1291e-005
$M = 1$					
$\alpha\|z$	0.40436	1.1150	0.0013566	*0.98856*	*0.54185*
$\beta\|z$	0.16207	1.1429	0.00054035	0.00022024	*0.53330*
$\sigma^2\|z$	0.031452	0.81728	4.8755e-005	1.9335e-005	5.9681e-006
$M = 3$					
$\alpha\|z$	0.46317	1.2305	0.0026897	*0.99218*	*0.68820*
$\beta\|z$	0.18467	1.2050	0.0010667	0.00042976	*0.67840*
$\sigma^2\|z$	0.039716	1.0100	0.00012994	5.1199e-005	1.3253e-005
$M = 5$					
$\alpha\|z$	0.48435	2.9272	0.0029436	*0.99234*	*0.75960*
$\beta\|z$	0.19274	2.8637	0.0011658	0.00046890	*0.75262*
$\sigma^2\|z$	0.043112	3.4651	0.00017067	6.7491e-005	1.7150e-005
$M = 10$					
$\alpha\|z$	0.51020	3.2109	0.0031933	*0.99134*	*0.75791*
$\beta\|z$	0.20257	3.2473	0.0012596	0.00050556	*0.75184*
$\sigma^2\|z$	0.047446	3.7755	0.00020091	7.9299e-005	2.2005e-005
$M = 20$					
$\alpha\|z$	0.51063	11.035	0.0037585	*0.99268*	*0.80889*
$\beta\|z$	0.20264	11.189	0.0014996	0.00060718	*0.80235*
$\sigma^2\|z$	0.048186	14.682	0.00026019	0.00010373	2.7529e-005
$M = 30$					
$\alpha\|z$	0.50255	14.210	0.0033777	*0.99158*	*0.78960*
$\beta\|z$	0.19932	14.063	0.0013367	0.00053798	*0.78177*
$\sigma^2\|z$	0.047606	19.590	0.000222342	8.8281e-005	2.3704e-005

Table 6.1: *The posterior means, covariances and correlations are shown for the parameters of the CIR process. True values are* $\alpha = 0.5$, $\beta = 0.2$ *and* $\sigma^2 = 0.05$. *Data was obtained using* $T = 500$ *and* $\Delta^\dagger = 5$. *The MH algorithm was run with* $M = 0, 1, 10$ *(*$\lambda = 3$*),* 20 *(*$\lambda = 5$*) and* 30 *(*$\lambda = 9$*) for* $N = 10,000$ *iterations. The bandwidth used in the inefficiency factor computations,* $B_N = 100$ *in all computations.*

6.2.4 A Monte Carlo experiment for the CIR model

Table 6.2 shows the results from a small simulation study for the parameters of the CIR process. Ten sets of data with $T = 500$ and $\Delta^\dagger = 5$ were generated and the MH algorithm was run with $M = 0, 1, 10, 20$ and 30 for $N = 10,000$ iterations. In general, $N = 10,000$ is used if the simulation inefficiency factor is around 20, on the basis that posterior noise is 500 times the simulation noise. The Monte Carlo results reported in Table 6.2, show that as M increases, there is a marked decrease in bias, though the difference between $M = 10, 20$ and $M = 30$ is negligible. This seems to suggest that though the initial increase in M gives more precise estimates of the parameters, there are diminishing returns after certain values, (for example $M > 20$). There is a corresponding increase in the standard deviations that accompanies

Monte Carlo Results (CIR process)					
	Mean	Bias	s.d.	Mean(s.e)	MSE(×100)
α (0.5)					
$M = 0$	0.310	-0.190	0.0161	0.0227	3.6190
$M = 1$	0.393	-0.107	0.0276	0.0359	1.2302
$M = 3$	0.449	-0.051	0.0375	0.0480	0.4032
$M = 5$	0.470	-0.031	0.0403	0.0515	0.2570
$M = 10$	0.496	-0.004	0.0461	0.0543	0.2136
$M = 20$	0.498	-0.0019	0.0451	0.0580	0.2035
$M = 30$	0.494	-0.0061	0.0445	0.0579	0.2021
β (0.2)					
$M = 0$	0.127	-0.074	0.0069	0.0095	5.4814
$M = 1$	0.159	-0.041	0.0115	0.0146	0.1806
$M = 3$	0.181	-0.019	0.0154	0.0194	0.0605
$M = 5$	0.189	-0.011	0.0165	0.0208	0.0400
$M = 10$	0.199	-0.000	0.0188	0.0219	0.0355
$M = 20$	0.200	-0.000	0.0185	0.0236	0.0341
$M = 30$	0.198	-0.002	0.0182	0.0234	0.0337
σ (0.05)					
$M = 0$	0.023	-0.027	0.0008	0.0056	0.0738
$M = 1$	0.032	-0.018	0.0013	0.0025	0.0325
$M = 3$	0.040	-0.010	0.0022	0.0036	0.0097
$M = 5$	0.044	-0.006	0.0026	0.0042	0.0044
$M = 10$	0.048	-0.002	0.0033	0.0048	0.0013
$M = 20$	0.049	-0.001	0.0033	0.0052	0.0011
$M = 30$	0.049	-0.001	0.0034	0.0051	0.0013

Table 6.2: *The results from a small simulation study for the parameters of the CIR process are shown. Ten sets of data with* $T = 500$, *and* $\Delta^{\dagger} = 5$ *were generated and the MH algorithm was run with* $M = 0, 1, 10, 20$ *and* 30 *for* $N = 10,000$ *iterations.*

an increase in M, as the estimation procedure is implemented over a larger number of latent states and so becomes harder. This increase, however, seems to level out for higher values of M. The MSE, (which can be used as a guide in applications in choosing the number of latent points to be imputed, given the model and discretisation interval in the observed data), is smallest for $M = 20$, when estimating σ^2.

To monitor the convergence of each parameter, we can additionally compute the *Gelman-Rubin* r statistic, see Gelman and Rubin (1992a) and Gelman, Carlin, Stern, and Rubin, (1995, pp. 331-332), which was briefly outlined in Section 3.3.6. For the parameters $(\alpha, \beta, \sigma^2)$ in the CIR model in one of the experiments, the corresponding $\sqrt{\hat{R}}$ statistic is given by $(0.99930, 0.99936, 0.99942)$, showing strong convergence. The M-H algorithm was run using $M = 3$ $(\lambda = 2)$ and $N = 5000$, with a burn-in of 1000 simulations. Computer time to run 100 iterations of the sampler for the case $M = 3$ with $T = 500$ observations is 258.60 seconds on a Pentium/133, using the matrix language Ox by Doornik (1996).

Although results from such a small simulation study are far from conclusive, they do suggest an improvement in the estimation of the parameters. This would seem to suggest that introducing auxiliary points gives a substantial improvement over results obtained that do not introduce additional points between the observed data.

6.3 UPDATING ALL M STATES IN ONE BLOCK

Instead of sampling the M latent states in smaller blocks, it is possible to update the whole block in one sweep, as was discussed in Section 5. A comparison is provided between the sampling of the M states using random block sizes, against treating all M states in one sweep, either iterating $\hat{y}_t^*$ towards the mode of the proposal using a Newton-Raphson (N-R) scheme or using the initial linear interpolation for $\hat{y}_t^*$, where the $(j+1)$'th element is given by $\hat{y}_{t,j+1}^* = y_t + \{(j+1)/(M+1)\}(y_{t+1} - y_t)$, for $j = -1, 0, \ldots, M$.

Using the notation defined in 5, the values of $\hat{y}_t^*$ can be iterated towards the mode of the proposal density, given by $q(y_t^*) \sim N(\tilde{\mu}, \tilde{\Sigma})$, which is just the proposal density from the MCMC scheme. Similarly, u is given by the $M \times 1$ vector with

$$u_{j+1} = \frac{\partial l(y_t^*)}{\partial y_{t,j}^*}.$$

The quantity $\hat{y}_t^{(k+1)*}$ is then iterated towards the mode using a N-R scheme, as

$$\hat{y}^{(k+1)*} = \hat{y}^{(k)*} + u\tilde{\Sigma}.$$

To verify the algorithm used in the N-R step, the OU process can be used due to its tractable form for the conditional transition density of y_t^*, given the neighbours. The expected path between the points can thus be evaluated and compared to the $\hat{y}_t^*$ obtained from the N-R scheme. The convergence to the mode, in the simulation studies reported below, occurred after one step of the algorithm.

Table 6.3 shows the comparison between the three MCMC schemes for $M = 2$, with $N = 4,000$ and $T = 500$ for the CIR process. Summary graphs are provided in the two plots in Figure 6.5. For small M, the difference between sampling all the auxiliary points in one sweep or using a random block size is negligible. There is nearly no difference between the posterior means and inefficiency factors are below one, indicating that the sampler is close to one drawing i.i.d. random variables from the posterior distribution. This reflects the fact that the correlation is low, when a small number of auxiliary points are introduced. Acceptance rates for the random block case is slightly higher, at 0.98 (as opposed to 0.95 for the case with no block size). For the CIR process, it would appear that for small M, the linear interpolation between the observed values for $\hat{y}_t^*$ is a good approximation to the mode of the conditional transition density, given by the $N(\tilde{\mu}, \tilde{\Sigma})$ density.

The results for the case where we increase the number of auxiliary points to $M = 5$, is shown in Table 6.4. The figures provide a comparison between the three MCMC schemes for $\Delta^\dagger = 1$, with $N = 4,000$ and $T = 500$ for the CIR process. Summary graphs are provided in the two plots in Figure 6.6. Posterior means are slightly more precise for the random block case, though inefficiency factors are similar in all three cases. The correlation between the drift parameters is around 0.97 and 0.38 between the drift and volatility coefficients. The average acceptance rate for the M-H algorithm, sampling all the latent states in one sweep, ranges between 0.84 and 0.91. A higher acceptance rate is reported for the case with block sampling. When the average random block size is $\lambda = 2$, acceptance rates are around 0.95 for the mid-state and around 0.96 at each end-point. This suggests that one area where the efficiency of the sampler can be increased, is in the placing of the auxiliary points within a given interval. The acceptance rates are higher for the states around the fixed end points, reflecting the availability of sample information. This suggests placing a large proportion of the auxiliary points towards the ends of the interval.

Summary statistics $M = 2$					
	Mean	Inefficiency	Covariance & *Correlation*		
block=M (iteration to mode)					
$\alpha\|z$ (0.5)	0.50237	0.64524	0.0055561	*0.97900*	*0.42164*
$\beta\|z$ (0.2)	0.19843	0.71596	0.0022157	0.00092189	*0.41276*
$\sigma^2\|z$(0.05)	0.049572	0.70785	0.00010593	4.2242e-05	1.13160e-05
block=M (no iteration to mode)					
$\alpha\|z$ (0.5)	0.50257	0.63709	0.0055713	*0.97904*	*0.42276*
$\beta\|z$ (0.2)	0.19851	0.70638	0.0022210	4.2596e-05	*0.41341*
$\sigma^2\|z$ (0.05)	0.049590	0.70952	0.00010698	4.2596e-05	1.1493e-05
random block size $\lambda = 1$					
$\alpha\|z$ (0.5)	0.50126	0.63515	0.0051547	*0.97749*	*0.28854*
$\beta\|z$ (0.2)	0.19811	0.70174	0.0020163	0.00082543	*0.27306*
$\sigma^2\|z$ (0.05)	0.049425	0.98348	6.4671e-05	2.4491e-05	9.7455e-06

Table 6.3: *A comparison between three MCMC algorithms for the CIR process is shown, where the first case considers updating all M latent states in one sweep, with $\hat{y}_t$ being iterated towards the mode of the density, for all t. The second shows the same scenario, except without any iteration towards the mode, while the third case uses a random block size of $\lambda = 1$. In all three cases, the number of auxiliary states is $M = 2$ and the algorithm is run for $N = 4,000$ iterations. True values of the parameters are shown in brackets. The data is generated using a discretisation bias of $\Delta^\dagger = 1$.*

Summary statistics $M = 5$, $\Delta^\dagger = 1$					
	Mean	Inefficiency	Covariance & *Correlation*		
block=M (iteration to mode)					
$\alpha\|z$ (0.5)	0.51112	0.34944	0.0052394	*0.97442*	*0.37368*
$\beta\|z$ (0.2)	0.20196	0.47755	0.0020944	0.00088178	*0.34044*
$\sigma^2\|z$(0.05)	0.051293	1.5086	9.8123e-05	3.6673e-05	1.3160e-06
block=M (no iteration to mode)					
$\alpha\|z$ (0.5)	0.51103	0.37771	0.0052813	*0.97460*	*0.38104*
$\beta\|z$ (0.2)	0.20193	0.51911	0.0021103	0.00088771	*0.34805*
$\sigma^2\|z$ (0.05)	0.051286	1.4320	0.00010134	3.7950e-05	1.3392e-05
random block size $\lambda = 2$					
$\alpha\|z$ (0.5)	0.50309	0.67313	0.0052638	*0.97866*	*0.38998*
$\beta\|z$ (0.2)	0.19837	0.76862	0.0020677	0.00084804	*0.37282*
$\sigma^2\|z$ (0.05)	0.050572	1.4595	9.8757e-05	3.7895e-05	1.1283e-05

Table 6.4: *A comparison between three MCMC algorithms for the CIR process is shown, where the first case considers updating all M latent states in one sweep, with $\hat{y}_t$ being iterated towards the mode of the density, for all t. The second shows the same scenario, except without any iteration towards the mode, while the third case uses a random block size of $\lambda = 1$. In all three cases, the number of auxiliary states is $M = 5$ and the algorithm is run for $N = 4,000$ iterations. True values of the parameters are shown in brackets. The data is generated using a discretisation bias of $\Delta^\dagger = 1$.*

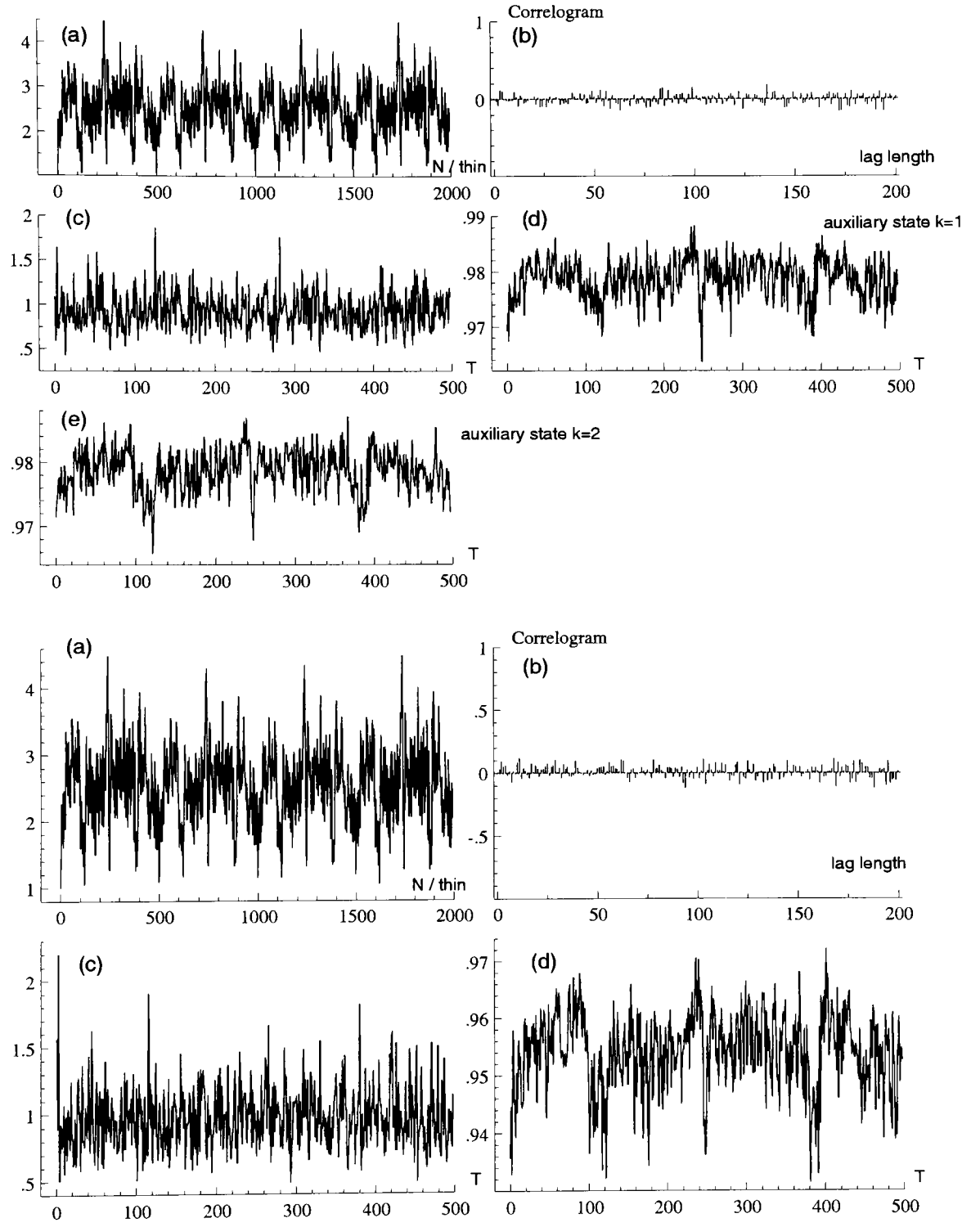


Figure 6.5: *Summary graphs for the sampling of the auxiliary states in the CIR process are shown. The algorithm was run with* $M = 2$*, for* $N = 4000$ *iterations, with every second simulation recorded. The* $T = 500$ *data points were generated with a discretisation bias of* $\Delta^{\dagger} = 1$*. Top panel shows the results using block sampling, while the bottom panel depicts the case for updating all states in one sweep. For both panels, from top left: average path of simulations for states (a), correlogram (b), inefficiency factors plotted over* T *(c) and acceptance rates plotted over* T *(d); top panel: plotted for auxiliary states* $k = 1$ *(d) and* $k = 2$ *(e). As the graphs for updating all latent states in one sweep with or without iteration to the mode were practically identical, only the former is shown.*

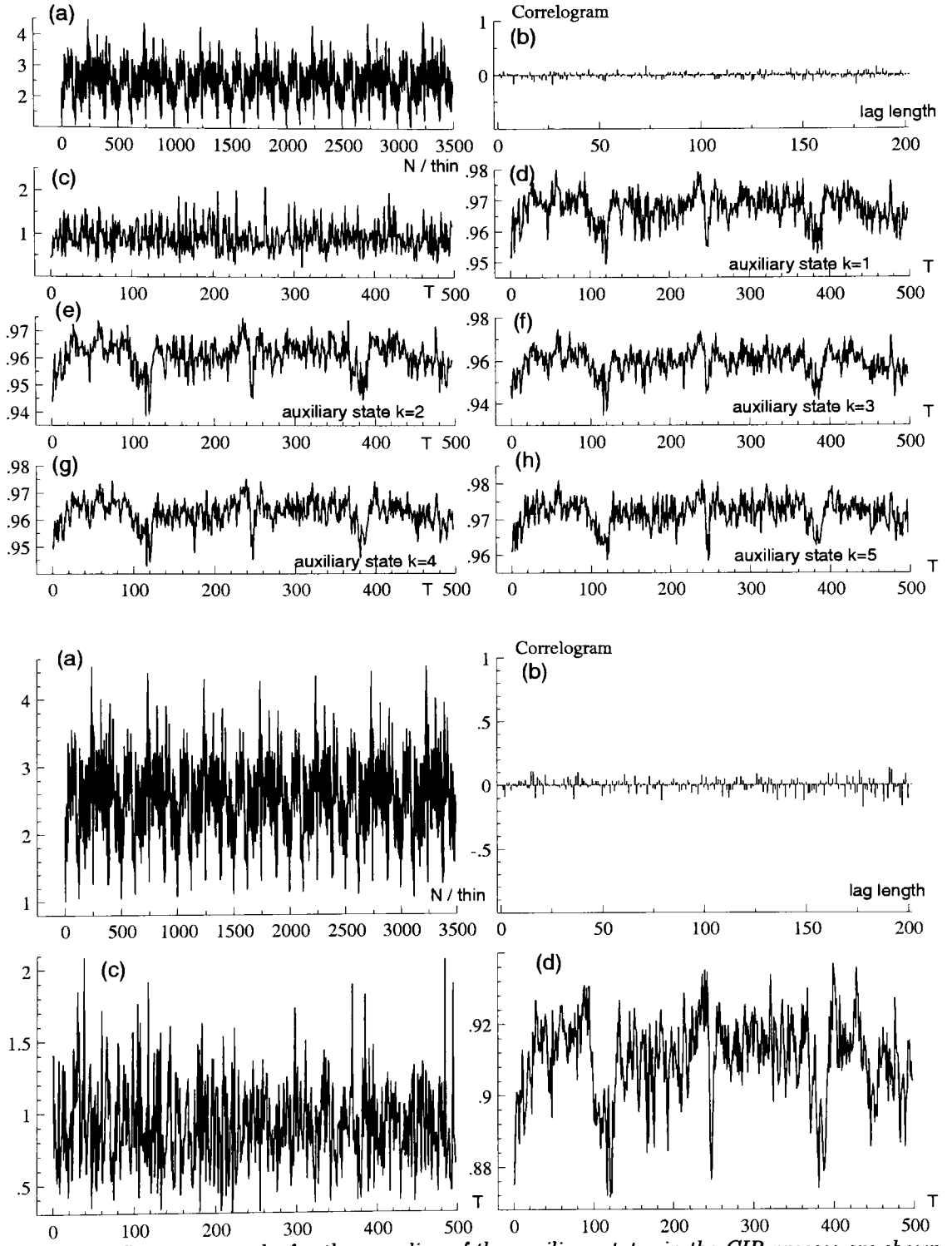


Figure 6.6: *Summary graphs for the sampling of the auxiliary states in the CIR process are shown. The algorithm was run with* $M = 2$, *for* $N = 4000$ *iterations, with a burn-in of 500 and each simulation recorded. The* $T = 500$ *data points were generated with a discretisation bias of* $\Delta^{\dagger} = 1$. *Top panel shows the results using block sampling, while the bottom panel depicts the case for updating all states in one sweep. For both panels, from top left: average path of simulations for states (a), correlogram (b), inefficiency factors plotted over* T *(c) and acceptance rates plotted over* T *(d); top panel: auxiliary state* $k = 1$ *(d),* $\ldots$,$k = 5$ *(h) are shown. As the graphs for updating all latent states in one sweep with or without iteration to the mode were practically identical, only the former is shown.*

Summary statistics $M = 2$, $\Delta^\dagger = 5$					
	Mean	Inefficiency	Covariance & *Correlation*		
block=M (iteration to mode)					
$\alpha\|z$ (0.5)	0.37220	0.85502	0.0013362	*0.98731*	*0.60577*
$\beta\|z$ (0.2)	0.14885	0.92063	0.00053602	0.00022058	*0.59324*
$\sigma^2\|z$(0.05)	0.035098	0.89084	5.9483e-05	2.3668e-05	7.2150e-06
block=M (no iteration to mode)					
$\alpha\|z$ (0.5)	0.37200	0.64025	0.0013297	*0.98724*	*0.60373*
$\beta\|z$ (0.2)	0.14876	0.70495	0.00053460	0.00022053	*0.59260*
$\sigma^2\|z$ (0.05)	0.035065	0.77445	5.9394e-05	2.3742e-05	7.2785e-05
random block size $\lambda = 1$					
$\alpha\|z$ (0.5)	0.37638	1.1608	0.0011204	*0.98492*	*0.52263*
$\beta\|z$ (0.2)	0.15062	1.3006	0.00044403	0.00018141	*0.50523*
$\sigma^2\|z$ (0.05)	0.035154	1.1025	4.3946e-05	1.7095e-05	6.3109e-06

Table 6.5: *A comparison between three MCMC algorithms for the CIR process is shown, where the first case considers updating all M latent states in one sweep, with $\hat{y}_t$ being iterated towards the mode of the density, for all t. The second shows the same scenario, except without any iteration towards the mode, while the third case uses a random block size of $\lambda = 1$. In all three cases, the number of auxiliary states is $M = 2$ and the algorithm is run for $N = 4,000$ iterations. True values of the parameters are shown in brackets. The data is generated using a discretisation bias of $\Delta^\dagger = 5$.*

We now increase the inherent bias in the simulated data by increasing $\Delta^\dagger$ to 5. Table 6.5 shows the comparison between the three MCMC schemes for $M = 2$, with $N = 4,000$ and $T = 500$ for the CIR process. The bias is more pronounced as the posterior means are around 0.37 for α, 0.15 for β and 0.035 for σ^2. Correlation between the drift and volatility coefficients has increased to 0.6.

When we increase the number of auxiliary points within each pair of observations to 10, the benefits from block sampling become more apparent. Table 6.6 shows the results from a comparison between the three MCMC schemes for this case, with $N = 4000$. The data was generated using $\Delta^\dagger = 5$, for $T = 500$ observations. Acceptance rates, for sampling all of the auxiliary states in one sweep within the M-H algorithm, range between 0.6 and 0.72. Splitting the block into smaller pieces, where the average block size is $\lambda = 3$, shows an increase in the average acceptance rate to around 0.925 for the mid-state and around 0.9 at each end-point. These can be seen in the plots in Figure 6.7. Notice the band structure in the average paths of the simulations as we increase the number of auxiliary points. In addition, for the state around 350, the inefficiency factor increases from around 1 to 6 and the acceptance rate decreases from 0.7 to 0.5. This shows that we are rejecting the proposed move for that state as often as we are accepting the move. Inefficiency factors are higher for the case with no blocks; we are up to 12 times less efficient than an i.i.d. sampler for drawing posterior values σ^2 as opposed to 3.9 times for the case with an average block size of $\lambda = 3$.

Summary statistics for the parameters of the CIR process are shown in Figure 6.7, including a measure of the inefficiency of the sampler, the posterior mean and covariance and correlation statistics. There are clear costs to increasing the number of auxiliary points. Decreases in returns to scale can be seen for both cases $M = 20$ and M =30. The posterior means are more precise for $M = 20$ and inefficiency factors are lower. For σ^2 with $M = 20$ auxiliary points, we are up to 28 times less efficient than an i.i.d. sampler, as opposed to 54 for the case $M = 30$. Correlation between the drift parameters is again around 0.98-0.99 for all M, while the correlation between the drift and volatility coefficients increases from -0.09 for $M = 0$ to 0.7 for $M = 30$. Note that in the case of block sampling, with a random block size of $\lambda = 3$ in Table 6.7, we are only up to 19 times less efficient for σ^2 using $M = 30$ imputed points.

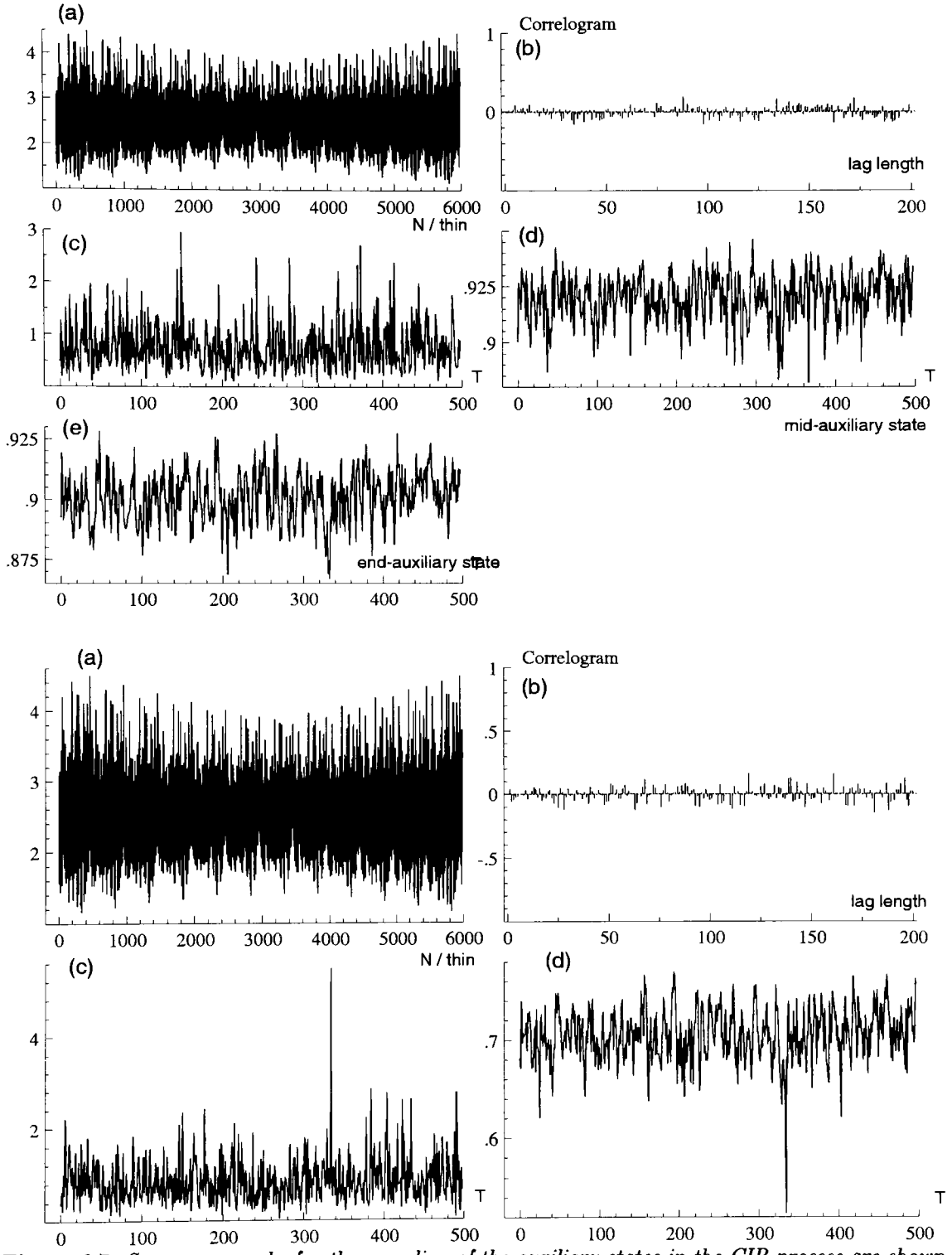


Figure 6.7: *Summary graphs for the sampling of the auxiliary states in the CIR process are shown. The algorithm was run with* $M = 10$*, for* $N = 6000$ *iterations, where each simulation was recorded. The* $T = 500$ *data points were generated with a discretisation bias of* $\Delta^{\dagger} = 5$*. Top panel shows the results using block sampling, while the bottom panel depicts the case for updating all states in one sweep. For both panels, from top left: average path of simulations for states (a), correlogram (b), inefficiency factors plotted over* T *(c) and acceptance rates, plotted against* T *(d); top panel: mid-auxiliary state (d) and end-state (e). As the graphs for updating all latent states in one sweep with or without iteration to the mode were practically identical, only the former is shown.*

Summary statistics $M = 10$, $\Delta^{\dagger} = 5$					
	Mean	Inefficiency	Covariance & *Correlation*		
block=M (iteration to mode)					
$\alpha\|z$ (0.5)	0.41264	7.9877	0.0020614	*0.98959*	*0.69655*
$\beta\|z$ (0.2)	0.16410	7.9069	0.00082476	0.00033697	*0.68423*
$\sigma^2\|z$(0.05)	0.042457	12.616	0.00011214	4.4538e-05	1.2574e-05
block=M (no iteration to mode)					
$\alpha\|z$ (0.5)	0.41554	8.1170	0.0023289	*0.99080*	*0.73195*
$\beta\|z$ (0.2)	0.16527	8.3673	0.00093177	0.00037975	*0.72145*
$\sigma^2\|z$ (0.05)	0.042743	11.647	0.00013673	5.4422e-05	1.4984e-05
random block size $\lambda = 3$					
$\alpha\|z$ (0.5)	0.42738	4.3704	0.0018617	*0.98804*	*0.66008*
$\beta\|z$ (0.2)	0.16981	4.4576	0.00074508	0.00030546	*0.65648*
$\sigma^2\|z$ (0.05)	0.044060	3.8939	0.00010043	4.0460e-05	1.2435e-05

Table 6.6: *A comparison between three MCMC algorithms for the CIR process is shown, where the first case considers updating all M latent states in one sweep, with $\hat{y}_t$ being iterated towards the mode of the density, for all t. The second shows the same scenario, except without any iteration towards the mode, while the third case uses a random block size of $\lambda = 3$. In all three cases, the number of auxiliary states is $M = 10$ and the algorithm is run for $N = 4,000$ iterations. True values of the parameters are shown in brackets. the data is generated using a discretisation bias of $\Delta^{\dagger} = 5$.*

This again points to the fact that for large M, a randomisation device should be used to increase the efficiency of the sampler.

6.4 Additional models

This section considers estimation of the Affine CIR and a general parametric model considered in Aït-Sahalia (1996b). These are applied to the 7-day Eurodollar deposit spot-rate in section 7.

6.4.1 Affine CIR

The model can be expressed as

$$dy(t) = \{\alpha_0 + \alpha_1 y(t)\}dt + \{\beta_0 + \beta_1 y(t)\}^{\frac{1}{2}} dW(t).$$

We work with the log transformation, (a special case of the above),

$$\begin{aligned} dx(t) &= \left[\frac{\alpha_0}{\exp\{x(t)\}} + \alpha_1 - \frac{\beta_0}{2\exp\{2x(t)\}} - \frac{\beta_1}{2\exp\{x(t)\}}\right] dt \\ &\quad + \left[\frac{\beta_0}{\exp\{2x(t)\}} + \frac{\beta_1}{\exp\{x(t)\}}\right]^{\frac{1}{2}} dW(t), \end{aligned} \tag{6.14}$$

Summary statistics for the parameters (CIR process)					
	Posterior Mean	Inefficiency	Covariance& *Correlation*		
$M = 0$					
$\alpha\|z$	0.31760	0.77854	0.0054081	*0.97279*	*-0.092305*
$\beta\|z$	0.12770	0.84724	0.00021815	9.2992e-005	*-0.21829*
$\sigma^2\|z$	0.022571	1.1696	-1.2008e-005	-1.1775e-005	3.1291e-005
$M = 1$					
$\alpha\|z$	0.34983	3.3544	0.00095933	*0.98281*	*0.43782*
$\beta\|z$	0.14031	3.1476	0.00038242	0.00015782	*0.42118*
$\sigma^2\|z$	0.030813	4.7331	2.9205e-005	1.1395e-005	4.6383e-006
$M = 3$					
$\alpha\|z$	0.48255	2.9676	0.0024796	*0.98956*	*0.68356*
$\beta\|z$	0.19366	3.0439	0.00099131	0.00040472	*0.67254*
$\sigma^2\|z$	0.042190	4.8709	0.00012444	4.9462e-005	1.3365e-005
$M = 5$					
$\alpha\|z$	0.40983	17.425	0.0017418	*0.98794*	*0.63409*
$\beta\|z$	0.16317	17.087	0.00069554	0.00028457	*0.62261*
$\sigma^2\|z$	0.041096	32.788	8.9695e-005	3.5599e-005	1.1488e-005
$M = 10$					
$\alpha\|z$	0.41337	28.558	0.0018831	*0.98855*	*0.69104*
$\beta\|z$	0.16440	28.256	0.00074956	0.00030531	*0.67973*
$\sigma^2\|z$	0.042388	51.746	0.00010681	4.2303e-005	1.2686e-005
$M = 20$					
$\alpha\|z$	0.48340	12.827	0.0019663	*0.98812*	*0.66389*
$\beta\|z$	0.19302	12.357	0.00078681	0.00032245	*0.64892*
$\sigma^2\|z$	0.045660	27.946	9.0782e-05	3.5933e-05	9.5094e-05
$M = 30$					
$\alpha\|z$	0.46221	26.411	0.0019990	*0.98752*	*0.70251*
$\beta\|z$	0.18467	26.215	0.00079281	0.00032243	*0.69024*
$\sigma^2\|z$	0.043128	54.256	0.00011692	4.6136e-05	1.3856e-05

Table 6.7: *The posterior means, covariances and correlations are shown for the parameters of the CIR process. True values are* $\alpha = 0.5$, $\beta = 0.2$ *and* $\sigma^2 = 0.05$. *Data was obtained using the strong solution of the process, with* $T = 500$ *and* $\Delta^\dagger = 5$. *The MH algorithm was run with* $M = 0, 1, 10,$ 20 *and* 30 *for* $N = 4000$ *iterations. The bandwidth used in the inefficiency factor computations,* $B_N = 100$ *in all computations.*

where the transformation $x(t) = \log y(t)$ has been applied. On using the Euler discretisation scheme on the transformed variable, we obtain

$$x^*_{t,k+1}|x^*_{t,k} \sim N\left[\left\{\frac{\alpha_0}{\exp(x^*_{t,k})} + \alpha_1 - \frac{\beta_0}{2\exp(2x^*_{t,k})} - \frac{\beta_1}{2\exp(x^*_{t,k})}\right\}\Delta + x^*_{t,k}, \left(\frac{\beta_0}{\exp(2x^*_{t,k})} + \frac{\beta_1}{\exp(x^*_{t,k})}\right)\Delta\right].$$

The parameters of the proposal distribution are given by equations B.5.7–B.5.9 in appendix 5.6.1, where

$$\begin{aligned}
a_{t,j} &= \frac{\alpha_0}{\exp(x^*_{t,k+j})} + \alpha_1 - \frac{\beta_0}{2\exp(2x^*_{t,k+j})} - \frac{\beta_1}{2\exp(x^*_{t,k+j})} \\
a'_{t,j} &= -\frac{\alpha_0}{\exp(x^*_{t,k+j})} + \frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{2\exp(x^*_{t,k+j})} \\
a''_{t,j} &= \frac{\alpha_0}{\exp(x^*_{t,k+j})} - 2\frac{\beta_0}{\exp(2x^*_{t,k+j})} - \frac{\beta_1}{2\exp(x^*_{t,k+j})} \\
g_{t,j} &= \left\{\frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})}\right\}^{-1} \\
g'_{t,j} &= \left\{\frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})}\right\}^{-2}\left\{\frac{2\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})}\right\} \\
g''_{t,j} &= 2\left\{\frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})}\right\}^{-3}\left\{\frac{2\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})}\right\}^{2} \\
&\quad - \left\{\frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})}\right\}^{-2}\left\{\frac{4\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})}\right\}
\end{aligned}$$

Suppose we define $\alpha = (\alpha_0, \alpha_1)'$ and $\beta = (\beta_0, \beta_1)$, then the log-likelihood for the observed and simulated values is given by

$$\begin{aligned}
\log f(\log z|\alpha, \beta) &= -\frac{(n-2)}{2}\log 2\pi\Delta - \frac{1}{2}\sum_{t=2}^{n}\log\left(\frac{\beta_0}{z^2_{t-1}} + \frac{\beta_1}{z_{t-1}}\right) \\
&\quad -\frac{1}{2\Delta}\sum_{t=2}^{n}\left[\left(\frac{\beta_0}{z^2_{t-1}} + \frac{\beta_1}{z_{t-1}}\right)^{-1}\{\log z_t - \log z_{t-1} \right. \\
&\quad \left. - \left(\frac{\alpha_0}{z_{t-1}} + \alpha_1 - \frac{\beta_0}{2z^2_{t-1}} - \frac{\beta_1}{2z_{t-1}}\right)\Delta\right\}^2\Bigg].
\end{aligned}$$

Suppose we set the prior for α as

$$f(\alpha) = \exp\left\{-\frac{1}{2}(\alpha - \alpha_{01})'\Sigma_0^{-1}(\alpha - \alpha_{01})\right\},$$

and further denote

$$\begin{aligned}
\sigma_t &= \left(\frac{\beta_0}{z_t^2} + \frac{\beta_1}{z_t)}\right) \\
c_t &= \log z_{t+1} - \log z_t + \frac{\beta_0 \Delta}{2z_t^2} + \frac{\beta_1 \Delta}{2z_t} \\
V &= \Delta \sum_{t=1}^{n-1} \begin{pmatrix} \sigma_t^{-1} z_t^{-2} & \sigma_t^{-1} z_t^{-1} \\ \sigma_t^{-1} z_t^{-1} & \sigma_t^{-1} \end{pmatrix} \\
\tilde{\alpha} &= \sum_{t=1}^{n-1} \begin{pmatrix} \frac{c_t}{\sigma_t z_t} \\ \frac{c_t}{\sigma_t} \end{pmatrix},
\end{aligned}$$

then

$$f(\alpha|z;\beta) \sim N(\Sigma\hat{\alpha}, \Sigma^{-1}),$$

where

$$\begin{aligned}
\hat{\alpha} &= \tilde{\alpha} + \Sigma_0^{-1}\alpha_{01} \\
\Sigma^{-1} &= V + \Sigma_0^{-1}.
\end{aligned}$$

For the remaining parameters β, the ARMS procedure described at the end of Section 6 can be used. Typically non-informative priors are set for β; $f(\beta_0)$ is an inverse Gamma and $f(\beta_1)$ follows a normal distribution in the applications considered below. Let the previous value of β be denoted by $\beta_i^{(j)}$, for $i = 1, 2$ and $j = 1, \ldots, N$. Initialisation values for the ARMS algorithm, (denoted by $\hat{\beta}_{\text{left}}$ for the left value and $\hat{\beta}_{\text{right}}$ for the right value in the interval), at each round of the iterations typically takes the following form:

- if $\beta_i^{(j)} > 0$, then $\hat{\beta}_{\text{left}} = 0.5\beta_i^{(j)}$ and $\hat{\beta}_{\text{right}} = 2\beta_i^{(j)}$
- if $\beta_i^{(j)} < 0$, then $\hat{\beta}_{\text{left}} = -2|\beta_i^{(j)}|$ and $\hat{\beta}_{\text{right}} = 2|\beta_i^{(j)}|$ or $-2|\beta_i^{(j)}|$ and $-0.5|\beta_i^{(j)}|$
- if $\beta_i^{(j)} = 0$, then $\hat{\beta}_{\text{left}} = \epsilon_1$ and $\hat{\beta}_{\text{right}} = \epsilon_2$,

where typically $\epsilon_1 = -1e-011$ and $\epsilon_2 = 1e-011$.

6.4.2 General parametric model

The general parametric model is denoted by

$$dy(t) = \{\alpha_0 + \alpha_1 y(t) + \alpha_2 y(t)^2 + \frac{\alpha_3}{y(t)}\}dt + \{\beta_0 + \beta_1 y(t) + \beta_2 y(t)^{\beta_3}\}^{\frac{1}{2}} dW(t).$$

We work with the log transformation $x(t) = \log y(t)$ to obtain,

$$\begin{aligned}
dx(t) &= \left[\frac{\alpha_0}{\exp\{x(t)\}} + \alpha_1 + \alpha_2 \exp\{x(t)\} + \frac{\alpha_3}{\exp\{2x(t)\}} - \frac{\beta_0}{2\exp\{2x(t)\}} - \frac{\beta_1}{2\exp\{x(t)\}}\right. \\
&\quad \left. - \frac{\beta_2 \exp\{(\beta_3 - 2)x(t)\}}{2}\right] dt + \left[\frac{\beta_0}{\exp\{2x(t)\}} + \frac{\beta_1}{\exp\{x(t)\}} + \frac{\beta_2}{\exp\{(\beta_3 - 2)x(t)\}}\right]^{\frac{1}{2}} dW(t).
\end{aligned} \tag{6.15}$$

On using the Euler discretisation scheme on the transformed variable,

$$x^*_{t,k+1}|x^*_{t,k} \sim N(\mu_x + x^*_{t,k}, \Sigma_x),$$

where

$$\begin{aligned}
\mu_x &= \left\{ \frac{\alpha_0}{\exp(x^*_{t,k})} + \alpha_1 + \alpha_2 \exp(x^*_{t,k}) + \frac{\alpha_3}{\exp(2x^*_{t,k})} - \frac{\beta_0}{2\exp(2x^*_{t,k})} \frac{\beta_1}{2\exp(x^*_{t,k})} \right. \\
&\quad \left. - \frac{\beta_2 \exp\{(\beta_3 - 2)x^*_{t,k}\}}{2} \right\} \Delta \\
\Sigma_x &= \left(\frac{\beta_0}{\exp(2x^*_{t,k})} + \frac{\beta_1}{\exp(x^*_{t,k})} + \beta_2 \exp\{(\beta_3 - 2)x^*_{t,k}\} \right) \Delta.
\end{aligned}$$

The parameters of the proposal distribution are again given by equations B.5.7–B.5.9 in Appendix 5.6.1, where

$$\begin{aligned}
a_{t,j} &= \frac{\alpha_0}{\exp(x^*_{t,k+j})} + \alpha_1 + \alpha_2 \exp(x^*_{t,k+j}) + \frac{\alpha_3}{\exp(2x^*_{t,k+j})} - \frac{\beta_0}{2\exp(2x^*_{t,k+j})} - \frac{\beta_1}{2\exp(x^*_{t,k+j})} \\
&\quad - \frac{\beta_2 \exp\{(\beta_3 - 2)x^*_{t,k+j}\}}{2} \\
a'_{t,j} &= -\frac{\alpha_0}{\exp(x^*_{t,k+j})} + \alpha_2 \exp(x^*_{t,k+j}) - 2\frac{\alpha_3}{\exp(2x^*_{t,k+j})} + \frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{2\exp(x^*_{t,k+j})} \\
&\quad - \frac{\beta_2(\beta_3 - 2)\exp\{(\beta_3 - 2)x^*_{t,k+j}\}}{2} \\
a''_{t,j} &= \frac{\alpha_0}{\exp(x^*_{t,k+j})} + \alpha_2 \exp(x^*_{t,k+j}) + 4\frac{\alpha_3}{\exp(2x^*_{t,k+j})} - 2\frac{\beta_0}{\exp(2x^*_{t,k+j})} - \frac{\beta_1}{2\exp(x^*_{t,k+j})} \\
&\quad - \frac{\beta_2(\beta_3 - 2)^2 \exp\{(\beta_3 - 2)x^*_{t,k+j}\}}{2} \\
g_{t,j} &= \left\{ \frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})} + \beta_2 \exp\{(\beta_3 - 2)x^*_{t,k+j}\} \right\}^{-1} \\
g'_{t,j} &= -\left\{ \frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})} + \beta_2 \exp\{(\beta_3 - 2)x^*_{t,k+j}\} \right\}^{-2} \\
&\quad \times \left[-\frac{2\beta_0}{\exp(2x^*_{t,k+j})} - \frac{\beta_1}{\exp(x^*_{t,k+j})} + \beta_2 \exp\{(\beta_3 - 2)x^*_{t,k+j}\} \right] \\
g''_{t,j} &= 2\left\{ \frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})} + \beta_2 \exp\{(\beta_3 - 2)x^*_{t,k+j}\} \right\}^{-3} \\
&\quad \times \left[-\frac{2\beta_0}{\exp(2x^*_{t,k+j})} - \frac{\beta_1}{\exp(x^*_{t,k+j})} + \beta_2 \exp\{(\beta_3 - 2)x^*_{t,k+j}\} \right]^2 \\
&\quad - \left\{ \frac{\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})} + \beta_2 \exp\{(\beta_3 - 2)x^*_{t,k+j}\} \right\}^{-2} \\
&\quad \times \left[\frac{4\beta_0}{\exp(2x^*_{t,k+j})} + \frac{\beta_1}{\exp(x^*_{t,k+j})} + \beta_2(\beta_3 - 2)^2 \exp\{(\beta_3 - 2)x^*_{t,k+j}\} \right].
\end{aligned}$$

Suppose we define $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)'$ and $\beta = (\beta_0, \beta_1, \beta_1, \beta_3)'$, then the log-likelihood for the observed and simulated values is given by

$$\begin{aligned}\log f(\log z|\alpha,\beta) &= -\frac{(n-2)}{2}\log 2\pi\Delta - \frac{1}{2}\sum_{t=2}^{n}\log\left(\frac{\beta_0}{z_{t-1}^2}+\frac{\beta_1}{z_{t-1}}+\beta_2 z_{t-1}^{\beta_3-2}\right)\\ &\quad -\frac{1}{2\Delta}\sum_{t=2}^{n}\left[\left(\frac{\beta_0}{z_{t-1}^2}+\frac{\beta_1}{z_{t-1}}+\beta_2 z_{t-1}^{\beta_3-2}\right)^{-1}\left\{\log z_t - \log z_{t-1}\right.\right.\\ &\quad \left.\left.-\left(\frac{\alpha_0}{z_{t-1}}+\alpha_1+\alpha_2 z_{t-1}+\frac{\alpha_3}{z_{t-1}^2}-\frac{\beta_0}{2z_{t-1}^2}-\frac{\beta_1}{2z_{t-1}}-\frac{\beta_2 z_{t-1}^{\beta_3-2}}{2}\right)\Delta\right\}^2\right].\end{aligned}$$

Suppose we set the prior for α as

$$f(\alpha) = \exp\left\{-\frac{1}{2}(\alpha-\alpha_{01})'\Sigma_0^{-1}(\alpha-\alpha_{01})\right\},$$

and further denote

$$\begin{aligned}\sigma_t &= \left(\frac{\beta_0}{z_t^2}+\frac{\beta_1}{z_t)}+\beta_2 z_t^{\beta_3-2}\right)\\ c_t &= \log z_{t+1} - \log z_t + \frac{\beta_0\Delta}{2z_t^2}+\frac{\beta_1\Delta}{2z_t}+\frac{\beta_2 z_t^{\beta_3-2}\Delta}{2}\\ V &= \Delta\sum_{t=1}^{n-1}\frac{\Delta}{\sigma_t}\begin{pmatrix} z_t^{-2} & z_t^{-1} & 1 & z_t^{-3}\\ z_t^{-1} & 1 & z_t & z_t^{-2}\\ 1 & z_t & z_t^2 & z_t^{-1}\\ z_t^{-3} & z_t^{-2} & z_t^{-1} & z_t^{-1}\end{pmatrix}\\ \tilde{\alpha} &= \sum_{t=1}^{n-1}\frac{c_t}{\sigma_t}\begin{pmatrix} z_t^{-1}\\ 1\\ z_t\\ z_t^{-2}\end{pmatrix},\end{aligned}$$

then

$$f(\alpha|z;\beta) \sim N(\Sigma\hat{\alpha}, \Sigma^{-1}),$$

where

$$\begin{aligned}\hat{\alpha} &= \tilde{\alpha} + \Sigma_0^{-1}\alpha_{01}\\ \Sigma^{-1} &= V + \Sigma_0^{-1}.\end{aligned}$$

For the remaining parameters β, the ARMS procedure is used. Typically non-informative priors, for the cases considered in Section 7 are set for β; $f(\beta_0)$ and $f(\beta_2)$ follow inverse Gamma distributions while $f(\beta_1)$ and $f(\beta_3)$ are set to normal distributions. Initialisation values are chosen as above. Note that for both the general parametric and affine CIR models, the diffusion coefficient will be zero before the states become negative.

6.5 HYPERBOLIC DIFFUSIONS

The distribution of stock returns is typically heavier tailed than the normal distribution, which has been documented in the literature by authors such as Mandelbrot (1963) and Fama (1965). The most famous model proposed in the literature to model this phenomenon is the Black-Scholes model (Black and Scholes 1973) which assumes normally distributed log returns. The logarithm of stock prices has been shown in empirical studies to have increments with very low correlation but which are not independent, see Fama (1965) and Taylor, (Ch.2, 1986). An additional study of stylised facts can be found in Rydberg (1997). We will consider the hyperbolic distribution, (first introduced by Barndorff-Nielsen 1977 as a model for sand) as it has been known to provide a good fit to many financial series, see Eberlein and Keller (1995) and Küchler, Neumann, Sørensen, and Streller (1994). Bibby and Sørensen (1997) have noted that the marginal distribution for financial returns is well approximated by the hyperbolic distribution.

Due to the findings that the stock return process itself has low correlation, but not independent increments, Taylor, (Ch.2, 1986) suggests the following form for the model

$$S_t = \exp(\kappa t + X_t), \tag{6.16}$$

where

$$X_t = X_0 + \int_0^t v(X_s)\, dW_s, \tag{6.17}$$

and S_t is the stock price. By Itô's formula

$$dS_t = S_t \left[\left\{ \kappa + \frac{1}{2} v^2(\log S_t - \kappa t) \right\} dt + v(\log S_t - \kappa t) dW_t \right]. \tag{6.18}$$

A more extensive coverage of the properties can be found in Bibby and Sørensen (1997). Briefly, one can obtain different models on appropriate choice of $v(x)$. We will choose the diffusion

$$v^2(x) = \sigma^2 \exp\left\{ \alpha \sqrt{\delta^2 + (x-\mu)^2} - \beta(x-\mu) \right\}, \tag{6.19}$$

as suggested by Bibby and Sørensen (1997), which corresponds to the logarithm of the stock price (eventually) approximately following a hyperbolic distribution. This representation is chosen as it ensures ergodicity with hyperbolic invariant measure. The density of the hyperbolic is given by

$$h(x; \alpha, \beta, \delta, \mu) = \frac{\sqrt{\alpha^2 - \beta^2}}{2\alpha\delta K_1(\delta\sqrt{\alpha^2 - \beta^2})} \exp\left\{ -\alpha\sqrt{\delta^2 + (x-\mu)^2} + \beta(x-\mu) \right\}, \quad x \in \mathbb{R}.$$

The parameters $(\alpha, \beta, \delta, \mu)$ are such that $\mu \in \mathbb{R}$, $\delta > 0$ and $0 \leq |\beta| < \alpha$. K_τ denotes the modified Bessel function of the third kind with index τ. For definitions and properties of Bessel functions, see Jørgensen (1982). The generalised hyperbolic distributions are mixtures of a normal and a generalised inverse Gaussian distribution with tail behaviour in the one-dimensional case taking the form $a|x|^b \exp(-c|x|)$, where a, b and c are positive constants. This accounts for the heavier tail characteristic as compared to a normal which has the form proportional to $\exp(-x^2)$. The hyperbolic distribution is one example of the generalised hyperbolic family. It has been proposed and applied as a model for sand in Barndorff-Nielsen,

Blæsild, Jensen, and Sørensen (1985). The parameter α is a gradient parameter, (the distribution is steeper the larger the α), β is the parameter of asymmetry, ($\beta = 0$ represents symmetry around the mean), δ is a scale parameter and μ the location parameter. The disadvantage of the hyperbolic distribution is that it is not closed under convolution. Rydberg (1999) discusses the Normal Inverse Gaussian, another example of a generalised hyperbolic diffusion.

One parameterization that can be chosen which is invariant to location and scale transformations involves the shape triangle. By defining

$$\xi = \left(1 + \delta\sqrt{\alpha^2 - \beta^2}\right)^{-1/2} \quad \text{and} \quad \chi = \frac{\beta}{\alpha}\xi$$

the density can be re-expressed in terms of (ξ, χ, δ, μ) with the restriction that $0 \leq |\chi| < \xi < 1$.

Given a dataset $\{S_t\}_{t=1}^T$, we will consider the log-price of the stock adjusted for trend as is outlined in Bibby and Sørensen (1997), so that $X_t = \log S_t - \kappa t$. The model we are therefore considering is given by

$$dX_t = \sigma \exp\left[\frac{1}{2}\left\{\alpha\sqrt{\delta^2 + (X_t - \mu)^2} - \beta(X_t - \mu)\right\}\right] dW_t. \tag{6.20}$$

The expressions required for the Milstein and Euler schemes are given by

$$\begin{aligned}
a(X_t;\theta) &= 0 \\
b(X_t;\theta) &= \sigma \exp\left[\frac{1}{2}\left\{\alpha\sqrt{\delta^2 + (X_t - \mu)^2} - \beta(X_t - \mu)\right\}\right] \\
b'(X_t;\theta) &= \frac{\sigma}{2}\left[\frac{\alpha(X_t - \mu)}{\{\delta^2 + (X_t - \mu)^2\}^{\frac{1}{2}}} - \beta\right] \exp\left[\frac{1}{2}\left\{\alpha\sqrt{\delta^2 + (X_t - \mu)^2} - \beta(X_t - \mu)\right\}\right] \\
g(X_t;\theta) &= \frac{1}{\sigma^2} \exp\left\{\alpha\sqrt{\delta^2 + (X_t - \mu)^2} - \beta(X_t - \mu)\right\} \\
g'(X_t;\theta) &= g(X_t;\theta)\left\{\frac{\alpha(X_t - \mu)}{\sqrt{\delta^2 + (X_t - \mu)^2}} - \beta\right\} \\
g''(X_t;\theta) &= g(X_t;\theta)\left\{\frac{\alpha(X_t - \mu)}{\sqrt{\delta^2 + (X_t - \mu)^2}} - \beta\right\}^2 \\
&\quad + g(X_t;\theta)\left[\frac{\alpha(X_t - \mu)^2\{\delta^2 + (X_t - \mu)^2\}^{-1/2} - \alpha\sqrt{\delta^2 + (X_t - \mu)^2}}{\delta^2 + (X_t - \mu)^2}\right].
\end{aligned}$$

For numerical reasons, we will work with the re-parameterised version given by the transformation $\phi = \alpha + \beta$ and $\gamma = \alpha - \beta$, so that (6.20) becomes

$$dX_t = \sigma \exp\left[\frac{1}{2}\left\{\left(\frac{\phi + \gamma}{2}\right)\sqrt{\delta^2 + (X_t - \mu)^2} - \left(\frac{\phi - \gamma}{2}\right)(X_t - \mu)\right\}\right] dW_t.$$

The relation between α and β is such that

$$\alpha > |\beta| \geq 0$$

which will be incorporated into the prior structure for (α, β). Suppose we postulate an inverse Gamma

distribution for the ratio

$$r = \frac{\alpha}{|\beta|} - 1 > 0, \quad \beta \neq 0$$

so that $r \sim$ Inverse Gamma (a, b). Then

$$f(r) = \frac{b^a}{\Gamma(a)} r^{-(a+1)} \exp\left(-\frac{b}{r}\right).$$

Suppose we express β as the product

$$\beta = s \times \gamma_2$$

where $\gamma_2 = |\beta|$ and

$$s = \begin{cases} 1 & \text{with probability } p \\ -1 & \text{with probability } (1-p) \end{cases}$$

so that

$$f(s) = p^{\frac{1+s}{2}} (1-p)^{\frac{1-s}{2}}.$$

The distribution of (γ_2, r, s) can then be expressed as

$$\begin{aligned} f(\gamma_2, r, s) &= f(r|\gamma_2, s)\, f(\gamma_2|s)\, f(s) \\ &= f(r)\, f(\gamma_2)\, f(s). \end{aligned}$$

Transforming

$$\begin{pmatrix} \gamma_2 \\ r \end{pmatrix} \rightarrow \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

we observe that the Jacobian is equal to unity. If we postulate $\beta \sim N(\beta_0, \Sigma_0)$, then γ_2 will represent a folded normal which can be expressed as a a function of a non-central chi-squared distribution with one degree of freedom as

$$f(\gamma_2) = \frac{\Sigma_0}{2} \exp\left\{-\frac{(\lambda + \gamma_2)}{2}\right\} \left(\frac{\gamma_2}{\lambda}\right)^{-(1/4)} \sqrt{\frac{2}{\pi\sqrt{\lambda\gamma_2}}} \cosh\left(\sqrt{\lambda\gamma_2}\right)$$

where $\lambda = (\beta_0/\Sigma_0)^4$, the square of the non-centrality parameter. The joint distribution of (α, β) is then

given by

$$f(\alpha,\beta) = \frac{b^a}{\Gamma(a)}\left(\frac{\alpha s}{\beta}-1\right)^{-(a+1)} \exp\left\{-\frac{b}{\left(\frac{\alpha s}{\beta}-1\right)}\right\} \frac{\Sigma_0}{2} \exp\left\{-\frac{\lambda+(\beta/s)}{2}\right\} \left(\frac{\beta}{s\lambda}\right)^{-(1/4)}$$
$$\times \sqrt{\frac{2}{\pi\sqrt{\lambda\beta/s}}} \cosh\left\{\sqrt{\lambda\beta/s}\right\}.$$

In theory,

$$f(\alpha,\beta|s) = p\, f(\alpha,\beta|s=1) + (1-p)\, f(\alpha,\beta|s=-1)$$

though as the value of β will be known when computing the prior density at that point, we will only have to work with the previous expression.

6.5.1 MCMC methods for the Hyperbolic using simulated data

Data is simulated using the Milstein scheme with precision specification $R = 10,000$. $T = 500$ observations were generated using $\Delta^\dagger = 1$, $\alpha = 4.4875$, $\beta = -3.8412$, $\delta = 1.1949$, $\mu = 7.2915$ and $\sigma = 0.0047$, see Bibby and Sørensen (1997). When the MCMC method is applied, one finds that the posterior distribution is multi-modal depending on the levels chosen for δ and μ. An illustration of this point is shown in Figure 6.8. The graph shows a plot of the logarithm of the volatility coefficient given by

$$v(x) = \sigma \exp\left\{\frac{\alpha}{2}\sqrt{\delta^2+(x-\mu)^2} - \frac{\beta}{2}(x-\mu)\right\}.$$

The graph shows that for certain ranges of the data, there exist observationally equivalent structures of θ. The sixteen values of θ given in this example are shown in Table 6.8. This shows that for a given set of (α,β,σ), various pairs of (δ,μ) can be found, giving similar values for the log-likelihood; around 1256.8 for the fifteen values mentioned above. It would seem to suggest that for certain ranges of x, a pair of values for (δ,μ) should be considered before estimating the remaining parameters as different pairs of (δ,μ) represent alternative modes of the distribution. It should also be noted that in Bibby and Sørensen (1997), the range of the Danish stock price data considered spans $(5.4, 6.6)$ suggesting that similar trajectories and likelihood statistics can be obtained for various values of (δ,μ); no standard errors are reported in their paper.

These results would seem to suggest that either one chooses a parameterization for (δ,μ) over the relevant range of x, where the shape of the exponential term can be differentiated from similar shapes resulting through different values of θ or if we are confined to a certain range of the x values, one uses a more parsimonious representation. A Taylor expansion of the exponential term can be derived, using

$$\begin{aligned} f(x) &= \log\sigma + \frac{\alpha}{2}\sqrt{\delta^2+(x-\mu)^2} - \frac{\beta}{2}(x-\mu) \\ f'(x) &= \frac{\alpha(x-\mu)}{2\sqrt{\delta^2+(x-\mu)^2}} - \frac{\beta}{2} \\ f''(x) &= \frac{\alpha}{2\sqrt{\delta^2+(x-\mu)^2}} - \frac{\alpha(x-\mu)^2}{2\{\delta^2+(x-\mu)^2\}^{3/2}} \end{aligned}$$

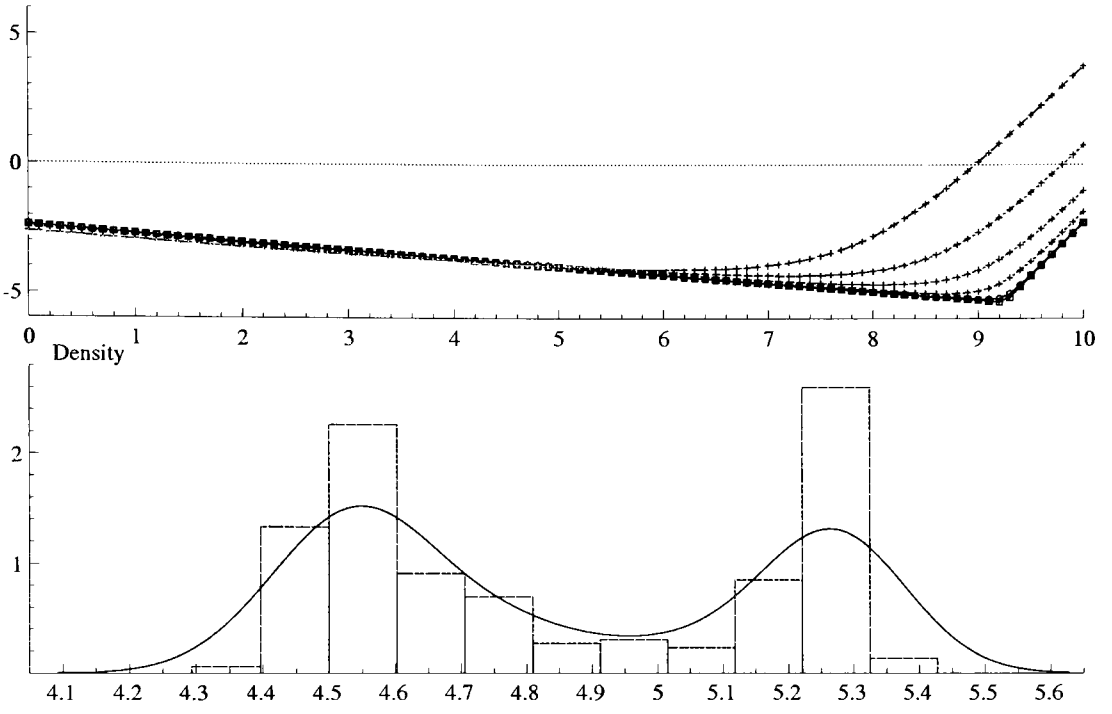


Figure 6.8: *Top panel shows a plot of the logarithm of the volatility function for eight different values of θ over a range of x values, where $x \in [0, 10]$. Bottom panel shows the distribution of the simulated data where $x \in [4.1, 5.6]$. For certain ranges of the data, there are observationally equivalent structures of θ.*

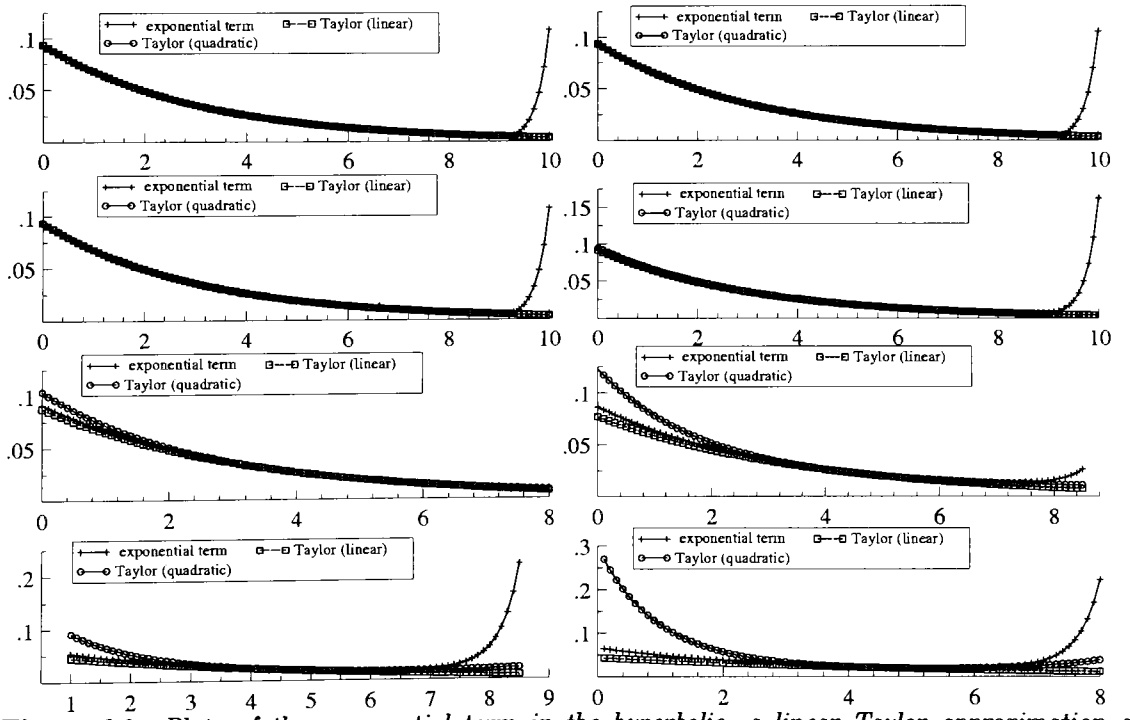


Figure 6.9: *Plots of the exponential term in the hyperbolic, a linear Taylor approximation and the quadratic Taylor approximation are shown for different values of θ drawn against x. For data within a certain range, it may be more parsimonious to use a Taylor approximation and thereby reducing the number of parameters to be estimated.*

OE structures of θ for the hyperbolic					
α	β	δ	μ	σ	$l(x;\theta)$
4.4875	-3.8412	1e-5	9.2472	0.0047	1256.8
4.4875	-3.8412	1e-4	9.2522	0.0047	1256.8
4.4875	-3.8412	0.001	9.2612	0.0047	1256.8
4.4875	-3.8412	0.101	9.2525	0.0047	1256.8
4.4875	-3.8412	0.201	9.2173	0.0047	1256.8
4.4875	-3.8412	0.301	9.1780	0.0047	1256.8
4.4875	-3.8412	0.401	9.1174	0.0047	1256.8
4.4875	-3.8412	0.501	9.0473	0.0047	1256.8
4.4875	-3.8412	0.601	8.9384	0.0047	1256.8
4.4875	-3.8412	0.701	8.8245	0.0047	1256.8
4.4875	-3.8412	0.801	8.6844	0.0047	1256.7
4.4875	-3.8412	0.901	8.4835	0.0047	1256.6
4.4875	-3.8412	1.001	8.2631	0.0047	1256.5
4.4875	-3.8412	1.101	7.9401	0.0047	1256.2
4.4875	-3.8412	1.1949	7.5397	0.0047	1255.6
4.4875	-3.8412	1.201	7.5070	0.0047	1254.2

Table 6.8: *Values of θ which are observationally equivalent; values of μ were estimated using reduced MCMC runs for specified values of δ, while (α, β, σ) were held fixed. The approximate log-likelihood for each specification is around 1256.8. The data was simulated with true values* $(4.4875, -3.8412, 1.1949, 7.2915, 0.0047)$, *using* $R = 10{,}000$, $\Delta^{\dagger} = 1$ *and* $T = 500$.

so that

$$f(x) \approx f(\hat{x}) + (x - \hat{x})f'(\hat{x}) + (x - \hat{x})^2 f''(\hat{x}),$$

where $\hat{x}$ is the mean of x. A linear or quadratic approximation is shown to approximate the more complicated structure in Figure 6.9 for certain ranges of the data, suggesting that for some datasets the exponential form $a + bx$ may suffice.

Figures 6.10–6.11 show the approximate profile log-likelihood plotted against δ. Data for Figure 6.10 is simulated using the Milstein scheme with $R = 10{,}000$, $\Delta^{\dagger} = 1$ and $T = 500$. The parameter values are given by $\alpha = 4.4878$, $\beta = -3.8412$, $\delta = 1.1949$, $\mu = 7.2915$ and $\sigma = 0.0047$, as in the paper by Bibby and Sørensen (1997). The minimum observation is given by 4.3682, the maximum by 5.3467 and the mean of the observations is 4.8604. Data for Figure 6.11 is simulated using parameter values $\alpha = 4.4878$, $\beta = -3.8412$, $\delta = 1.1949$, $\mu = 7.2915$ and $\sigma = 0.0047$. The minimum observation is given by 5.7160, the maximum by 7.0298 and the mean of the observations is 6.2995. In the estimation procedure, the parameters α, β and σ were held fixed as μ was estimated through reduced MCMC runs, (200 burn-in, 1000 iterations, saving every tenth), for different values of $\delta \in (0, 1.8]$. The posterior mean is used to compute the approximate log-likelihood using Pedersen's method with $M = 3$ for 1000 iterations. Figure 6.10 shows the case of a flat likelihood corresponding to different values of (δ, μ), whereas in Figure 6.11, unique values of (δ, μ) correspond to the maximum of the log-likelihood. As the distribution is clearly non-symmetric, we compute the Likelihood Ratio (LR) confidence region instead of the usual confidence

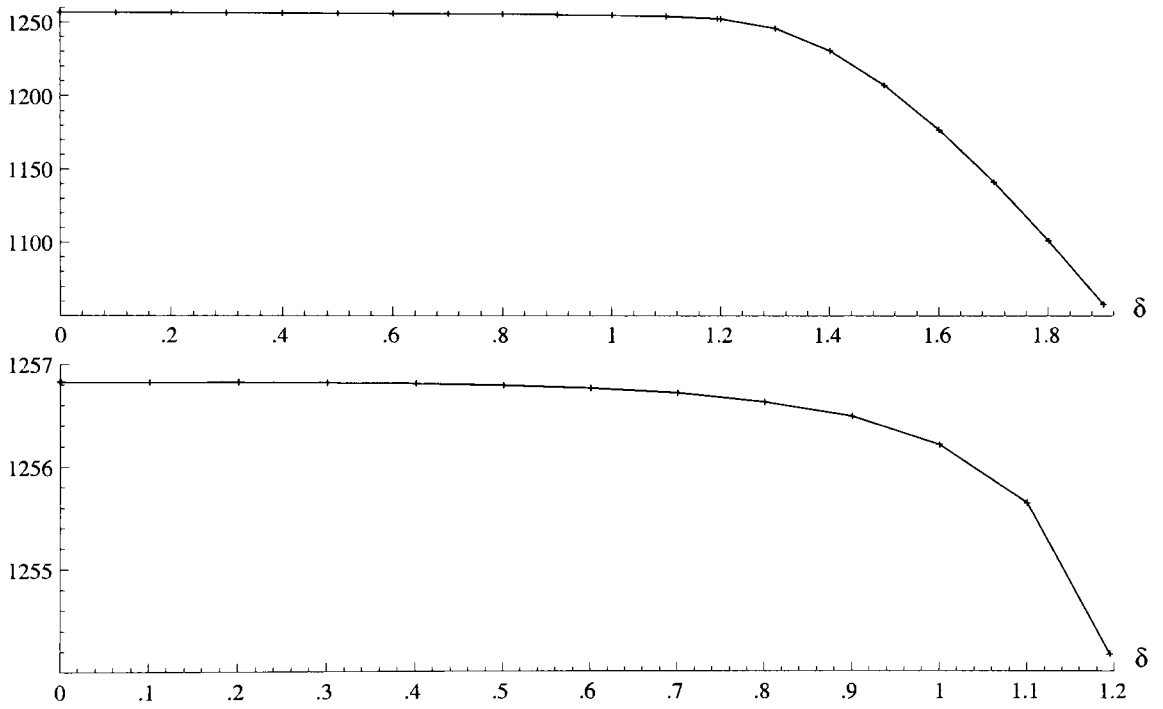


Figure 6.10: *Profile log-likelihood plotted against δ. The parameters α, β and σ were held fixed as μ was estimated through reduced MCMC runs, (200 burn-in, 1000 iterations, saving every tenth), for different values of $\delta \in (0, 1.8]$ in the top panel and $\delta \in (0, 1.2]$ in the bottom panel. The posterior mean is used to compute the approximate log-likelihood using Pedersen's method with $M = 3$ for 1000 iterations. Simulated data is in the region [4.3, 5.3].*

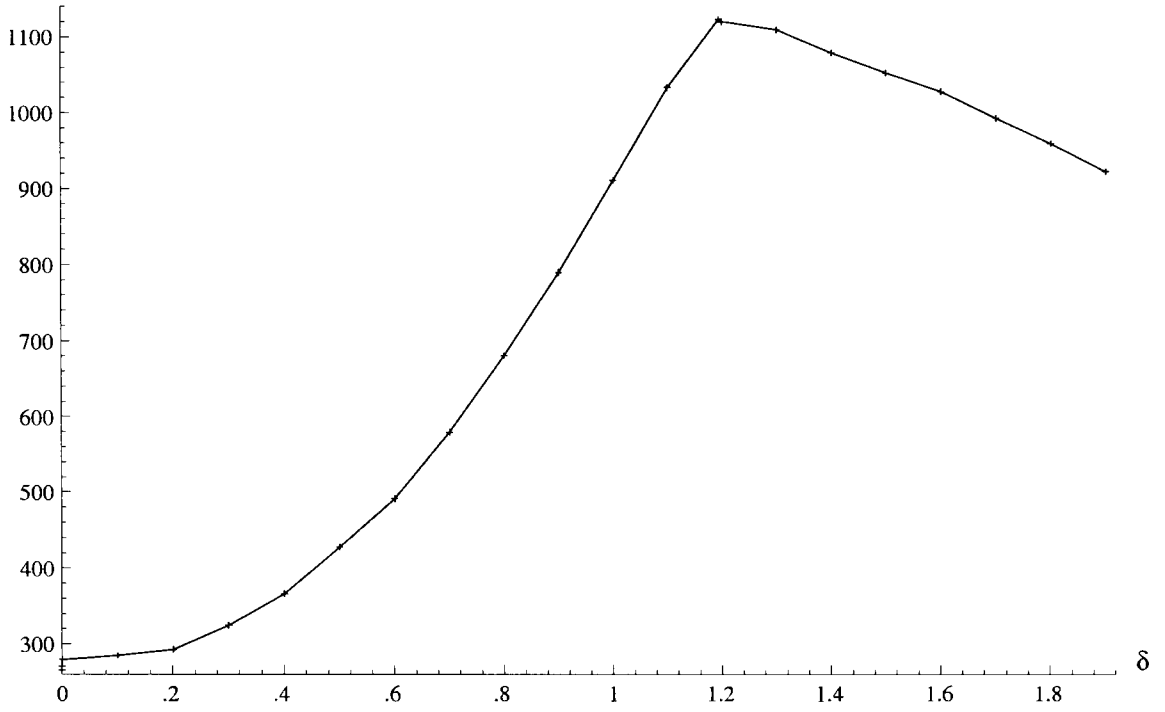


Figure 6.11: *Profile log-likelihood plotted against δ. The parameters α, β and σ were held fixed as μ was estimated through reduced MCMC runs, (200 burn-in, 1000 iterations, saving every tenth), for different values of $\delta \in (0, 1.8]$. The posterior mean is used to compute the approximate log-likelihood using Pedersen's method with $M = 3$ for 1000 iterations. Data was simulated in the region of [5.7, 7.02].*

interval. The LR test is based on the hypothesis

$$\begin{aligned} H_0 : \theta &= \theta_0 \\ H_1 : \theta &\neq \theta_0 \end{aligned}$$

and is constructed as

$$2\{l(x;\hat{\theta}) - l(x;\theta_0)\} < 3.841$$

where the right hand side is the value of χ^2_1 at the 95% level and $\hat{\theta}$ is the MLE. This implies that we do not reject θ for values of the likelihood greater than 1254.8795, which corresponds (see Table 6.8) to all values of $\delta \leq 1.1949$.

Further graphs are shown in 6.12–6.14, confirming the points made above. In 6.12, a contour plot of δ is drawn against μ, where α, β and σ are fixed at the true values for the simulated hyperbolic dataset, using the Milstein scheme with $\Delta^\dagger = 1$, $T = 500$ observations. These show that there may be a non-linear relationship between δ and μ. The graphs on the bottom panel show the same contour plots, for a different dataset; values generated in the region of 7. No clear relationship between δ and μ is implied from the graph. 6.13 uses a simulated version of the Baltica data,[2] where the true parameter values are given by, $\alpha = 4.4875$, $\beta = -3.8412$, $\delta = 1.1949$, $\mu = 7.2915$ and $\sigma = 0.0047$. The figure shows plots of α

[2]The Baltica dataset is comprised of daily stock returns from a Danish insurance company and Aarhus Oliefabrik, a large producer of cooking oils.

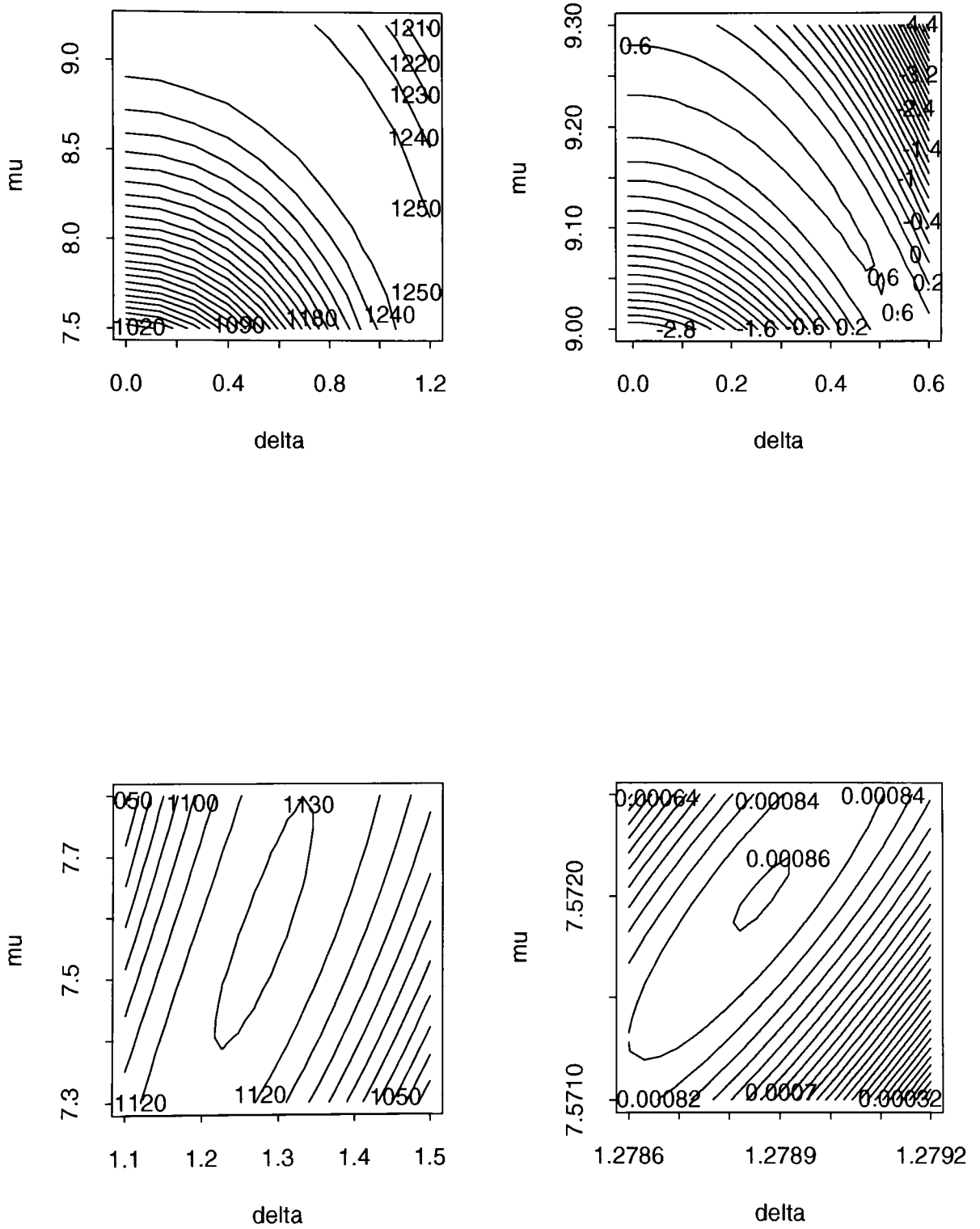


Figure 6.12: *Top graph: Contour plot of δ against μ, where α, β and σ are fixed at the true values for the simulated hyperbolic dataset, using the Milstein scheme with $\Delta^{\dagger} = 1$, $T = 500$ observations. Graph shows a non-linear relationship between δ and μ.Left graph shows the region $[0.00001, 7.5]$ to $[1.2, 9.3]$, right shows a smaller region $[0.00001, 9.0]$ to $[0.6, 9.3]$, where 1257 has been subtracted. Bottom graph: Contour plot of δ against μ, where α, β and σ are fixed at the true values for the simulated hyperbolic dataset, using the Milstein scheme with observations in the region 7, $\Delta^{\dagger} = 1$, $T = 500$ observations. Left graph*

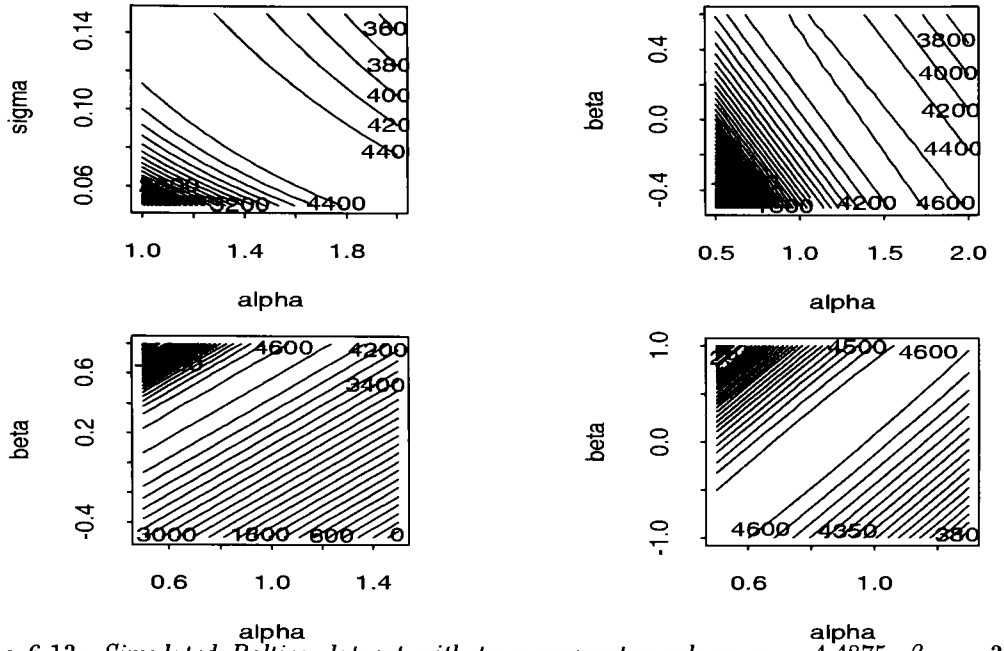


Figure 6.13: *Simulated Baltica dataset with true parameter values,* $\alpha = 4.4875$, $\beta = -3.8412$, $\delta = 1.1949$, $\mu = 7.2915$ *and* $\sigma = 0.0047$. *The figure shows plots of* α *against* β, *with the remaining parameters varying between plots, to show that the two parameters have a linear relationship between them except for first panel where dependence is slightly curved. Top panel uses true values, second panel uses* $\delta = 1.1949, \mu = 7.2915$ *and* $\sigma = 0.0047$, *third panel uses* $\delta = 0.1, \mu = 2$ *and* $\sigma = 0.01$ *and fourth panel uses* $\delta = 2, \mu = 5$ *and* $\sigma = 0.008$.

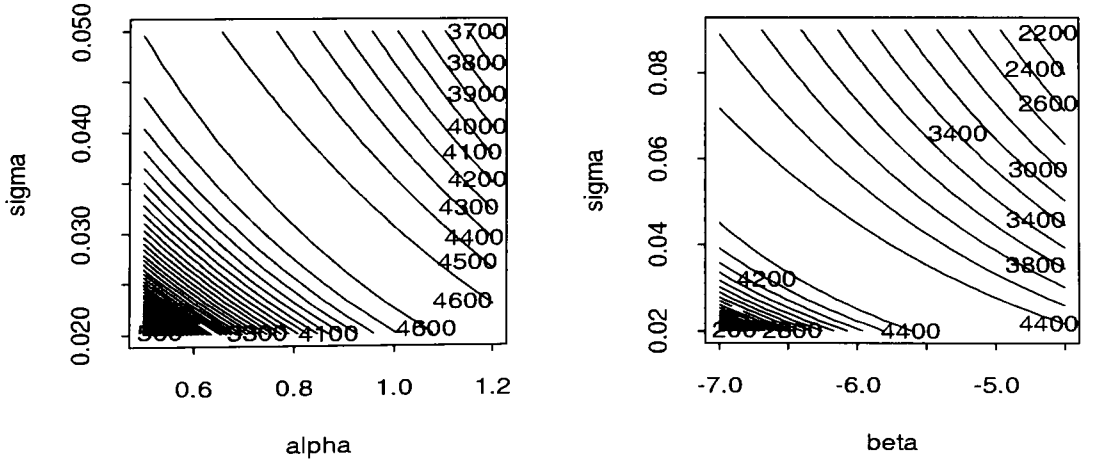


Figure 6.14: *Contour plot of* σ *against* α *(left) and* σ *against* β *(right), where* δ *and* μ *are fixed, but* β *varies linearly with* α *according to the relationship* $\beta = (2/3)\alpha - (7/3)$. *These graphs show, that* σ *will be difficult to obtain, eventhough the number of parameters has been decreased by one.*

against β, with the remaining parameters varying between each plot, but fixed within plots. The graphs seem to suggest a linear relationship between α and β, except in the first panel where dependence is slightly curved. Figure 6.14 shows contour plots of σ against α (left) and σ against β (right), where δ and μ are fixed, but β varies linearly with α according to the relationship $\beta = (2/3)\alpha - (7/3)$. These graphs show, that σ may be difficult to obtain, eventhough the number of parameters has been decreased by one.

Results show that for certain ranges of the data, the model is overparameterised and obtaining unique estimates of the parameters may not be possible. Dependence structures between δ and μ, and α and β seem to suggest that the model could be reduced to 3 parameters. However, as shown in figure 6.14, there may still be a range of values for σ, consistent with certain parameterisations of the remaining parameters.

6.6 Conclusion

The procedure for obtaining the posterior estimates of the parameters with applications using simulated data was outlined above. An additional discussion on updating all latent states and the hyperbolic diffusion was also provided. The Affine CIR and general parametric model were defined in Section 6.4 and will be considered in the next Chapter when we apply the methods developed so far to the Eurodollar deposit spot-rate.

CHAPTER 7

APPLICATION TO INTEREST-RATE DATA

This Chapter considers various parametric models for the 7-day Eurodollar deposit spot-rate presented in Section 7.1. Models applied to the data are presented in Section 7.2 and include the Vasicek model, the CIR, the Affine CIR and a general parametric model considered as an alternative to the one studied in Aït-Sahalia (1996b). The models are non-nested, though the Vasicek and Affine CIR are nested within the general parametric model.

The results are presented in Section 7.3. The aim is to capture the correlation and clustering structures found in the data. Both features are typical of financial data and have consistently been reported in the literature. Clustering implies that large volatility movements are followed by a volatility change similar in size and direction to the previous one, (see for example Biais, Hillion, and Spatt 1995 for an analysis in relation to the Paris Bourse stock exchange). An additional experiment is carried out in Section 7.4, where the whole block of auxiliary points is treated in one sweep. Extensions of the ideas presented in the thesis, such as endogenising the choice of M is considered in Section 7.5 and the Chapter concludes with Section 7.6.

7.1 THE DATA

The data concerns the 7-day Eurodollar deposit spot rate; the bid-ask midpoint is the short-term rate used.[1] Figure 7.1 provides a time-series of the $T = 5505$ observations. The data depicts higher interest rate levels for the 1980-1982 period due to a shift in monetary policy. In Aït-Sahalia (1996b), the main source of misspecification is stated to arise from the requirement of linearity in the drift coefficient. A more plausible interpretation would specify a non-linear mean reversion. Under this specification, the drift is essentially zero when the interest rate is between $4 - 7$ percent, but is pulled back to this region if rates enter other parts of the domain. The volatility function is lower in the middle region, exhibiting the *smile* pattern, being higher at either end of the spectrum. The data would seem to support this view, as opposed to the uniformly increasing smile pattern, as in the *Constant Elasticity of Variance* (CEV) model, proposed in the literature. The data is typical of interest rate data with short maturity in its infrequent but considerably sized jumps, as noted in Tauchen (1997), which could be due to liquidity or microstructure effects. Tauchen (1997) proposes a semi non-parametric approach taking the form of a truncated Hermite series expansion. The leading term in this series is specified as an AR(1)-ARCH(r)

[1] The data can be downloaded from http://gsbwww.uchicago.edu/fac/yacine.aitsahalia.

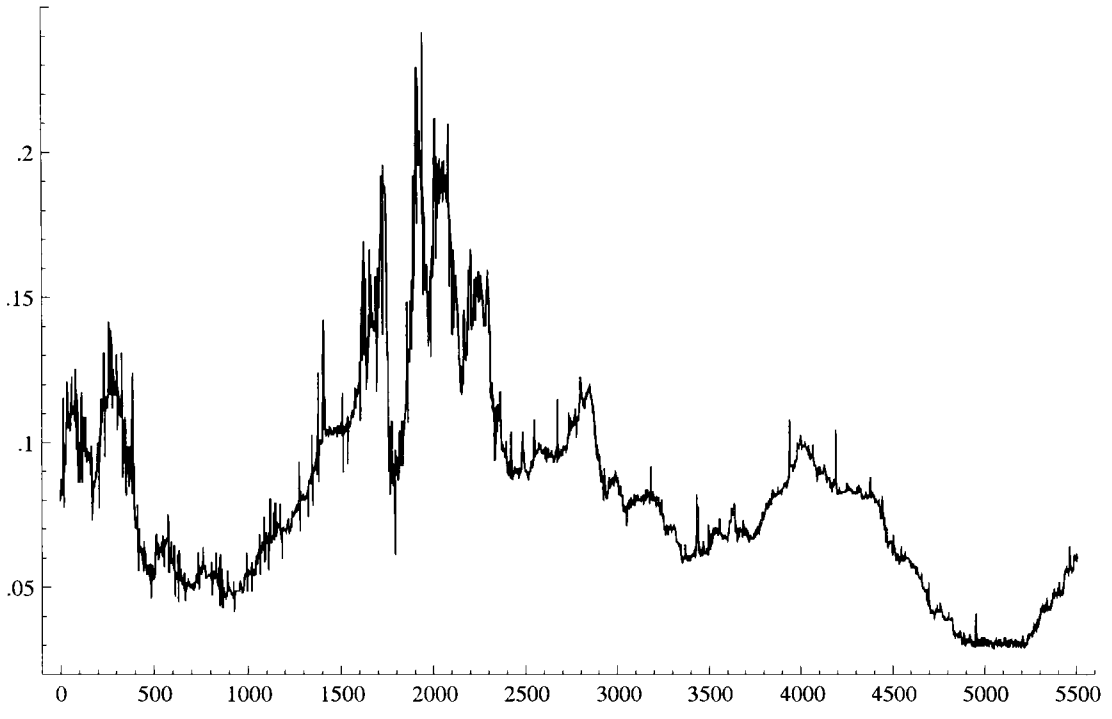


Figure 7.1: *Time-series plot of the 7-day Eurodollar deposit short-rate, from 1st of June 1973 to 25th of February, 1995.*

model, that is, a linear autoregression of length l and ARCH error structure of length r. An empirical study of various models applied to the short-term interest rate have been considered by Chan, Karolyi, Longstaff, and Sanders (1992).

7.2 Parametric specifications of the data to be considered

Suppose we wish to analyse the following parametric specification of the short rate process, studied in Aït-Sahalia (1996b). The diffusion functions in (6.15) are given by

$$\begin{aligned} a(y;\theta) &= \alpha_0 + \alpha_1 y + \alpha_2 y^2 + \frac{\alpha_3}{y} \\ b^2(y;\theta) &= \beta_0 + \beta_1 y + \beta_2 y^{\beta_3} \end{aligned} \tag{7.1}$$

where $\theta = (\alpha_0, \alpha_1, \alpha_3, \beta_0, \beta_1, \beta_2, \beta_3)'$ and

$\beta_0 \geq 0$ (and $\beta_2 > 0$ if $\beta_0 = 0$ and $0 < \beta_3 < 1$, or $\beta_1 > 0$ if $\beta_0 = 0$ and $\beta_3 > 1$)

$\beta_2 > 0$ if either $\beta_3 > 1$ or $\beta_1 = 0$, and $\beta_1 > 0$ if either $0 < \beta_3 < 1$ or $\beta_2 = 0$

$\alpha_2 \leq 0$ and $\alpha_1 < 0$ if $\alpha_2 = 0$

$\alpha_3 > 0$ and $2\alpha_3 \geq \beta_0 \geq 0$, or $\alpha_3 = 0, \alpha_0 > 0$, $\beta_0 = 0$, $\beta_3 > 1$ and $2\alpha_0 \geq \beta_1 > 0$.

The first two equations are necessary for b^2 to be positive, the third ensures that the drift is mean reverting at high interest values while the last equation rules out the zero case. Note that instead of letting β_3 be restricted to the region $(0,1) \cup (1,\infty)$, we could let $\beta_3 \in (1,\infty)$ given the characteristics of the data at hand (the sample data contains $T = 5505$ observations with a mean of 0.083621 and variance of 0.0012893. The observations range from 0.029150 to 0.24333). The other model specifications are outlined in Table 7.1. One reason we consider these alternative model specifications is because the above model is practically unidentified for these data, as will be shown below.

Model	Drift function $a(y;\theta)$	Volatility function $b^2(y;\theta)$
Vasicek	$\alpha_0 + \alpha_1 y$	β_0
CIR	$\alpha_0 + \alpha_1 y$	$\beta_1 y$
Affine CIR	$\alpha_0 + \alpha_1 y$	$\beta_0 + \beta_1 y$
Aït-Sahalia (1996b)	$\alpha_0 + \alpha_1 y + \alpha_2 y^2 + \frac{\alpha_3}{y}$	$\beta_0 + \beta_1 y + \beta_2 y^{\beta_3}$
General parametric model	$\alpha_0 + \alpha_1 y + \alpha_2 y^2 + \frac{\alpha_3}{y}$	$\beta_0 + \beta_1 y + \beta_2 y^2 + \beta_3 y^3$

Table 7.1: *Models to be considered for the short-rate process: Vasicek, CIR, Affine CIR, Aït-Sahalia (1996b) model, and general parametric model.*

7.3 Results

Our first set of results are for the Vasicek, CIR and Affine CIR models, where we have fit the model after taking logs and applying the Itô transformation. The results are given in Tables 7.2, 7.3 and 7.4 where we report the posterior means, Monte Carlo standard errors, inefficiency factors and the covariances and correlations of the parameters. The *Highest Probability Density* (HPD) regions are also reported.[2] These results are based on $M = 10$ ($\lambda = 3$) using $N = 2000$ MCMC iterations for the Vasicek and CIR models and $N = 5000$ for the Affine CIR model. For each of these models, the output paths, correlograms and histograms of $\theta = (\alpha_0, \alpha_1, \beta_1)'$ are given in Figures 7.2, 7.3 and 7.4. We selected $B_N = 100$ (400) in the Parzen computations of the Monte Carlo standard errors for α (β). We set standard diffuse normal priors on the alpha coefficients and diffuse inverse Gamma priors on β. In the Affine CIR and general parametric model, however, β_1 is given a diffuse normal prior.

Table 7.2: *The posterior means, HPD, Monte Carlo standard errors, inefficiency factors (Ineff, bandwidth $B_N = 100$), covariances and correlations (in italics) are shown for the parameters using the Eurodollar short-rate data applied to the Vasicek process. The M-H algorithm was run with $M = 10$ ($\lambda = 3$), for $N = 2000$ iterations.*

Parameters of Vasicek applied to Eurodollar short-rate data								
	Mean	HPD		MCse	Ineff	Covariance & *Correlation*		
$M = 10$								
$\alpha_0 \vert y$	0.000563	0.000271	0.000821	8.74e-08	1.2	1.98e-08	*-0.917*	*0.065*
$\alpha_1 \vert y$	-0.00673	-0.00972	-0.00377	8.85e-07	1.2	-1.97e-07	2.33e-06	*-0.080*
$\beta_0 \vert y$	1.66e-05	1.60e-05	1.71e-05	3.37e-09	22.4	2.73e-12	-3.67e-11	9.04e-14

[2]The α% HPD region represents the shortest interval that contains α% of the points of highest posterior density, for example, see Besag, Green, Higdon, and Mengersen (1995).

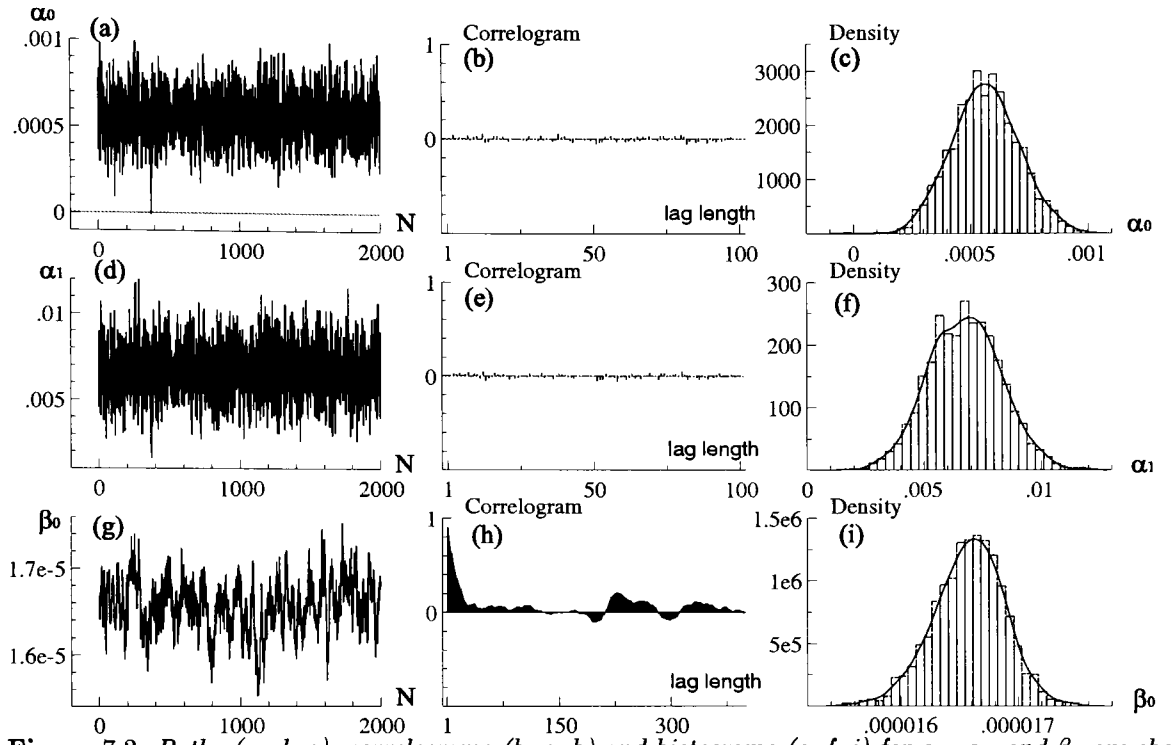


Figure 7.2: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for* α_0, α_1 *and* β_0 *are shown using the Vasicek process applied to the Eurodollar short-rate. The M-H algorithm was run with* $M = 10$ *(*$\lambda = 3$*), for* $N = 2000$ *iterations.*

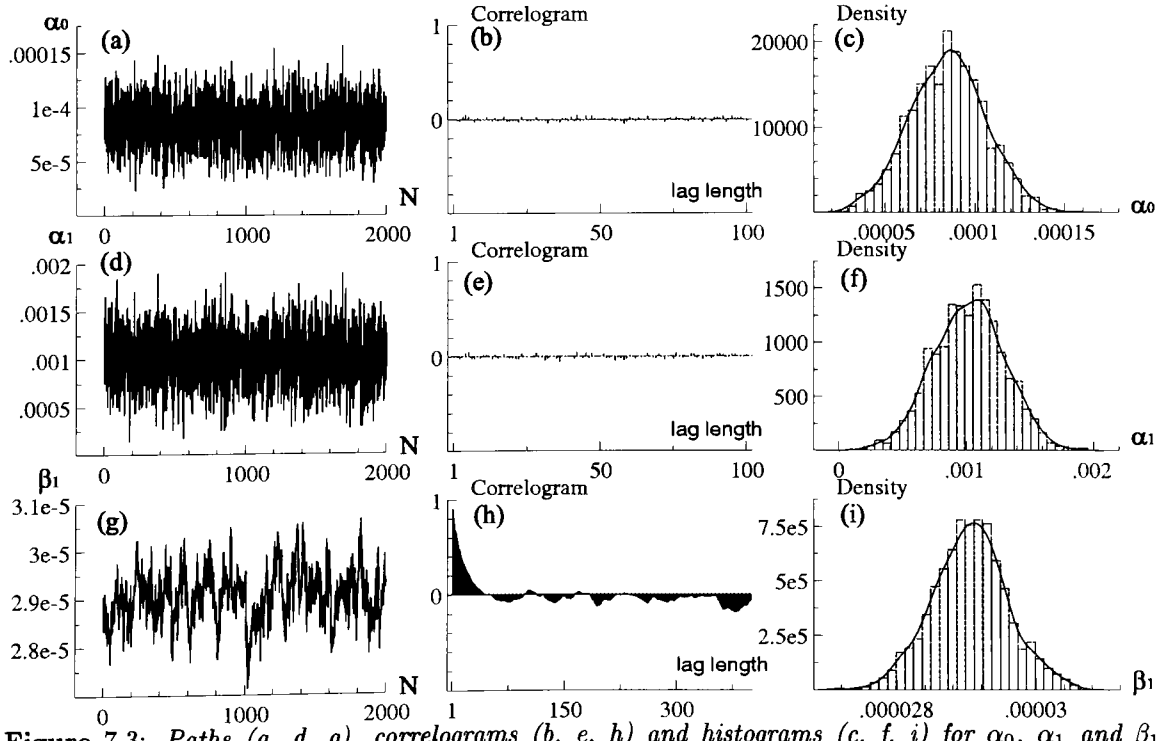


Figure 7.3: *Paths (a, d, g), correlograms (b, e, h) and histograms (c, f, i) for* α_0, α_1 *and* β_1 *are shown using the CIR process applied to the Eurodollar short-rate data. The M-H algorithm was run with* $M = 10$ *(*$\lambda = 3$*), for* $N = 2000$ *iterations.*

Table 7.3: *The posterior means, HPD, Monte Carlo standard errors, inefficiency factors (Ineff, bandwidth $B_N = 100$), covariances and correlations (in italics) are shown for the parameters using the CIR process applied to the Eurodollar short-rate data. The M-H algorithm was run with $M = 10$ ($\lambda = 3$), for $N = 2000$ iterations.*

Parameters of CIR applied to Eurodollar short-rate data								
	Mean	HPD		MCse	Ineff	Covariance & *Correlation*		
$M = 10$								
$\alpha_0 \mid y$	0.00042392	0.00022572	0.00063118	4.6733e-008	0.9	1.09e-08	*-0.911*	*0.085*
$\alpha_1 \mid y$	-0.0051101	-0.00246	-0.00779	5.99e-07	0.9	-1.32e-07	1.92e-06	*0.0797*
$\beta_1 \mid y$	0.000145	0.000140	0.000150	3.06e-08	22.8	2.38e-11	2.96e-10	7.19e-12

It will be seen that for each model the correlogram of α dissipates quickly, though that of β_1 in the Affine CIR model shows persistence up to lag 100. In all models, the posterior mean of α_0 is positive and that of α_1 is negative. The results on β_0 in the Vasicek model and those of β_1 in the CIR model are consistent with those in the Affine CIR model. The Figures show that β_0 is close to zero compared to β_1, but also clearly distinct from zero given the histogram, and that the variance is mainly determined by $\beta_1 y$. The α and β parameters are close to being uncorrelated. One point to note is that the inefficiency factors are low for α. For β, we are up to 25 times less efficient.

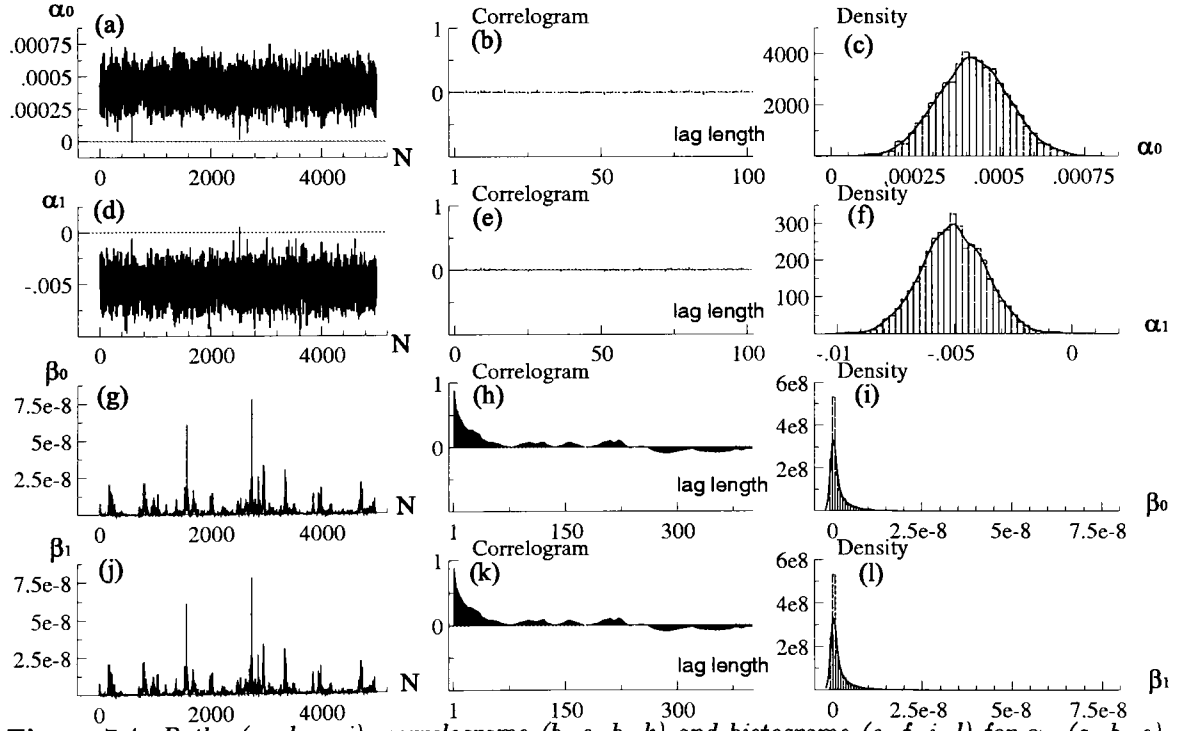


Figure 7.4: *Paths (a, d, g, j), correlograms (b, e, h, k) and histograms (c, f, i, l) for α_0 (a, b, c), α_1 (d, e, f), β_0 (g, h, i) and β_1 (j, k, l) are shown using the Affine CIR process applied to the Eurodollar short-rate data. The M-H algorithm was run with $M = 10$ ($\lambda = 3$), for $N = 5000$ iterations.*

In Figure 7.5 we plot the standardised forecast errors, defined in Section 3.3.3, for all models. A plot of the autocorrelation functions is provided in Figure 7.6 with the corresponding QQ plots. The residuals and their reflected versions are computed using the one-step-ahead prediction distribution. The standardised forecast errors show that the extremes in the data are not being picked up for a large part of the data. This is further confirmed by the correlograms, which show remaining structure not being accounted for.

Table 7.4: *The posterior means, HPD, Monte Carlo standard errors, inefficiency factors (Ineff, bandwidth $B_N = 100$), covariances and correlations (in italics) are shown for the parameters using the Affine CIR process applied to the Eurodollar short-rate data. The M-H algorithm was run with $M = 10$ ($\lambda = 3$), for $N = 5000$ iterations.*

Parameters of Affine CIR applied to Eurodollar short-rate data									
	Mean	HPD		MCse	Ineff	Covariance & *Correlation*			
$M = 10$									
$\alpha_0 \mid y$	0.000420	0.000202	0.000607	2.84e-08	1.4	1.07e-08	*-0.908*	*0.00202*	*0.0766*
$\alpha_1 \mid y$	-0.00506	-0.00770	-0.00245	3.39e-07	1.3	-1.27e-07	1.82e-06	*-0.00439*	*-0.0728*
$\beta_0 \mid y$	1.92e-09	4.10e-12	7.95e-09	1.92e-11	25.4	7.81e-16	-2.24e-14	1.43e-17	*0.0323*
$\beta_1 \mid y$	0.000145	0.000140	0.000150	1.21e-08	20.2	2.37e-11	-2.94e-10	3.65e-16	8.96e-12

The amount of activity present signifies a lack of fit, reflecting the fact that these Markov models do not account for the volatility clustering present in this series.

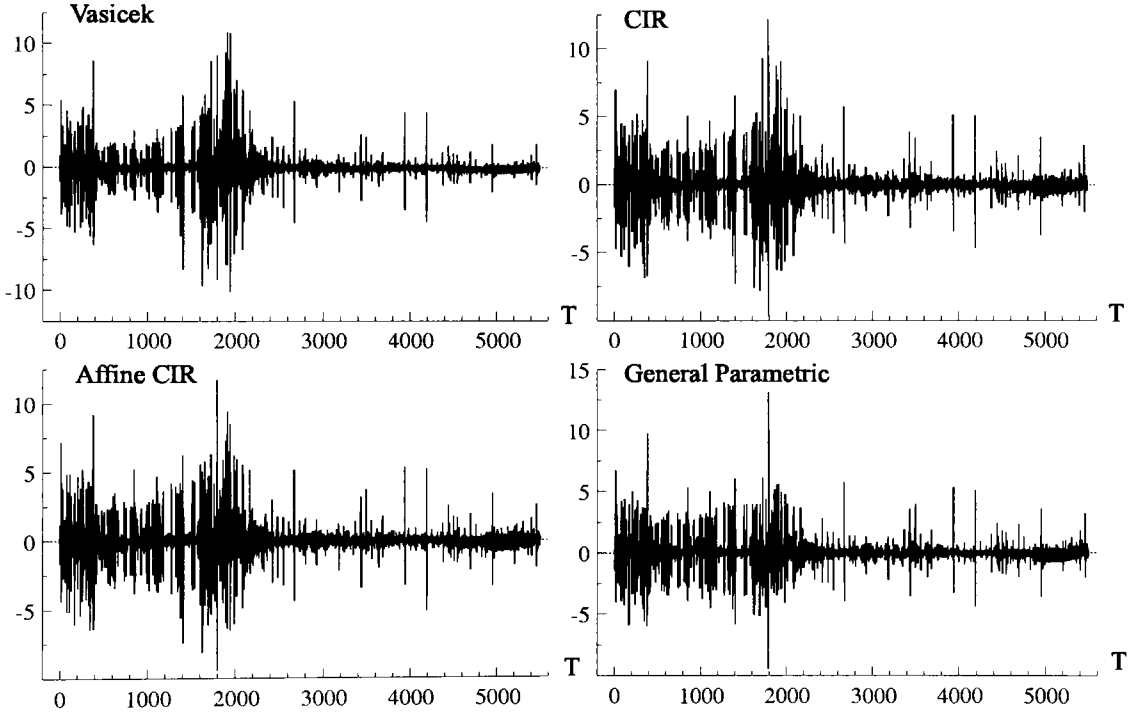


Figure 7.5: *Standardized one-step-ahead forecast errors for the Vasicek, CIR, Affine CIR and general parametric models. We graph* $y_{t+1} - \hat{E}(y_{t+1}|y_t;\theta)/\sqrt{var(y_{t+1}|y_t,\theta)}$, *taking* $\theta = E(\theta|y)$.

We now turn our attention to the general parametric model given by equation 6.15. We first show that β_2 and β_3 are difficult to identify separately from our data. In Figure 7.7 we show a plot of $\beta_2 y^{\beta_3}$ against a range of y values determined from the data. The function is plotted for five pairs of values of (β_2, β_3) : $(4.511 \times 10^{-4}, 1.5)$, $(6.9338 \times 10^{-4}, 1.7)$, $(7.021 \times 10^{-4}, 1.8)$, $(1.312 \times 10^{-3}, 2)$ and $(1.5177 \times 10^{-3}, 2.1)$. These were obtained from the MCMC algorithm fixing a value of β_3 and evaluating the posterior mean of β_2. In the bottom graph we plot the kernel density estimate of the y data. We see that for the range of y values with mass (essentially values of y below 0.16), the five different pairs of (β_2, β_3) values produce virtually the same value of $\beta_2 y^{\beta_3}$. On the basis of this analysis, see also Tauchen (1997), we estimate an alternative model in which the volatility specification is cubic, preserving the quadratic shape exhibited

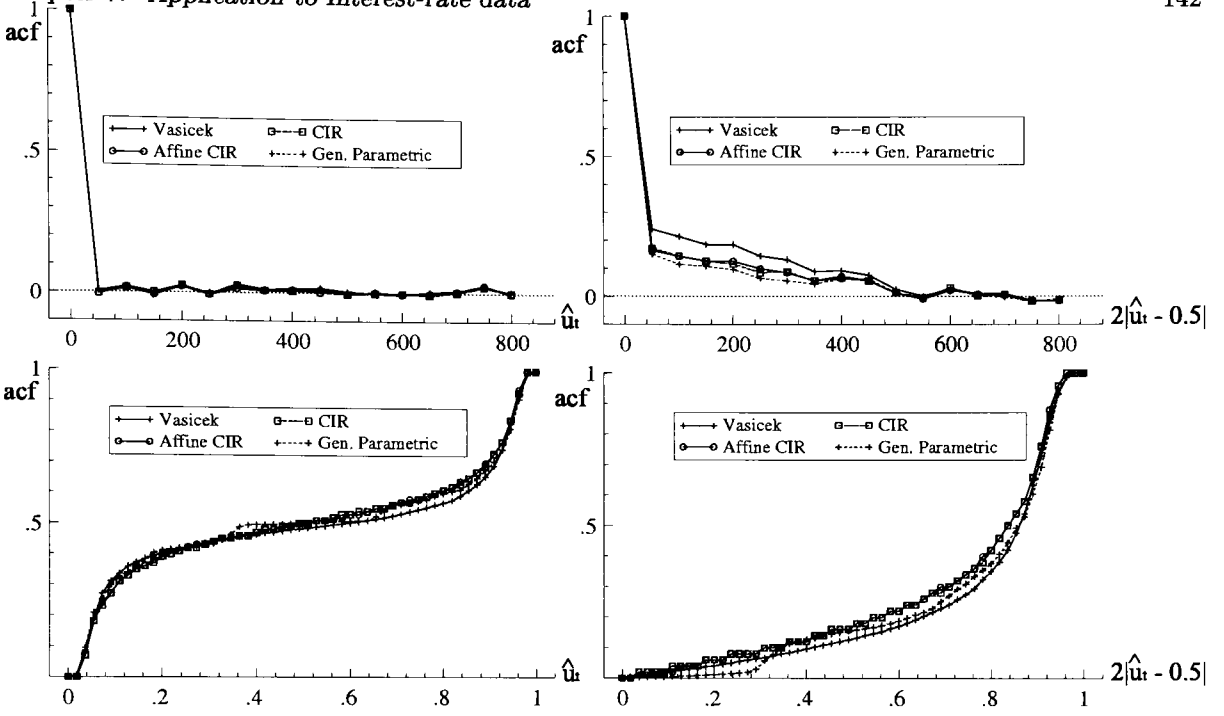


Figure 7.6: *Top graphs: Autocorrelation functions of residuals, $\{\hat{u}_t\}$ (left panel) and reflected residuals, $\{2|\hat{u}_t - 0.5\}$ (right panel) for the Vasicek, CIR, Affine CIR and general parametric process. Bottom graphs: The corresponding QQ plots against observations.*

above

$$\beta_0 + \beta_1 y + \beta_2 y^2 + \beta_3 y^3,$$

listed in Table 7.1, as suggested by Figure 7.7. Results are shown in Table 7.5. Similar diagnostics were obtained on setting $\beta_3 = 0$. Corresponding posterior means, correlograms and histograms are shown in Figures 7.8 and 7.9. Standardized forecast errors in Figure 7.5 show that the extremes in the data are not being picked up, and the autocorrelation functions in Figure 7.6 confirm that the models have not provided a good fit to the data.

To complete the analysis we also compute the marginal likelihoods of these various models, where the method is outlined in Section 8. These are computed using the importance sampler to evaluate $f(y|\theta)$, using $M = 10$ auxiliary points, where all M are updated in one sweep. The log marginal likelihood computations are shown in Table 7.6. We infer that there is almost equal support in the data for the CIR and the Affine CIR models but that the general parametric model receives overwhelming support in relation to the other three models. Nonetheless, volatility clustering in the residuals suggests that each of these models should be elaborated to include a heavy tailed stochastic volatility component, perhaps along the lines of Andersen and Lund (1997).

7.4 MCMC ESTIMATION: UPDATING THE WHOLE BLOCK IN ONE SWEEP

The analysis is repeated except that the whole block of M auxiliary points is now updated in one sweep. We again consider the Vasicek, CIR, Affine CIR models and general parametric model, where we have fit the model after taking logs and applying the Itô transformation. The results are given in Tables

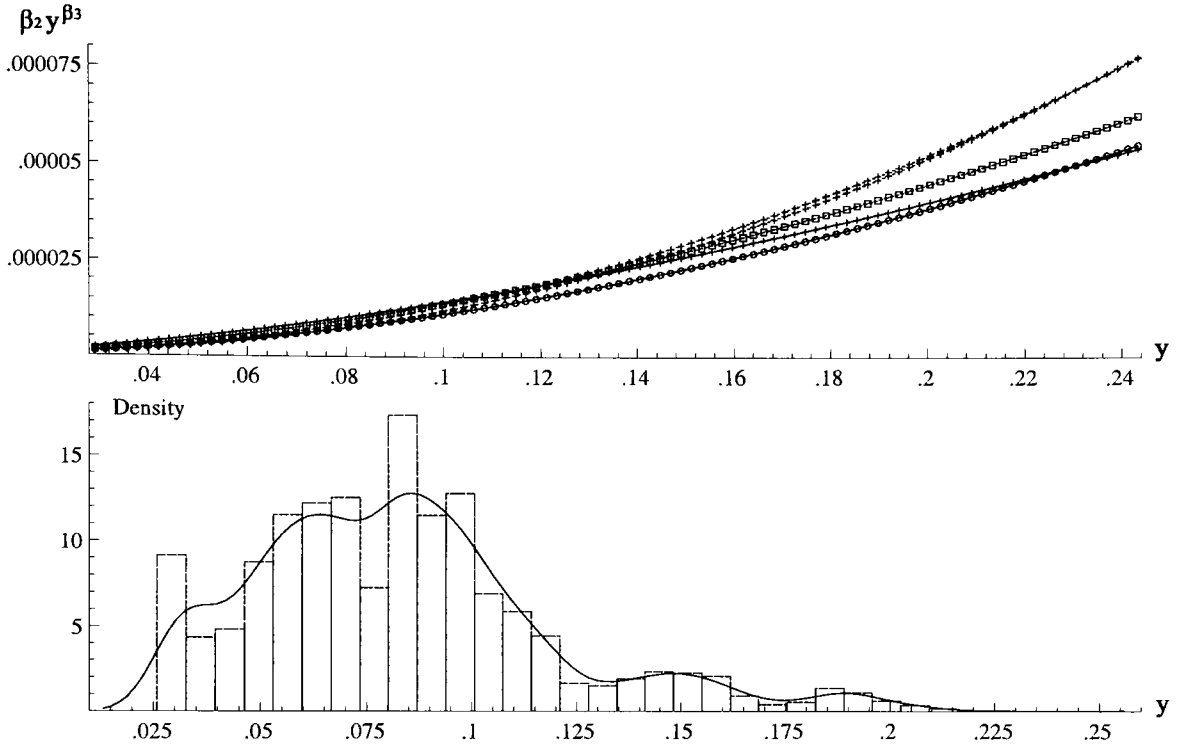


Figure 7.7: *Top graph plots $\beta_2 y^{\beta_3}$ against a range of values for y between the minimum and maximum values observed from the data, for five pairs of values of (β_2, β_3). Bottom graph plots the kernel estimate of the density using the interest data.*

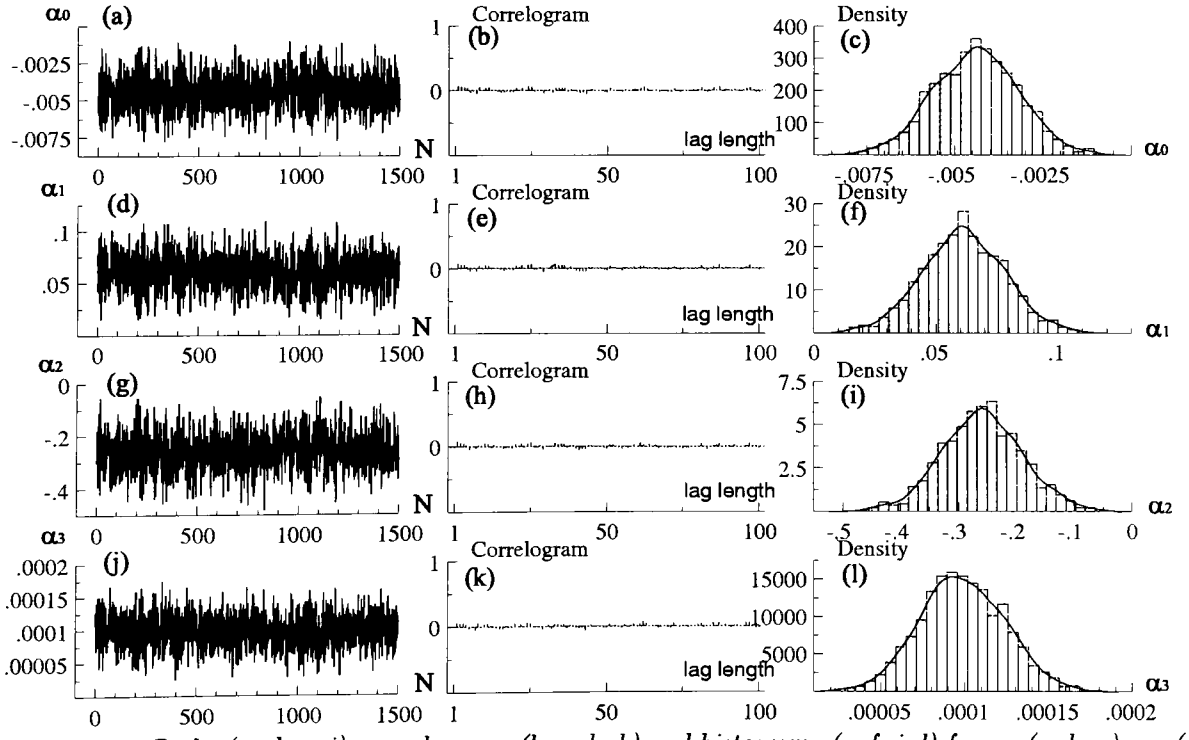


Figure 7.8: *Paths (a, d, g, j), correlograms (b, e, h, k) and histograms (c, f, i, l) for α_0 (a, b, c), α_1 (d, e, f), α_2 (g, h, i) and α_3 (j, k, l) are shown using the Eurodollar short-rate data applied to the general parametric process. The M-H algorithm was run with $M = 10$ $(\lambda = 3)$, for $N = 1500$ iterations.*

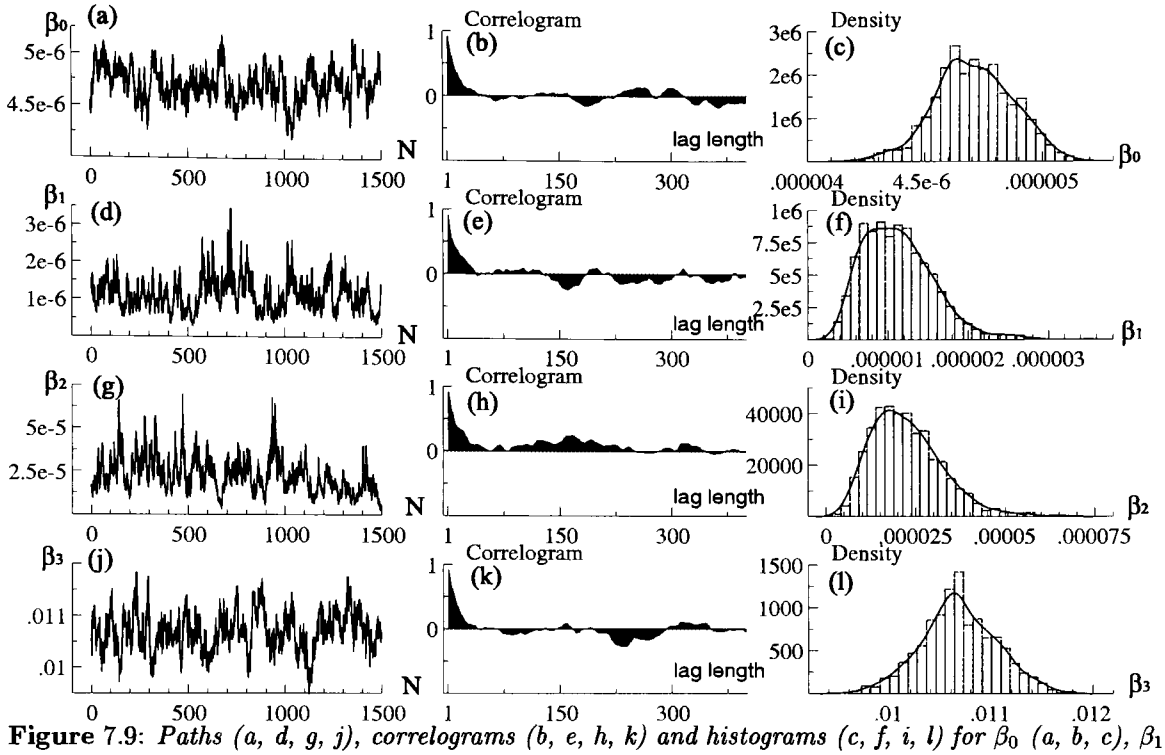


Figure 7.9: *Paths (a, d, g, j), correlograms (b, e, h, k) and histograms (c, f, i, l) for β_0 (a, b, c), β_1 (d, e, f), β_2 (g, h, i) and β_3 (j, k, l) are shown using the Eurodollar short-rate data applied to the general parametric process. The M-H algorithm was run with $M = 10$ ($\lambda = 3$), for $N = 1500$ iterations.*

7.7– 7.10, where we report the posterior means, Monte Carlo standard errors, inefficiency factors, HPD regions and the covariances and correlations of the parameters. These results are based on $M = 5$ using $N = 4000$ MCMC iterations. The bandwidth $B_N = 100$ (400) is used in the Parzen computations of the Monte Carlo standard errors for α (β). We set standard diffuse normal priors on the alpha coefficients and diffuse inverse Gamma priors on β. In the Affine CIR and general parametric model, however, β_1 is given a diffuse normal prior centred at zero.

Tables 7.7–7.9 show the posterior mean, MCse, HPD, inefficiency factors and covariance and correlation structures for the Vasicek, CIR and Affine CIR models applied to the Eurodollar data using $M = 5$, $N = 4000$. These results are very similar to those for the case where $M = 10$ with $\lambda = 3$. Inefficiency factors are lower for the volatility coefficients - for example, 2.14 for β_1 in the Vasicek model for $M = 5$, no blocks, as opposed to 22.4 for the case with blocks. Higher inefficiency factors are mainly due to the increased number of auxiliary points. For the general parametric model, (results shown in Table 7.10), we set $\beta_3 = 0$, (as there is little difference between the quadratic pattern exhibited by each), and estimate remaining parameters as in the case for sampling with blocks. Results are quite different for this new specification and on computing the marginal likelihood statistics, the model performs the worst in the class of models considered. These are shown in Table 7.11. The importance sampler was used to evaluate $f(x|\theta^*)$, using reduced MCMC runs with $M = 5$. θ^* is set at the posterior mean, (computed using $M = 10$, $\lambda = 3$ in the top panel and $M = 5$ in the bottom panel). In the latter case β_3 is set to zero.

Table 7.5: *The posterior means, standard errors, HPD, inefficiency factors (Ineff, bandwidth $B_N = 100$), covariances and correlations (in italics) are shown for the parameters using the Eurodollar short-rate data applied to the general parametric process. The M-H algorithm was run with $M = 10$ ($\lambda = 3$), for $N = 1500$ iterations.*

	Mean	MCse	HPD		Ineff
$M = 10$					
$\alpha_0\|y$	-0.00441	1.25e-06	-0.00674	-0.00215	1.6
$\alpha_1\|y$	0.0612	1.83e-05	0.0280	0.0924	1.7
$\alpha_2\|y$	-0.259	7.72e-05	-0.387	-0.118	1.7
$\alpha_3\|y$	9.85e-05	2.53e-08	4.97e-05	0.000145	1.5
$\beta_0\|y$	4.71e-06	2.32e-09	4.40e-06	5.04e-06	21.2
$\beta_1\|y$	1.13e-06	6.32e-09	3.90e-07	1.95e-07	21.6
$\beta_2\|y$	2.25e-05	1.61e-07	6.62e-06	4.21e-05	24.1
$\beta_3\|y$	0.00106	5.21e-06	0.00997	0.0115	21.0

Covariance & *Correlation*

	α_0	α_1	α_2	α_3	β_0	β_1	β_2	β_3
α_0	1.40e-06	*-0.983*	*0.928*	*-0.982*	*0.00393*	*-0.0208*	*0.0190*	*-0.0488*
α_1	-1.93e-05	0.000274	*-0.978*	*0.937*	*-0.0173*	*0.0347*	*-0.0230*	*0.0549*
α_2	7.61e-05	-0.00112	0.00480	*-0.861*	*0.0238*	*-0.0448*	*0.0284*	*-0.0594*
α_3	-2.88e-08	3.84e-07	-1.48e-06	6.13e-10	*0.0190*	*0.00429*	*-0.0170*	*0.0409*
β_0	7.68e-13	-4.72e-11	2.72e-10	7.76e-014	2.72e-14	*-0.185*	*-0.0962*	*-0.286*
β_1	-1.08e-11	2.52e-10	-1.36e-09	4.67e-14	-1.34e-14	1.93e-13	*-0.0182*	*-0.136*
β_2	2.25e-10	-3.80e-09	1.96e-08	-4.21e-12	-1.59e-13	-7.98e-14	9.99e-11	*-0.266*
β_3	-2.15e-08	3.38e-07	-1.53e-06	3.77e-10	-1.76e-11	-2.23e-11	-9.90e-10	1.39e-07

7.5 Extensions

A natural question arises as to the optimal choice of M or M_t for each t. Endogenising M_t, (depending on the conditional variance of y_t^*, the parameter values and the end-points), is discussed briefly in Section 9.6. Another method would be to compare the fit of the model with a more refined model based on $M_0 = 2M$.

Suppose we define $y_t^* = (y_t^{-\prime}, y_t^{+\prime})'$. Then

$$\begin{aligned} f(y_{t+1}, y_t^*|y_t; M_0) &= f(y_{t+1}|y_{t,M}^*)\left\{\prod_{j=2}^{M} f(y_{t,j}^*|y_{t,j-1}^*)\right\} f(y_{t,1}^*|y_t) \\ &= f(y_{t+1}|y_{t,M}^+)\left\{\prod_{j=2}^{M} f(y_{t,j}^+|y_{t,j}^-) f(y_{t,j}^-|y_{t,j-1}^+)\right\} f(y_{t,1}^+|y_{t,1}^-) f(y_{t,1}^-|y_t). \end{aligned}$$

If we further assume that the y_t^+ correspond to the missing data when M auxiliary points are used, then

$$\frac{f_{M_0}(y_{t+1}|y_t)}{f_M(y_{t+1}|y_t)} = \frac{\int\int f(y_{t+1}, y_t^+, y_t^-|y_t; M_0)\, dy_t^+ dy_t^-}{\int f(y_{t+1}, y_t^+|y_t; M)\, dy_t^+}.$$

Figure 7.10 shows the discretisation scheme using the more refined model. Since

$$\begin{aligned} f_M(y_{t+1}|y_t) &= \int f(y_{t+1}, y_t^+|y_t; M)\, dy_t^+ \\ &= \frac{f(y_{t+1}, y_t^+|y_t; M)}{f(y_t^+|y_{t+1}, y_t; M)} \end{aligned}$$

Table 7.6: *The log marginal likelihood computations are shown for* $x = \log y$. $f(x|\theta^*)$ *and* $\pi(\theta^*|x)$ *are computed using* $M = 10$, *burn-in=200 and* $N = 2000$. *Priors used are Inverse Gamma* $\left\{\frac{\theta^*}{(15\sigma^*)^2} + 2, \theta^*(\sigma^* - 1)\right\}$ *for* β *and Normal* $\{\theta^* + 4\sigma^*, (15\sigma^*)^2\}$ *for* α, *(*β_1 *uses normal specification in general parametric model), where* θ^* *and* σ^* *are the respective posterior means and standard deviations, based on a training sample data set. Numerical standard error of marginal likelihood (log scale) is denoted by (numerical se).*

Model	$\log m(x)$	Numerical se
Vasicek	8328.9	0.52857
CIR	9430.5	0.25093
Affine CIR	9436.9	0.26587
General parametric model	9716.8	0.60919

	Mean	MCse	HPD		Ineff
$M = 5$					
$\alpha_0\|y$	0.00059	6.1329e-08	0.00032	0.00086	1.69
$\alpha_1\|y$	-0.00703	8.3086e-07	-0.01010	-0.00431	2.15
$\beta_0\|y$	1.6579e-05	3.1393e-07	1.5969e-05	1.7209e-05	2.14

Covariance & *Correlation*		
2.0972e-08	*-0.92479*	*0.08214*
-2.0675e-07	2.3832e-06	*-0.081002*
3.7343e-12	-3.9256e-11	9.8550e-14

Table 7.7: *The posterior means, standard errors, HPD, inefficiency factors (Ineff), covariances and correlations (in italics) are shown for the parameters using the Eurodollar short-rate data applied to the Vasicek process. The MH algorithm was run with* $M = 5$, *for* $N = 4000$ *iterations with* 1000 *burn-in. The bandwidth used in the Parzen computations,* $B_N = 100$ *in all computations.*

is constant,

$$\frac{f_{M_0}(y_{t+1}|y_t)}{f_M(y_{t+1}|y_t)} = \int\int \frac{f(y_{t+1}, y_t^+, y_t^-|y_t; M_0)}{f(y_t^+, y_{t+1}|y_t, M)} f(y_t^+|y_{t+1}, y_t; M)\, dy_t^+ dy_t^- . \tag{7.2}$$

Suppose we denote the importance sampler for this problem by $r(y_t^-|y_t, y_t^+, y_{t+1}; M_0)$, then we can estimate the integral (7.2) by simulation as

$$\frac{1}{R}\sum_{j=1}^{R} \frac{f(y_{t+1}, y_t^{+(j)}, y_t^{-(j)}|y_t; M_0)}{f(y_{t+1}, y_t^{+(j)}|y_t; M)\; r(y_t^{-(j)}|y_t, y_t^{+(j)}, y_{t+1}; M_0)}$$

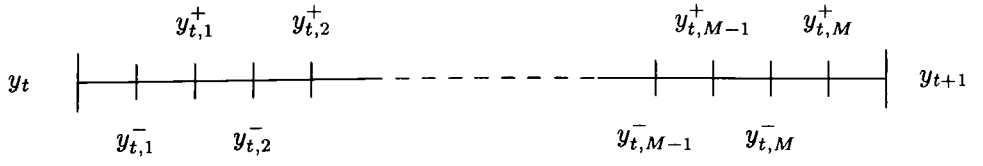


Figure 7.10: *Discretisation scheme using the more refined model, where* $y_t^* = \left(y_t^{-\prime}, y_t^{+\prime}\right)'$ *are introduced between observed values* y_t *and* y_{t+1}.

	Mean	MCse	HPD		Ineff
$M = 5$					
$\alpha_0 \vert y$	0.00042	6.3935e-09	0.00021	0.00060	0.25
$\alpha_1 \vert y$	-0.00502	1.4018e-07	-0.00245	-0.00779	0.42
$\beta_1 \vert y$	0.00015	1.3270e-09	0.00014	0.00015	1.79

Covariance & *Correlation*		
1.0560e-08	*-0.90502*	*0.16528*
-1.2526e-07	1.8139e-06	*0.11547*
5.0337e-11	4.6090e-10	8.7835e-12

Table 7.8: *The posterior means, standard errors, HPD, inefficiency factors (Ineff), covariances and correlations (in italics) are shown for the parameters using the Eurodollar short-rate data applied to the CIR process. The MH algorithm was run with* $M = 5$, *for* $N = 4000$ *iterations with* 1000 *burn-in. The bandwidth used in the Parzen computations,* $B_N = 100$ *in all computations.*

	Mean	MCse	HPD		Ineff
$M = 5$					
$\alpha_0 \vert y$	0.00043	5.9e-08	0.00023	0.00063	2.13
$\alpha_1 \vert y$	-0.00523	9.43e-07	-0.00840	-0.00289	2.68
$\beta_0 \vert y$	1.3589e-09	2.16e-12	6.8321e-12	5.7623e-09	4.09
$\beta_1 \vert y$	0.00015	8.55e-10	0.00014	0.00015	1.10

Covariance & *Correlation*			
1.2270e-08	*-0.91925*	*-0.027523*	*-0.030340*
-1.4322e-07	1.9781e-06	*0.020189*	*-0.016086*
-6.4413e-15	5.9994e-14	4.4639e-18	*-0.10886*
-1.0423e-11	-7.0164e-11	-7.1333e-16	9.6181e-12

Table 7.9: *The posterior means, standard errors, HPD, inefficiency factors (Ineff), covariances and correlations (in italics) are shown for the parameters using the Eurodollar short-rate data applied to the Affine CIR process. The MH algorithm was run with* $M = 5$, *for* $N = 4000$ *iterations with* 1000 *burn-in. The bandwidth used in the Parzen computations,* $B_N = 100$ *in all computations.*

where $y_t^{+(j)}$ and $y_t^{-(j)}$ are drawn from $r(y_t^- | y_t^+, y_{t+1}; M_0)\ f(y_t^+ | y_t, y_{t+1}; M)$. The quantity $f(y_t^+ | y_t, y_{t+1}; M)$ can be obtained from the MCMC output and

$$r(y_t^- | y_t^+, y_{t+1}; M_0) = \prod_{j=1}^{M_0} f(y_{t,j}^- | y_{t,j}^+; M_0).$$

7.6 Conclusion

Various parametric models have been considered for the 7-day Eurodollar deposit spot-rate. Diagnostic tests show that the correlation and clustering structures found in the data have not been captured. Volatility clustering in the residuals suggest that these models should be extended to incorporate an underlying model for the volatility, such as a heavy tailed latent stochastic volatility component, which will be considered in Part III of the thesis.

The next Chapter considers existing methods and presents a new method for obtaining an expression for the log-likelihood within a stochastic differential equation system.

	Mean	MCse	HPD		Ineff
$M = 5$					
$\alpha_0 \mid y$	-0.015828	7.94e-07	-0.01979	-0.01243	1.64
$\alpha_1 \mid y$	0.16917	9.00e-06	0.11958	0.21473	1.45
$\alpha_2 \mid y$	-0.56317	3.71e-05	-0.72858	-0.39041	1.61
$\alpha_3 \mid y$	4.34e-04	2.01e-08	3.5917e-04	5.1960e-04	1.88
$\beta_0 \mid y$	4.5079e-05	1.53e-09	4.3157e-05	4.6886e-05	6.05
$\beta_1 \mid y$	-1.7964e-18	1.15e-19	-9.9997e-16	8.9796e-16	0.79
$\beta_2 \mid y$	1.0451e-04	9.44e-09	4.2038e-05	1.7449e-04	1.08

Covariance & *Correlation*						
3.7461e-06	*-0.97713*	*0.92326*	*-0.96587*	*-0.05307*	*0.00000*	*0.06589*
-4.7003e-05	0.00062	*-0.98090*	*0.89815*	*0.02589*	*0.00000*	*-0.06525*
1.6504e-04	-0.00225	0.00853	*-0.81951*	*-0.01996*	*0.00000*	*0.06275*
-8.0094e-08	9.5636e-07	-3.2427e-06	1.8356e-09	*0.09741*	*0.00000*	*-0.05800*
-1.0369e-10	6.4968e-10	-1.8612e-09	4.2132e-12	1.0192e-12	*0.00000*	*-0.36379*
-8.1442e-20	1.1474e-18	-4.2214e-18	1.3745e-21	-2.2426e-23	3.4021e-31	*0.00000*
4.4483e-09	-5.6565e-08	2.0217e-07	-8.6674e-11	-1.2811e-11	-1.6921e-22	1.2168e-09

Table 7.10: *The posterior means, standard errors, HPD, inefficiency factors (Ineff), covariances and correlations (in italics) are shown for the parameters, (fixing $\beta_3 = 0$), using the Eurodollar short-rate data applied to the general parametric process. The MH algorithm was run with $M = 5$, for $N = 4000$ iterations with 1000 burn-in. The bandwidth used in the Parzen computations, $B_N = 100$ in all computations.*

Model	$\log m(x)$	**Numerical se**
Vasicek	8328.9	0.52857
CIR	9430.5	0.25093
Affine CIR	9436.9	0.26587
General parametric model	9716.8	0.60919
Model	$\log m(x)$	**Numerical se**
Vasicek	8339.3	0.19086
CIR	9434.9	0.03003
Affine CIR	9444.2	0.03166
General parametric model	7323.8	0.03630

Table 7.11: *The marginal likelihood computations are shown for $x = \log y$. $f(\theta|x)$ is computed using using $M = 10$, $\lambda = 3$ burn-in=200 and $N = 2000$ in the top panel; general parametric model does not restrict β_3 to be zero. Bottom panel uses $M = 5$ updated in one block and burn-in=1000, $N = 4000$; general parametric model restricts $\beta_3 = 0$. Priors used are Inverse Gamma $\left\{\frac{\theta^*}{(15\sigma^*)^2} + 2, \theta^*(\sigma^* - 1)\right\}$ for β, and Normal $\{\theta^* + 4\sigma^*, (15\sigma^*)^2\}$ for α, (β_1 uses normal specification in general parametric model), where θ^* and σ^* are the respective posterior means and standard deviations, based on a training sample data set. Numerical standard error of $f(x|\theta^*)$ (log scale) is denoted by (numerical se). Both use the importance sampling method to compute $f(x|\theta^*)$ where all M are updated in one sweep.*

CHAPTER 8

ESTIMATING THE LIKELIHOOD FOR STOCHASTIC DIFFERENTIAL EQUATIONS

This Chapter outlines the existing methods for obtaining an expression for the log-likelihood within a stochastic differential equation system. A new method is presented and compared to existing methods. Obtaining an estimate of the likelihood may be important for certain applications, useful for inference and necessary from a classical perspective.

Section 8.1 will first introduce the key concepts. The aim will be to generate random vectors from a multivariate distribution with a density function which is known up to a constant of proportionality and to evaluate it at a certain point. These integrals can often be computationally intractable when the integrals cannot be evaluated explicitly. Multivariate proposal distributions are covered in Section 8.2. Existing methods for obtaining an expression for the log-likelihood are then first outlined and a new method is presented. The existing methods are based on simulated likelihood methods and have been used in numerous approaches, including filtering techniques and methods based on estimating functions. Pedersen's method, (Pedersen 1995b, outlined in Section 8.3) uses the one-step ahead conditional transition density implied by the Euler scheme, and is shown to be inefficient as it induces high Monte Carlo variance. An alternative method, presented in Section 8.4, is based on a simulated likelihood ratio statistic, see Geyer (1999). The new method discussed in Section 8.5 is based on the concept of importance sampling for obtaining the likelihood. Using simulated data, a comparison between the three approaches is presented in Section 8.6; the importance sampling approach is shown to dominate the other methods in finding the MLE. The Chapter concludes with Section 8.7.

8.1 MONTE CARLO TECHNIQUES FOR ESTIMATING INTEGRALS

We wish to approximate the integral

$$J(y) = \int f(y|x)h(x)dx \tag{8.1}$$

where $h(x)$ and $f(y|x)$ are known. If $J(y)$ cannot be evaluated analytically, one possibility is to use a numerical integration method, for examples, see Smith, Skene, Shaw, Naylor, and Dransfield (1985) and Crouch and Spiegelman (1990). However, numerical integration performs better for small-dimensional

integrals as opposed to high-dimensional ones. Alternatively, one can approximate $J(y)$ using a simulation technique known as *Monte Carlo Integration*. For more details, see Kloek and van Dijk (1978), van Dijk and Kloek (1980), Ripley (1987) and Geweke (1989).

Given that h is a probability density in (8.1), one can sample $x_1, \ldots, x_N$ independently from h and form

$$\hat{J}(y) = \frac{1}{N} \sum_{i=1}^{N} f(y|x_i).$$

The resulting $\hat{J}(y)$ unbiasedly estimates $J(y)$, for

$$\mathrm{E}_h\{\hat{J}(y)\} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{E}\{f(y|x_i)\} = \mathrm{E}\{f(y|x_1)\} = \int hf dx = J(y).$$

However, when the second moment exists,

$$\mathrm{var}\{\hat{J}(y)\} = \frac{1}{N} \int \{f(y|x) - J(y)\}^2 h(x) dx,$$

so that the standard error (as we are interested in the precision of the estimate) of $\hat{J}(y)$ is proportional to $1/\sqrt{N}$.

8.1.1 Importance sampling

We will first concentrate on the general integral to introduce the concept of importance sampling. For a more detailed exposition, see Tanner, (1996, Section 3.3.3), Gilks, Richardson, and Spiegelhalter, (1996, Section 6.4.1) and Tanner and Wong (1987). Suppose we are interested in (8.1), but are unable to sample explicitly from $h(x)$. Suppose also that $I(x)$ is a density that can be evaluated, is easy to sample from and approximates $h(x)$. Then the method of importance sampling approximates $J(y)$, where $w_i = h(x_i)/I(x_i)$, as

1. Draw $x_1, \ldots, x_N \overset{\text{i.i.d.}}{\sim} I(x)$,
2. $\tilde{J}(y) = \sum_{i=1}^{N} w_i f(y|x_i) / \sum_{i=1}^{N} w_i$,

giving more weight to regions where $I(x) < h(x)$ and less weight to regions where $I(x) > h(x)$. Geweke (1989) points out that if the support of $I(x)$ includes the support of $h(x)$, the x_i's are an i.i.d. sample from $I(x)$ and $J(y)$ exists and is finite, then the strong law of large numbers implies

$$\tilde{J}(y) \xrightarrow{\text{a.s.}} J(y).$$

The rate of convergence depends on how closely $I(x)$ approximates $h(x)$. Geweke (1989) also showed that the Monte Carlo standard error of $\tilde{J}(y)$ can be estimated by

$$\sqrt{\frac{\sum_{i=1}^{N} \{f(y|x_i) - \tilde{J}(y)\}^2 w_i}{\sum_{i=1}^{N} w_i}}.$$

$I(x)$ should be chosen so that it is a good approximation to $h(x)$. The tails of $I(x)$ should not decay faster than those of $h(x)$. Typically $I(x)$ is chosen so that $h(x)/I(x)\, f(y|x)$ has a small variance. The *multivariate-Student t* distribution is often used as a basis for the importance functions, especially if the distribution is heavy tailed.

8.2 MULTIVARIATE PROPOSAL DENSITIES

The method described above utilises random variates drawn from multivariate distributions. These can be obtained using random variables drawn from common univariate distributions. Two examples include the multivariate-normal and multivariate$-t$ distribution.

8.2.1 Sampling from a multivariate-normal distribution

Suppose we denote by $\boldsymbol{\mu}$ a $m \times 1$ vector and by $\boldsymbol{\Sigma}$, a $m \times m$ symmetric and positive definite matrix. Let

$$\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}'$$

correspond to the Choleski decomposition of $\boldsymbol{\Sigma}$, where $\boldsymbol{C}$ represents a lower triangular matrix. The properties of $\boldsymbol{C}$ are such that it is unique and has positive diagonal elements, see Lütkepohl (1996). The algorithm for sampling from $y \sim N(\mu, \Sigma)$ then proceeds as follows

1. Generate m independent standard normal variates $Q_1, \ldots, Q_m$ so that their joint distribution is $\boldsymbol{Q} \sim N_m(\boldsymbol{0}, \boldsymbol{I}_m)$.
2. Compute $\boldsymbol{V} = \boldsymbol{\mu} + \boldsymbol{C}\boldsymbol{Q}$, so that $\boldsymbol{V} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

8.2.2 Sampling from a multivariate$-t$ distribution

As the case of the multivariate-normal, the multivariate$-t$ distribution can be obtained from univariate$-t$ distributions. Suppose we let $\boldsymbol{V}$ denote a m dimensional$-t$ distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with ν degrees of freedom. That is, $\boldsymbol{V} \sim t_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$. Then the density function is

$$p(\boldsymbol{V}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+m}{2})}{(\nu\pi)^{m/2}\Gamma(\frac{\nu}{2})|\Sigma|^{1/2}}\{1 + \nu^{-1}(\boldsymbol{V} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{V} - \boldsymbol{\mu})\}^{-(\nu+m)/2} \tag{8.2}$$

with mean and variance of $\boldsymbol{V}$ given by $\boldsymbol{\mu}$ and $\nu(\nu-2)^{-1}\boldsymbol{\Sigma}$ respectively.

Variates from a multivariate distribution can then be generated using, for example, Anderson, (1984, Problem 7.27), as

1. Generate $\boldsymbol{Q}$ and P independently from $N_m(\boldsymbol{0}, \boldsymbol{\Sigma})$ and χ^2_ν distributions
2. Compute $\sqrt{\nu/P}\boldsymbol{Q} + \boldsymbol{\mu}$, which is distributed as $t_m(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$.

8.3 PEDERSEN'S METHOD

By definition, the approximate likelihood can be expressed in terms of the transition densities

$$\begin{aligned} f^M(y_{t+1}|y_t,\theta) &= \int f(y_{t+1}|y_t^*,y_t,\theta)f(y_t^*|y_t,\theta)dy_t^* \\ &= \int f(y_{t+1}|y_{t,M_t}^*;\theta)\, f(y_{t,M_t}^*|y_t)dy_{t,M_t}^* \quad \text{see (8.1)} \\ &= \mathrm{E}\{f(y_{t+1}|y_{t,M_t}^*;\theta)|y_t\}, \end{aligned} \tag{8.3}$$

where the expectations is taken with respect to the density

$$\begin{aligned} f(y_{t,M_t}^*|y_t) &= \int f(y_t^*|y_t)\, dy_{t,1}^*,\ldots,dy_{t,M_t-1}^* \\ &= \int f(y_{t,M_t}^*|y_{t,M_t-1}^*)f(y_{t,M_t-1}^*|y_{t,M_t-2}^*)\cdots f(y_{t,1}^*|y_t)\, dy_{t,1}^*,\ldots,dy_{t,M_t-1}^*, \end{aligned} \tag{8.4}$$

where dependence on θ has been suppressed. Monte Carlo integration is used to estimate (8.3) and (8.4) is used to generate the values of y_{t,M_t}^*, where we set $M_t = M$ for all t. As each $y_{t,j}^*$ is drawn from $f(y_{t,j}^*|y_{t,j-1}^*;\theta)$, the resulting $y_{t,M}^*$ will be a variate drawn from $f(y_{t,M}^*|y_t;\theta)$. A sequence of N by M standard normal random variates, $U_i^{(k)}$ are generated for $k = 1,\ldots,N$ and $i = 1,\ldots,M$; to reduce storage requirements, the same $U_i^{(k)}$ are used for all time points t. Keeping the same seed for all iterations has the advantage that the method is computationally less intensive. If different seeds are used, then the estimate of θ will not converge to a single point, (see Wang and Robins 1998 for a discussion in the context of the E-M algorithm). Each point $y_{t,i+1}^{(k)*}$ is simulated forwards where $y_{t,i+1}^{(k)*}|y_{t,i}^{(k)*}$ is normal given the Euler approximation

$$y_{t,i+1}^{*(k)} = y_{t,i}^{*(k)} + a\{y_{t,i}^{*(k)};\theta\}\Delta + b\{y_{t,i}^{*(k)};\theta\}\sqrt{\Delta}U_{i+1}^{(k)} \tag{8.5}$$

for $i = 0,\ldots,M-1$. An estimate of $f^M(y_{t+1}|y_t;\theta)$ is then given by

$$\begin{aligned} \hat{f}^M(y_{t+1}|y_t;\theta) &= \frac{1}{N}\sum_{k=1}^{N} f\{y_{t+1}|y_{t,M}^{*(k)};\theta\} \\ &= \frac{1}{N}\sum_{k=1}^{N} \frac{1}{\sqrt{2\pi b^2\{y_{t,M}^{*(k)};\theta\}\Delta}} \exp\left(-\frac{[y_{t+1}-y_{t,M}^{*(k)}-a\{y_{t,M}^{*(k)};\theta\}\Delta]^2}{2\Delta b^2\{y_{t,M}^{*(k)};\theta\}}\right) \end{aligned} \tag{8.6}$$

and the estimated log-likelihood is given by

$$\hat{l}_T^{(M)}(\theta) = \sum_{t=1}^{T-1} \log \hat{f}^M(y_{t+1}|y_t;\theta). \tag{8.7}$$

Note that $\hat{f}^M$ is unbiased for f^M, but the log is biased.

An illuminating way to view the Pedersen approach is to consider his method as a particular importance sampler for $f^M(y_{t+1}|y_t;\theta)$. In general, if $I(y_t^*|y_t,y_{t+1};\theta)$ represents an importance sampler for this

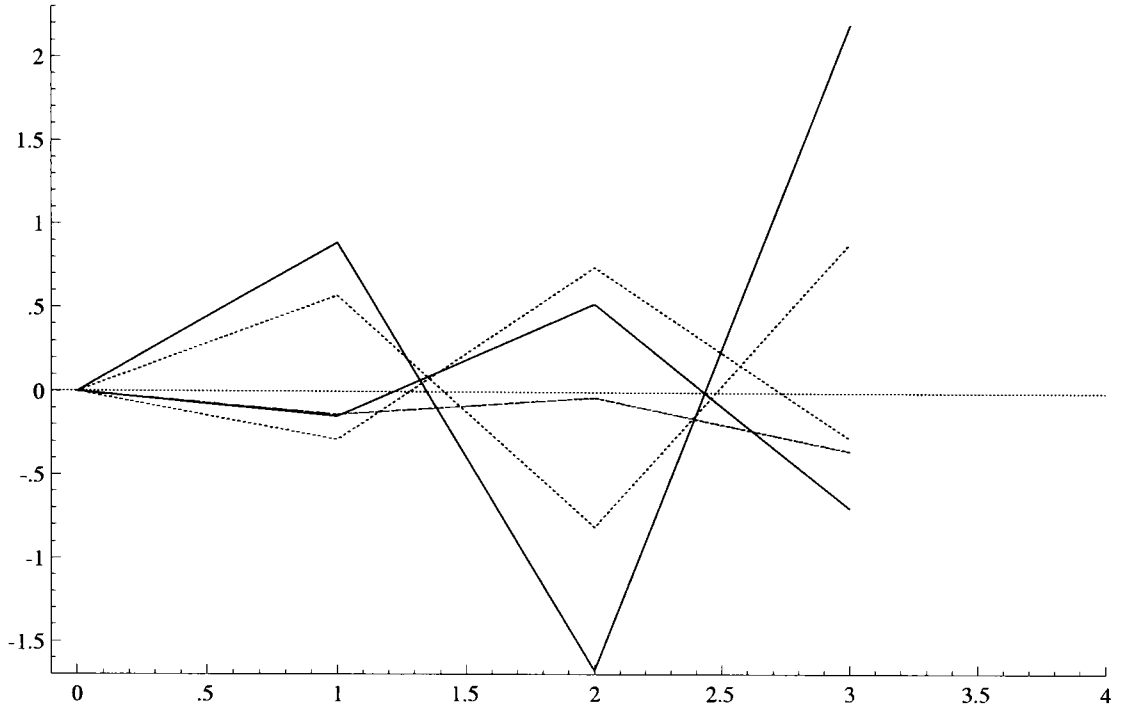


Figure 8.1: *Five simulated paths are shown for an OU process, with* $M = 3$ *and* $\mu = -0.5$ *and* $\sigma^2 = 0.01$. *Points are simulated one-step ahead with last value* $y^*_{t,M}$.

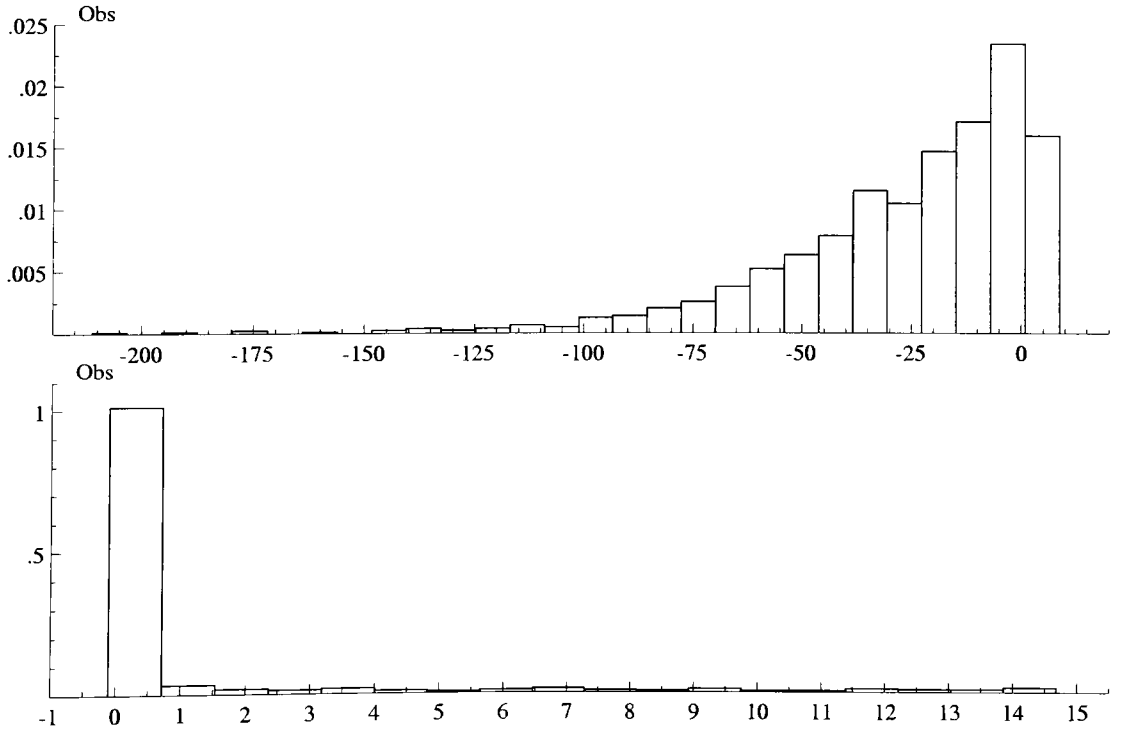


Figure 8.2: *Histogram is shown for the log density estimate (top) and density estimate (bottom) of* $f(y_{t+1}|y_t;\theta)$, *with* $M = 200$, *using* $N = 1000$ *paths. The mean in the top graph is given by* -26.625.

problem, we can write

$$f^M(y_{t+1}|y_t;\theta) = \int \frac{f(y_{t+1}, y_t^*|y_t;\theta)}{I(y_t^*|y_t, y_{t+1};\theta)} I(y_t^*|y_t, y_{t+1};\theta)\, dy_t^*,$$

where in the notation of Section 8.1.1, $f(y|x)\, h(x) = f(y_{t+1}, y_t^*|y_t;\theta)$. The best importance function (in terms of minimising the Monte Carlo variance of the estimate) is $f(y_t^*|y_t, y_{t+1};\theta)$ since then

$$\frac{f(y_{t+1}, y_t^*|y_t;\theta)}{f(y_t^*|y_t, y_{t+1};\theta)} = f^M(y_{t+1}|y_t;\theta)$$

and $f^M(y_{t+1}|y_t;\theta)$ can be computed using one simulated value of y_t^* . In practice, however, an explicit form for $f(y_t^*|y_t, y_{t+1};\theta)$ is not available and the optimal estimate is not realisable.

Considering the importance sampler for this problem, Pedersen approximates $f^M(y_{t+1}|y_t;\theta)$ using Monte Carlo Integration for

$$\int f(y_{t+1}|y_t^*, y_t;\theta) f(y_t^*|y_t;\theta) dy_t^* = \int \frac{f(y_{t+1}, y_t^*|y_t;\theta)}{f(y_t^*|y_t;\theta)} f(y_t^*|y_t;\theta) dy_t^*$$

and suggests that this transition density be estimated by importance sampling as

$$\frac{1}{J}\sum_{j=1}^{J} \frac{f\{y_{t+1}, y_t^{*(j)}|y_t;\theta\}}{f\{y_t^{*(j)}|y_t;\theta\}} = \frac{1}{J}\sum_{j=1}^{J} f\{y_{t+1}|y_{t,M}^{*(j)};\theta\}$$

where $y_t^{*(j)}$ is drawn from the importance function $f(y_t^*|y_{t;,\theta})$. This estimate, although easily available, is poorly behaved if M is large since the importance function conditions only on the first observation and knowledge of y_{t+1} is not being utilised. As M increases, the Monte Carlo bias may decrease but the Monte Carlo variance increases substantially.

This is illustrated in Figure 8.1, which shows simulated paths, $y_{t,i}^{*(k)}$ (for $k = 1, \ldots, 5$ and $i = 1, \ldots, 3$), starting at $y_t = 0$ and continuing to $y_{t,M}^{*(k)}$ for the OU process, simulated one-step ahead with last value $y_{t,M}^*$. Notice the range of values for $y_{t,M}^*$, resulting from only conditioning on y_t. The estimate $\hat{f}(y_{t+1}|y_{t,M}^*;\theta)$ is an average of $f\{y_{t+1}|y_{t,M}^{*(k)};\theta\}$. As M, the number of auxiliary points increases, Δ decreases resulting in a decrease in the variance of $f(y_{t+1}|y_{t,M}^*;\theta)$. Those nearest to y_{t+1} are given higher weight than those further away. If the densities are fairly peaked, only a small number of them will have a non-negligible contribution to $\hat{f}^M(y_{t+1}|y_t;\theta)$, the sample mean of $f\{y_{t+1}|y_{t,M}^{*(k)}\}$ over k. The result is an increase in the Monte Carlo variance of the estimate of $f(y_{t+1}|y_t;\theta)$, which can also be seen from Figure 8.2. The Histograms are shown for the log density and density estimates of $f(y_{t+1}|y_t;\theta)$, with $M = 200$, using $N = 1000$ paths. Notice how most of the contribution is provided by a small number of the simulated paths, (given that the mean is given by -26.625). In this example, 612 out of 1000 of the $f(y_{t+1}|y_{t,M}^{*(k)})$ are less than 10^{-6}, while only 160 of them are more than 1. The estimate of $\hat{f}$ is thus overwhelmingly dominated by the contributions from a small number of paths where $y_{t,M}^{(k)*}$ is close to y_{t+1}. An additional graphical representation is provided in Figure 8.3.

Estimates of the θ are obtained by maximising the approximate log-likelihood with respect to θ. Suppose we consider the OU process as an example. The functions for the OU process are given by $a(y) = \mu y$

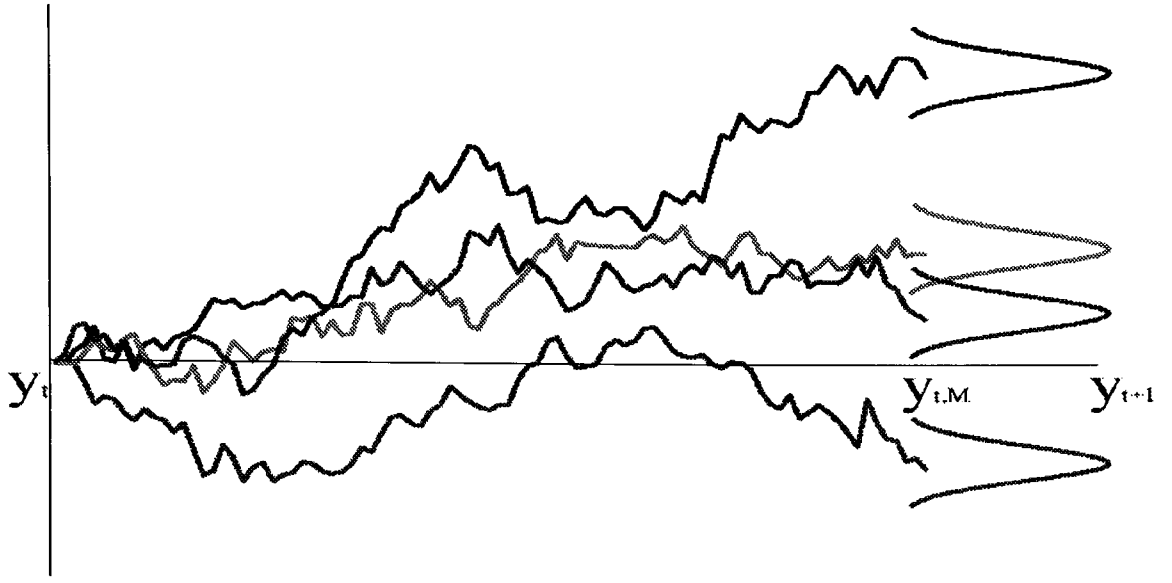


Figure 8.3: *Figure shows simulated paths for $f(y^*_{t,M}|y_t;\theta)$ using Pedersen's method. Initially, as M, the number of auxiliary points increases, Δ decreases resulting in a decrease in the variance of $f(y_{t+1}|y^*_{t,M};\theta)$. Those nearest to y_{t+1} are given higher weight than those further away. If the densities are fairly peaked, only a small number of them will contribute to $\hat{f}^M(y_{t+1}|y_t;\theta)$, the sample mean of $f\{y_{t+1}|y^{*(k)}_{t,M}\}$.*

and $b(y) = \sigma$, so if we choose $M = 1$, equations (8.5) and (8.6) become

$$\begin{aligned} y^{*(k)}_{t,1} &= y_t(1+\mu\Delta) + \sigma\sqrt{\Delta}U^{(k)}_{t,1} \\ \hat{f}^1(y_{t+1}|y_t) &= \frac{1}{N}\sum_{k=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2\Delta}}\exp\left[-\frac{1}{2\sigma^2\Delta}\{y_{t+1} - y^{*(k)}_{t,1}(1+\mu\Delta)\}^2\right] \\ &= \frac{1}{N}\sum_{k=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2\Delta}}\exp\left[-\frac{1}{2\sigma^2\Delta}\{y_{t+1} - y_t(1+\mu\Delta)^2 - (1+\mu\Delta)\sigma\sqrt{\Delta}U^{(k)}_1\}^2\right]. \end{aligned}$$

In general,

$$\hat{f}^M(y_{t+1}|y_t;\theta) = \frac{1}{N}\sum_{k=1}^{N}\frac{1}{\sqrt{2\pi\sigma^2\Delta}}\exp\left[-\frac{\{y_{t+1} - y_t(1+\mu\Delta)^{M+1}}{2\sigma^2\Delta} - \sigma\sqrt{\Delta}\sum_{i=1}^{M}(1+\mu\Delta)^{M+1-i}U^{(k)}_i\}^2\right].$$

Due to the form of $a(y)$, a fairly simple closed form expression can be obtained for $\hat{f}^M(y_{t+1}|y_t;\theta)$. However, closed-form estimates for μ and σ^2 are not available, essentially due to the fact that $\hat{f}$ is of the form $\log(N^{-1}\sum_{k=1}^{N} f_{t,k})$, that is, the logarithm of a sum. So even in this simple case, the approximate log-likelihood $\hat{l}^M(\theta)$ needs to be maximised numerically. For more complicated models, closed-form expressions of $y^{*(k)}_{t,M}$ are also inconvenient to work with as they involve a nested sequence of evaluations of $a(\cdot)$ and $b(\cdot)$. Instead, $y^{*(k)}_{t,M}$ and its derivatives will be evaluated recursively as shown in the Appendix 8.8.1. In practice, it is easier to work with numerical derivatives.

For clarity, the convention for notation below will be the following: define $w^k_{t,i} = y^{*(k)}_{t,i}$ and $w^k_t = w^k_{t,M}$. Vectors will be in bold while scalars will be in normal font. The partial derivative of f with respect to θ will be $\frac{\partial f}{\partial\theta}$. The functions will be defined as $a = a(\theta, w)$, $b = b(\theta, w)$, $w = w(\theta)$ and $f_{t,k} =$

$f(y_{t+1}|w_t^k) = f_{t,k}(a,b,w)$. To obtain the approximate MLE's of θ, we will need to iterate the following Newton-Raphson procedure for each iteration t

$$\theta^{(j+1)} = \theta^{(j)} - \left[\frac{\partial^2 \hat{l}_T^{(M)}\{\theta^{(j)}\}}{\partial\theta\partial\theta'}\right]^{-1}\left[\frac{\partial \hat{l}_T^{(M)}\{\theta^{(j)}\}}{\partial\theta}\right]. \tag{8.8}$$

It is important here to use the same sequence of random numbers $U_i^{(k)}$ for each iteration, since otherwise the algorithm will not converge to a fixed point. Drawing a different set of $U_i^{(k)}$ would produce a different estimate of θ. In effect, $U_i^{(k)}$ then becomes part of the data.

Suppose we consider two examples. The first process to be considered is the OU. Data is generated using the Euler approximation, with $R = 100{,}000$, where $\Delta = 1/R$. $T = 500$ observations are generated using $\mu = -0.5$, and $\sigma^2 = 0.01$. Figure 8.4 shows estimates of μ and σ^2 for $M = 20$, $N = 1000$.

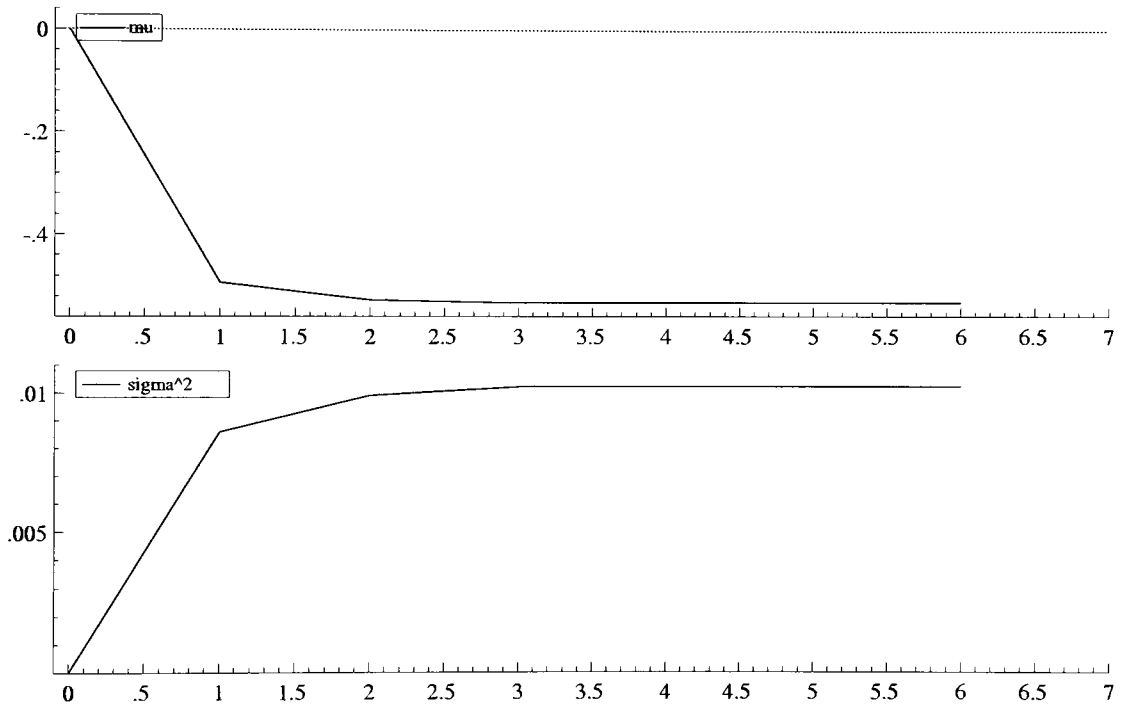


Figure 8.4: *The estimates of μ and σ^2 are shown, with $M = 20$, $N = 1000$. True values are $(-0.5, 0.01)'$. Iterations are shown on the x-axis.*

N must increase at least as fast as M, otherwise the Monte Carlo variance affects $\hat{f}^M$, distorting parameter values. It should also be noted that for some large values of M, given the same sequence $\{U_i^{(k)}\}_{i=1}^M$, $(k = 1,\ldots,N)$ is used for all t, few terms contributing to the value of $\hat{f}^M$ resulted in convergence problems.

Estimates corresponding to the approximate log-likelihoods in Figure 8.5				
M	1	2	10	25
$\hat{\theta}_T^M$	2.2	2.6	4.2	4.8

Figure 8.5 shows the approximate log-likelihoods for the one-dimensional SDE defined by

$$dy_t = -\theta y_t dt + \theta\sqrt{1 + \frac{y_t^2}{1+y_t^2}}dW_t \tag{8.9}$$

for the cases $M = 1, 2, 10$, and 25. The true value of θ is 5. Data was generated using $R = 100,000$ and $\Delta = 0.1/R$ for $T = 1000$. Corresponding estimates of θ for the approximate log-likelihood functions in Figure 8.5 are given in Table 8.3.

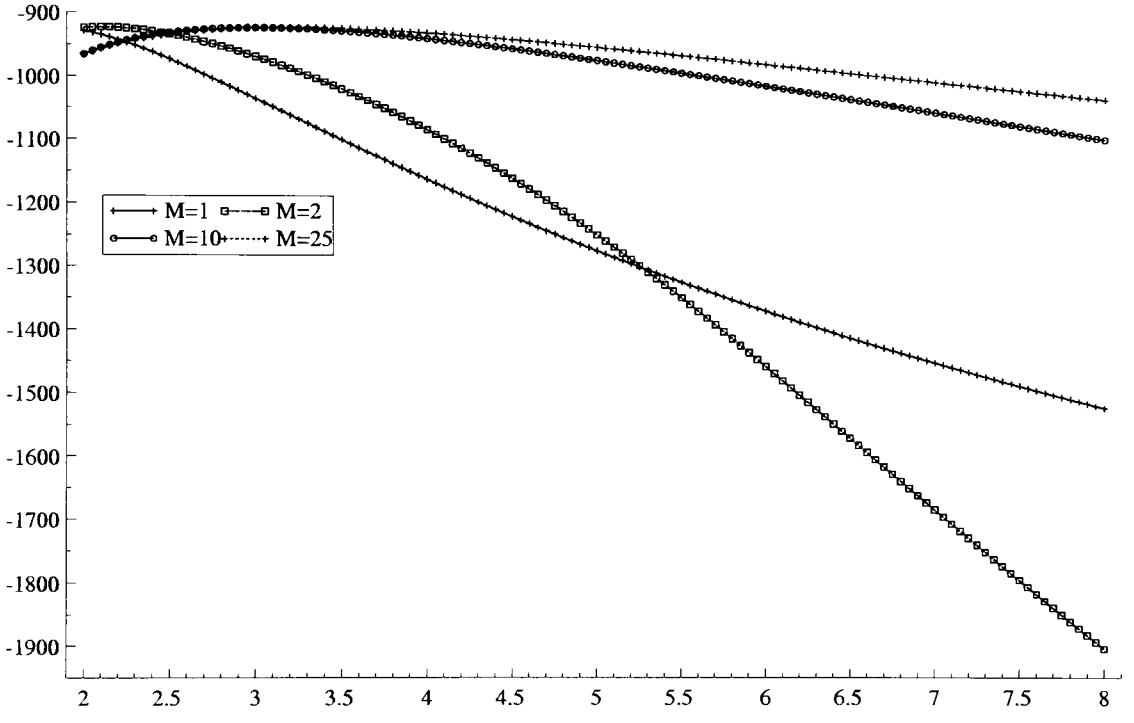


Figure 8.5: *The approximate log-likelihoods for the one-dimensional SDE given in (8.9) are shown for* $M = 1, 2, 10, 25$, *graphing the estimates against* θ. *The true value is* $\theta = 5$

8.4 Geyer's approach

Another approach, following Geyer (1999), is to work with the likelihood

$$f(y_{t+1}|y_t;\theta) = \int f(y_{t+1}, y_t^*|y_t;\theta)dy_t^*$$

which was outlined in Section 2.5.5. From the importance sampling interpretation above, we know that we are unable to explicitly evaluate the density we are interested in, $f(y_t^*|y_t, y_{t+1};\theta_0)$, but we are able to work with the likelihood ratio. The likelihood can be expressed as a ratio against a fixed parameter θ_0, so that

$$L(\theta) = \frac{f(y_{t+1}|y_t,\theta)}{f(y_{t+1}|y_t,\theta_0)} = \int \frac{f(y_{t+1}, y_t^*|y_t,\theta)}{f(y_{t+1}, y_t^*|y_t,\theta_0)} f(y_t^*|y_t, y_{t+1},\theta_0)dy_t^*$$

since $f(y_{t+1}|y_t;\theta_0) = \frac{f(y_{t+1},y_t^*|y_t;\theta_0)}{f(y_t^*|y_t,y_{t+1};\theta_0)}$.

The idea is to simulate from the smoothing density $f(y_t^*|y_t, y_{t+1};\theta)$ using MCMC methods. To approximate $L(\theta)$ by Monte Carlo, samples from the joint distribution of the observations and the auxiliary points, and the conditional distribution of the auxiliary points given the available observations (for the parameter θ_0 in both cases) need to be generated. Hence

$$L(\theta) = \mathrm{E}_{y_t^*}\left\{\frac{f(y_{t+1},y_t^*|y_t;\theta)}{f(y_{t+1},y_t^*|y_t;\theta_0)}|Y\right\} \tag{8.10}$$

can be estimated unbiasedly as

$$L_N(\theta) = \frac{1}{N}\sum_{j=1}^{N}\frac{f\{y_{t+1},y_t^{*(j)}|y_t;\theta\}}{f\{y_{t+1},y_t^{*(j)}|y_t;\theta_0\}}, \quad \text{where} \quad y_t^{*(j)} \sim f(y_t^*|y_t,y_{t+1};\theta_0). \tag{8.11}$$

Expression (8.11), which is smooth with respect to θ (since simulation is only considered over one value of θ_0), hence provides a Monte Carlo approximation of (8.10). MCMC sampling at θ_0 generates R random samples from the marginal distribution of y_t^* given y_t, y_{t+1} and θ_0, enabling us to estimate the likelihood ratio by taking the arithmetic mean of the sampled likelihood contributions over a grid of θ values. It also provides a classical interest in the density $\log f(y_t^*|y_t,y_{t+1};\theta)$ which is then summed over the T observations and maximised over θ yielding $\hat{\theta}_R$, the Monte Carlo approximation to the MLE $\hat{\theta}$ that maximises equation (8.10). Statistics such as the score and information matrix can easily be computed and the likelihood-ratio can be graphed against θ and used for diagnostic checks. The method, however, requires θ_0 to be close enough to the exact MLE $\hat{\theta}$. If convergence is slow, Geyer (1999) suggests repeating the optimisation procedure using different values of θ_0; that is, maximising the likelihood ratio to obtain $\hat{\theta}$ and using $\theta_0 = \hat{\theta}$ in the next optimisation run.

Though an efficient estimate of the likelihood function for θ in a neighbourhood of θ_0 can be obtained, evaluating the statistic is computationally intensive and samples may be highly correlated. Computations become increasingly unstable for values of θ some distance from θ_0 as values for $y_t^{*(j)}$ which are likely for θ far from θ_0 are unlikely to be sampled under θ_0. The method is also sensitive to initial values. Billio, Monfort, and Robert (1998) consider maximising (8.11) for a range of models. They state that provided R is large, the SLR method can be considered as asymptotically efficient, unlike the Method of Simulated Moments (MSM), the Simulated Pseudo Likelihood Method (SPLM) and the Indirect Inference method (II). It also only requires one optimisation (provided θ_0 is close to the exact MLE) which is not the case with simulated E-M methods.

8.5 Importance sampling approach

A more efficient importance function can be developed according to the approach of Chib, Greenberg, and Winkelmann (1998) and Shephard and Pitt (1997), which relies on a tailoring procedure that is analogous to that discussed above in connection with the M-H proposal density. As before, we are interested in the quantity

$$f^M(y_{t+1}|y_t;\theta) = \int \frac{f(y_{t+1},y_t^*|y_t;\theta)}{I(y_t^*|y_t,y_{t+1};\theta)} I(y_t^*|y_t,y_{t+1};\theta)\, dy_t^*$$

where $I(y_t^*|y_t, y_{t+1}; \theta)$ is the importance sampling density. This quantity is then approximated by

$$\hat{f}^M(y_{t+1}|y_t;\theta) = \frac{1}{R}\sum_{i=1}^{R} \frac{f\{y_{t+1}, y_t^{(j)*};\theta\}}{I\{y_t^{*(j)}|y_t, y_{t+1};\theta\}}. \tag{8.12}$$

Let $\hat{\mu}_t$ be the mode of $\log f(y_{t+1}, y_t^*|y_t;\theta)$ as a function of y_t^* and let $\hat{\Sigma}_t$ be the negative of the inverse of the Hessian of $\log f(y_{t+1}, y_t^*|y_t, \theta)$ evaluated at $\hat{\mu}$. Given $\hat{\mu}_t$ and $\hat{\Sigma}_t$, let the importance density be

$$q(y_t^*|y_t, y_{t+1};\theta) = f_T(y_t^*|\hat{\mu}_t, \hat{\Sigma}_t, \nu),$$

a multivariate-t density with ν degrees of freedom. Hence, y_t^* is sampled from a multivariate$-t$ distribution, where ν is chosen depending on the thickness required in the tails. This choice leads to the importance sampling density estimator of the form

$$\tilde{f}^M(y_{t+1}|y_t;\theta) = \frac{1}{R}\sum_{j=1}^{R} \frac{f\left(y_{t+1}, y_t^{*(j)}|y_t;\theta\right)}{f_T\left(y_t^{*(j)}|\hat{\mu}_t, \hat{\Sigma}_t, \nu\right)}, \tag{8.13}$$

$$y_t^{*(j)} \sim f_T(y_t^*|\hat{\mu}_t, \hat{\Sigma}_t, \nu). \tag{8.14}$$

The variance of $\tilde{f}^M$, when it exists, can be estimated by the method provided in Geweke (1989). Note that all M states can be integrated out by importance sampling in one sweep.

The Monte Carlo standard error (mcse) is given by

$$\text{mcse} = \sqrt{\sum_{t=1}^{T-1}\left(\frac{\text{mcse}_t}{\tilde{f}^M(y_{t+1}|y_t)}\right)}$$

where $\text{mcse}_t = \{\text{E}(\tilde{f}_t^M - \text{E}(\tilde{f}_t^M)\}^2/R$.

8.6 Comparisons

A comparison between Pedersen's method and Geyer's method is first presented in Tables 8.1 and 8.2. A comparison between Pedersen's method and the importance sampling approach is then provided in Figure 8.6, 8.7, 8.8 and Tables 8.3 and 8.4.

Ten Monte Carlo results for the OU process are shown in Table 8.1 for the parameter estimate μ and the simulated log-likelihoods under Pedersen's and Geyer's method. The true value of μ in the simulated datasets is -0.5. The average value of μ under Pedersen is given by -0.5695 and under Geyer -0.571. The average MSE is given by 0.0037575 and 0.005955 under Pedersen and Geyer respectively. There is much more fluctuation in the log-likelihood figures under Pedersen's method with an average of 594.884. Under Geyer, there is less discrepancy between estimates and the average is 597.945. Optimisation routines using both methods are then shown in Table 8.2 for different values of M and N. Estimates of μ and σ^2 and the resulting log-likelihood statistics are given for ($M = 3$, $N = 100$), ($M = 5$, $N = 100$), ($M = 5$, $N = 1000$) and ($M = 10, N = 100$). Both methods seem to perform equally well.

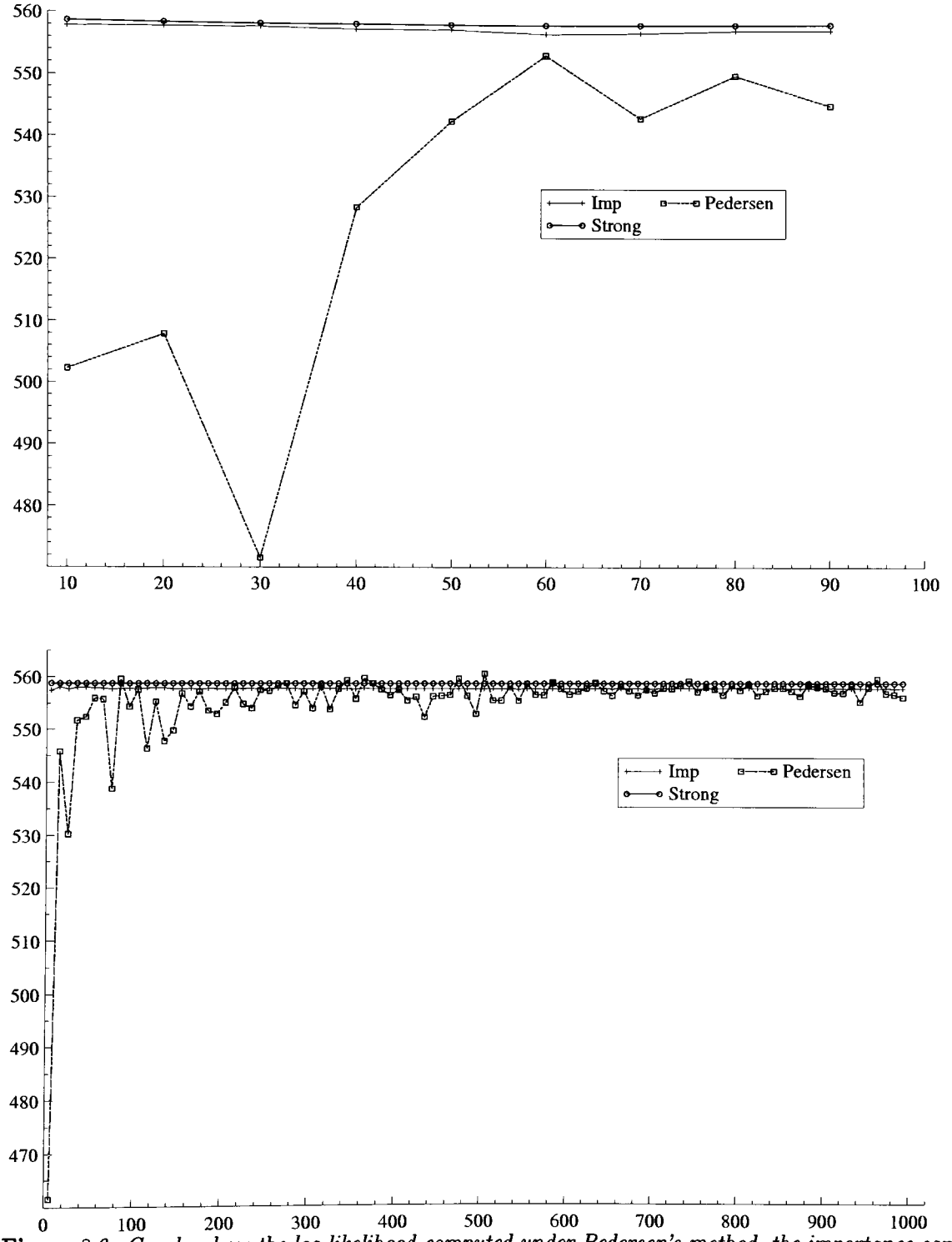


Figure 8.6: *Graphs show the log-likelihood computed under Pedersen's method, the importance sampling method and computed under the strong solution for the OU process. Top panel:* $M = 10$, $T = 500$, $\Delta^\dagger = 1$ *and* $N = 100$. *Bottom panel:* $M = 5$, $T = 500$, $\Delta^\dagger = 1$ *and* $N = 1000$.

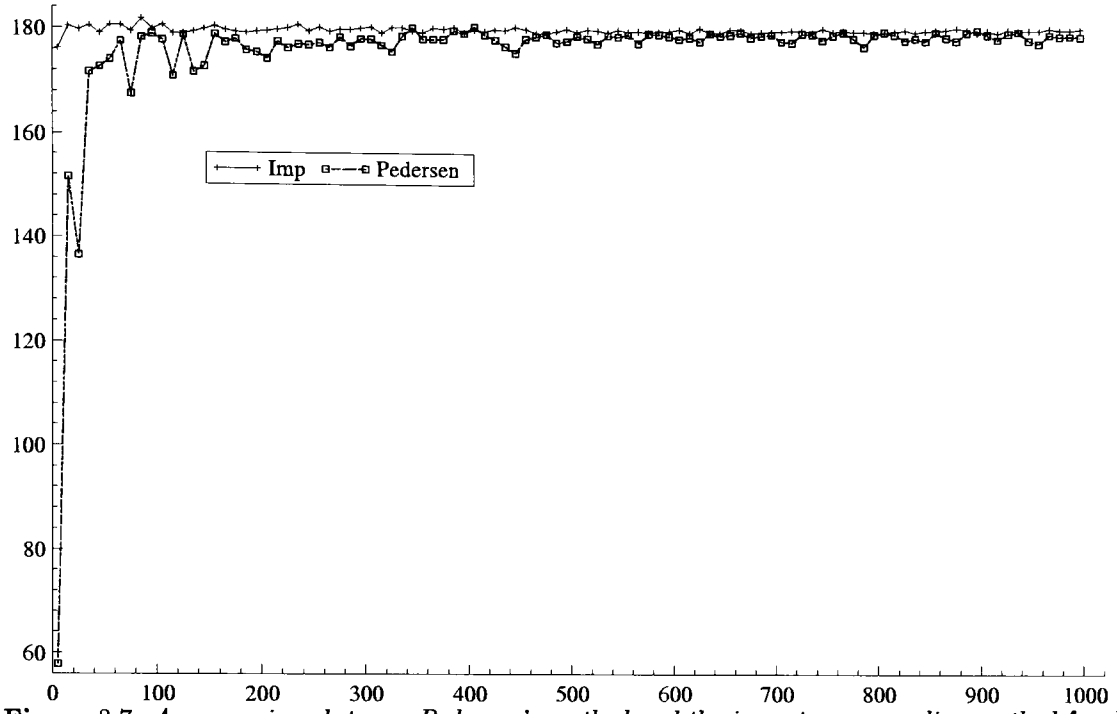


Figure 8.7: *A comparison between Pedersen's method and the importance sampling method for different values of R for the CIR process for $M = 5$, $T = 500$, $\Delta^{\dagger} = 2$ and $R = 5$ to $R = 1000$.*

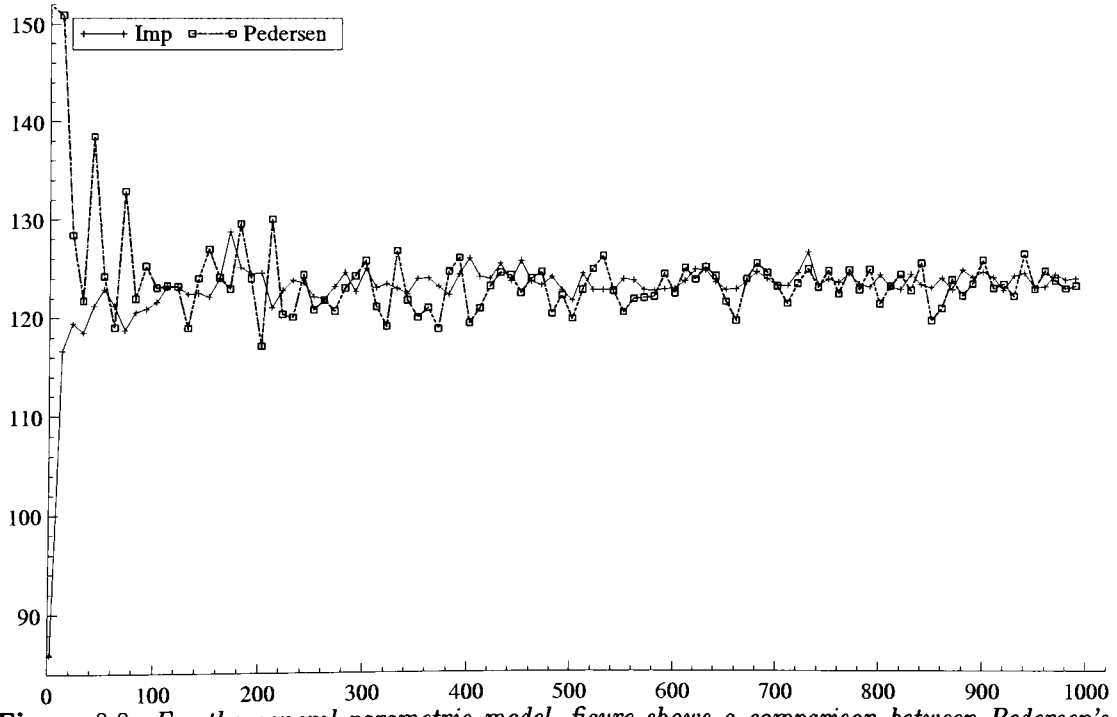


Figure 8.8: *For the general parametric model, figure shows a comparison between Pedersen's method and the importance sampling method for different values of R for the CIR process for $M = 2$, $T = 500$, $\Delta^{\dagger} = 2$ and $R = 2$ to $R = 1000$.*

Pedersen vs. Geyer, Ten Monte Carlos					
θ_P	θ_G	MSE in θ_P	MSE in θ_G	MSE in log-likelihood (P)	MSE in log-likelihood (G)
-0.54	-0.535	0.0016	0.001225	399.00	395.52
-0.59	-0.58	0.0081	0.0064	1251.1	643.65
-0.57	-0.575	0.0049	0.005625	872.80	629.05
-0.54	-0.55	0.0016	0.0025	486.71	482.56
-0.55	-0.56	0.0025	0.0036	503.81	511.61
-0.54	-0.545	0.0016	0.002025	588.45	459.77
-0.625	-0.635	0.015625	0.018225	2289.0	980.59
-0.62	-0.615	0.0144	0.013225	2209.3	856.90
-0.565	-0.565	0.004225	0.004225	841.31	546.05
-0.555	-0.55	0.003025	0.0025	507.36	472.90

Table 8.1: *The parameter estimate of μ, (true value -0.5) is shown for the OU, where Mean Square Error (MSE) in μ and the log-likelihood are computed under the Pedersen (P) and Geyer (G) methods.*

Pedersen numerical optimisation routine				
M	N	μ	σ^2	log-likelihood
3	100	-0.45772	0.0093774	540.469
5	100	-0.47551	0.0092892	540.930
5	1000	-0.45982	0.0098236	544.419
10	100	-0.51838	0.010745	539.286

Geyer numerical optimisation routine				
M	N	μ	σ^2	LR
3	100	-0.51304	0.0092746	0.52089
5	100	-0.47833	0.0091672	0.79747
5	1000	-0.49726	0.010174	0.64700
10	100	-0.47534	0.0090668	0.53490

Table 8.2: *Left: Pedersen; Right: Geyer; using an optimisation routine, true values for OU are $\mu = -0.5$ and $\sigma^2 = 0.01$, using one simulated dataset.*

A comparison between Pedersen's method, the strong solution and the importance sampling approach is provided by Figure 8.6 applied to the OU process. Top panel has $M = 10$, $T = 500$, $\Delta^\dagger = 1$ and $N = 100$ and $M = 5$, $N = 1000$ in the bottom panel. The importance sampler is shown to be superior to Pedersen's method for this problem, converging to the strong solution, already for $M = 10$. The estimate given by Pedersen is seen to fluctuate much more heavily. Figures for the log-likelihoods over different number of iterations under both sets of auxiliary points are shown in Tables 8.3 and 8.4. The strong solution is given by 558.71. Figures 8.7 and 8.8 extend the analysis to the general parametric model when $\Delta^\dagger = 2$ in the simulated data. The importance sampler is again shown to dominate Pedersen's method.

8.7 Conclusion

Existing methods for obtaining an expression for the log-likelihood within a stochastic differential equation system have been presented using three different methods. These include simulated likelihood methods, as in the case of Pedersen (1995b), a simulated likelihood ratio method based on the ideas of Geyer (1999) and an importance sampling method for obtaining the likelihood, which is shown to be superior to the other two methods in finding the MLE.

Part III next explores extensions to the work presented in Part II. Specifically, alternative schemes that can be used to approximate the transition density are considered. Extending the analysis to multivariate

A comparison in log-likelihoods for $M = 10$ and various values of R				
M	R	Pedersen	importance sampling	strong solution
10	10	502.26	557.88	558.71
10	20	507.93	558.09	558.71
10	30	471.54	558.20	558.71
10	40	528.80	557.88	558.71
10	50	542.96	557.93	558.71
10	60	553.79	557.30	558.71
10	70	543.58	557.43	558.71
10	80	550.47	557.77	558.71
10	90	545.54	557.82	558.71

A comparison in log-likelihoods for $M = 5$ and various values of R			
M	R	Pedersen	IS
5	05	461.52	557.40
5	75	538.78	557.63
5	115	546.35	557.80
5	185	552.90	557.71
5	285	554.59	557.76
5	355	555.78	557.69
5	445	556.29	557.77
5	565	556.63	557.76
5	655	556.41	557.71
5	755	557.05	557.70
5	875	557.80	557.81
5	995	556.03	557.73

Table 8.3: *A comparison between Pedersen's method and the importance sampling method for different values of R for the OU process. Left: $M = 10$, $T = 500$, $\Delta^\dagger = 1$ and various values of R. Right: $M = 5$, $T = 500$, $\Delta^\dagger = 1$ and various values of R. The strong solution in both cases gives 558.71.*

models is also discussed.

8.8 Appendix

8.8.1 Derivatives used in Pedersen's method

To compute the Newton-Raphson procedure given by equation 8.8, we will need the following:

$$\begin{aligned}
\hat{l}_T^{(M)}(\theta) &= \sum_{t=1}^{T-1} \log f_t \\
\frac{d\hat{l}_T^{(M)}(\boldsymbol{\theta})}{d\boldsymbol{\theta}} &= \sum_{t=1}^{T-1} \left(\frac{d\boldsymbol{f_t}}{d\boldsymbol{\theta}}\right) f_t^{-1} \\
\frac{d^2\hat{l}_T^{(M)}(\boldsymbol{\theta})}{d\boldsymbol{\theta}d\boldsymbol{\theta}'} &= \sum_{t=1}^{T-1} \left\{ f_t \frac{d^2 \boldsymbol{f_t}}{d\boldsymbol{\theta}d\boldsymbol{\theta}'} - \left(\frac{d\boldsymbol{f_t}}{d\boldsymbol{\theta}}\right)\left(\frac{d\boldsymbol{f_t}}{d\boldsymbol{\theta}}\right)' \right\} f_t^{-2}
\end{aligned}$$

where

$$\begin{aligned}
f_t &= \frac{1}{N}\sum_{k=1}^{N} f_{t,k} \\
\frac{d\boldsymbol{f_t}}{d\boldsymbol{\theta}} &= \frac{1}{N}\sum_{k=1}^{N} \frac{d\boldsymbol{f_{t,k}}}{d\boldsymbol{\theta}} \\
\frac{d^2\boldsymbol{f_t}}{d\boldsymbol{\theta}d\boldsymbol{\theta}'} &= \frac{1}{N}\sum_{k=1}^{N} \frac{d^2\boldsymbol{f_{t,k}}}{d\boldsymbol{\theta}d\boldsymbol{\theta}'}
\end{aligned}$$

A comparison in log-likelihoods for $M = 10$, for various values of R			
M	R	Pedersen	importance sampling
10	30	471.54	558.20
10	210	547.76	557.77
10	300	554.84	557.82
10	450	552.42	557.64
10	550	556.74	557.61
10	610	556.44	557.74
10	670	552.64	557.65
10	760	555.84	557.74
10	800	555.11	557.78
10	930	555.29	557.77
10	990	559.54	557.70

Table 8.4: *A comparison between Pedersen's method and the importance sampling method for different values of R for the OU process for $M = 10$, $T = 500$, $\Delta^\dagger = 1$ and $R = 10 - 1000$. The strong solution in both cases gives 558.71.*

and suppressing dependence on w_t^k in the functions $a(w_t^k)$ and $b(w_t^k)$,

$$f_{t,k} = f(y_{t+1}|w_t^k) = \frac{1}{\sqrt{2\pi\Delta b^2}} \exp\left\{-\frac{1}{2\Delta b^2}(y_{t+1} - a\Delta - w_t^k)^2\right\}. \tag{B.8.1}$$

We will also define

$$\begin{aligned} \frac{\boldsymbol{da}}{\boldsymbol{d\theta}} &= \frac{\boldsymbol{\partial a}}{\boldsymbol{\partial \theta}} + \frac{\partial a}{\partial w_t^k}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta}} \\ \frac{\boldsymbol{db}}{\boldsymbol{d\theta}} &= \frac{\boldsymbol{\partial b}}{\boldsymbol{\partial \theta}} + \frac{\partial b}{\partial w_t^k}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta}}. \end{aligned}$$

The values w_t^k and their derivatives are computed recursively using

$$\begin{aligned} w_{i+1} &= w_{t,i}^k + a\Delta + b\sqrt{\Delta}U_{i+1} \\ \frac{\boldsymbol{dw_{i+1}}}{\boldsymbol{d\theta}} &= \frac{\boldsymbol{dw_{t,i}^k}}{\boldsymbol{d\theta}} + \frac{\boldsymbol{da}}{\boldsymbol{d\theta}}\Delta + \frac{\boldsymbol{db}}{\boldsymbol{d\theta}}\sqrt{\Delta}U_{i+1} \\ \frac{\boldsymbol{d^2w_{i+1}}}{\boldsymbol{d\theta d\theta'}} &= \frac{\boldsymbol{d^2w_{t,i}^k}}{\boldsymbol{d\theta d\theta'}} + \frac{\boldsymbol{d^2a}}{\boldsymbol{d\theta d\theta'}}\Delta + \frac{\boldsymbol{d^2b}}{\boldsymbol{d\theta d\theta'}}\sqrt{\Delta}U_{i+1} \end{aligned}$$

where a and b are evaluated at $w_{t,i}^k$. Starting values take the form

$$\begin{aligned} w_1 &= y_t + a(y_t)\Delta + b(y_t)\sqrt{\Delta}U_1 \\ \frac{\boldsymbol{dw_1}}{\boldsymbol{d\theta}} &= \frac{\boldsymbol{\partial a(y_t)}}{\boldsymbol{\partial \theta}}\Delta + \frac{\boldsymbol{\partial b(y_t)}}{\boldsymbol{\partial \theta}}\sqrt{\Delta}U_1 \\ \frac{\boldsymbol{d^2w_1}}{\boldsymbol{d\theta d\theta'}} &= \frac{\boldsymbol{\partial^2 a(y_t)}}{\boldsymbol{\partial\theta\partial d\theta'}}\Delta + \frac{\boldsymbol{\partial^2 b(y_t)}}{\boldsymbol{\partial\theta\partial\theta'}}\sqrt{\Delta}U_1. \end{aligned}$$

Initial values of θ are found by maximising $\sum_{t=0}^{T-1} f(y_{t+1}|y_t)$. To obtain the total derivative of $f_{t,k}$ with respect to θ in equation (B.8.1), we use

$$\frac{\boldsymbol{df_{t,k}}}{\boldsymbol{d\theta}} = \frac{\partial f_{t,k}}{\partial a}\frac{\boldsymbol{da}}{\boldsymbol{d\theta}} + \frac{\partial f_{t,k}}{\partial b}\frac{\boldsymbol{db}}{\boldsymbol{d\theta}} + \frac{\partial f_{t,k}}{\partial w_t^k}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta}}$$

where

$$\begin{aligned}
\frac{\partial f_{t,k}}{\partial a} &= f_{t,k}\frac{(y_{t+1} - a\Delta - w_t^k)}{b^2} \\
\frac{\partial f_{t,k}}{\partial b} &= \frac{f_{t,k}}{\Delta b^3}\{(y_{t+1} - a\Delta - w_t^k)^2 - \Delta b^2\} \\
\frac{\partial f_{t,k}}{\partial w_t^k} &= f_{t,k}\frac{(y_{t+1} - a\Delta - w_t^k)}{\Delta b^2}.
\end{aligned}$$

Suppose we also define the partial second derivatives of $f_{t,k}$ with respect to a, b and w_t^k as

$$\begin{aligned}
\frac{\partial^2 f_{t,k}}{\partial a^2} &= \frac{f_{t,k}}{b^4}\{(y_{t+1} - a\Delta - w_t^k) - \Delta b^2\} \\
\frac{\partial^2 f_{t,k}}{\partial a \partial w_t^k} &= \frac{\partial^2 f_{t,k}}{\partial w_t^k \partial a} = \frac{f_{t,k}}{\Delta b^4}\{(y_{t+1} - a\Delta - w_t^k)^2 - \Delta b^2\} \\
\frac{\partial^2 f_{t,k}}{\partial a \partial b} &= \frac{\partial^2 f_{t,k}}{\partial b \partial a} = \frac{f_{t,k}}{\Delta b^5}(y_{t+1} - a\Delta - w_t^k)\{(y_{t+1} - a\Delta - w_t^k)^2 - 3\Delta b^2\} \\
\frac{\partial^2 f_{t,k}}{\partial b^2} &= \frac{f_{t,k}}{\Delta^2 b^6}[\{(y_{t+1} - a\Delta - w_t^k)^2 - \Delta b^2\}\{(y_{t+1} - a\Delta - w_t^k)^2 - 4\Delta b^2\} - 2\Delta^2 b^4] \\
\frac{\partial^2 f_{t,k}}{\partial b \partial w_t^k} &= \frac{\partial^2 f_{t,k}}{\partial w_t^k \partial b} = \frac{f_{t,k}}{\Delta^2 b^5}(y_{t+1} - a\Delta - w_t^k)\{(y_{t+1} - a\Delta - w_t^k)^2 - 3\Delta b^2\} \\
\frac{\partial^2 f_{t,k}}{(\partial w_t^k)^2} &= \frac{f_{t,k}}{\Delta^2 b^4}\{(y_{t+1} - a\Delta - w_t^k)^2 - \Delta b^2\}
\end{aligned}$$

Therefore, the total second derivative of $f_{t,k}$ with respect to θ can be expressed as

$$\begin{aligned}
\frac{\boldsymbol{d^2 f_{t,k}}}{\boldsymbol{d\theta d\theta'}} &= \frac{\boldsymbol{da}}{\boldsymbol{d\theta}}\left[\frac{\boldsymbol{d}}{\boldsymbol{d\theta'}}\left(\frac{\partial f_{t,k}}{\partial a}\right)\right] + \frac{\partial f_{t,k}}{\partial a}\frac{\boldsymbol{d^2 a}}{\boldsymbol{d\theta d\theta'}} + \frac{\boldsymbol{db}}{\boldsymbol{d\theta}}\left[\frac{\boldsymbol{d}}{\boldsymbol{d\theta'}}\left(\frac{\partial f_{t,k}}{\partial b}\right)\right] + \frac{\partial f_{t,k}}{\partial b}\frac{\boldsymbol{d^2 b}}{\boldsymbol{d\theta d\theta'}} \\
&\quad + \frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta}}\left[\frac{\boldsymbol{d}}{\boldsymbol{d\theta'}}\left(\frac{\partial f_{t,k}}{\partial w_t^k}\right)\right] + \frac{\partial f_{t,k}}{\partial w_t^k}\frac{\boldsymbol{d^2 w_t^k}}{\boldsymbol{d\theta d\theta'}}
\end{aligned} \tag{B.8.2}$$

where

$$
\begin{aligned}
\frac{\boldsymbol{d}}{\boldsymbol{d\theta'}}\left(\frac{\partial f_{t,k}}{\partial a}\right) &= \frac{\partial^2 f_{t,k}}{\partial a^2}\frac{\boldsymbol{da}}{\boldsymbol{d\theta'}} + \frac{\partial^2 f_{t,k}}{\partial a \partial b}\frac{\boldsymbol{db}}{\boldsymbol{d\theta'}} + \frac{\partial^2 f_{t,k}}{\partial a \partial w_t^k}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta'}} \\
\frac{\boldsymbol{d^2 a}}{\boldsymbol{d\theta d\theta'}} &= \frac{\boldsymbol{\partial^2 a}}{\boldsymbol{\partial\theta\partial\theta'}} + \frac{\boldsymbol{\partial^2 a}}{\boldsymbol{\partial\theta\partial w_t^k}}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta'}} + \frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta}}\frac{\boldsymbol{\partial^2 a}}{\boldsymbol{\partial w_t^k \partial\theta'}} + \frac{\partial^2 a}{(\partial w_t^k)^2}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta}}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta'}} + \frac{\partial a}{\partial w_t^k}\frac{\boldsymbol{d^2 w_t^k}}{\boldsymbol{d\theta d\theta'}} \\
\frac{\boldsymbol{d}}{\boldsymbol{d\theta'}}\left(\frac{\partial f_{t,k}}{\partial b}\right) &= \frac{\partial^2 f_{t,k}}{\partial a \partial b}\frac{\boldsymbol{da}}{\boldsymbol{d\theta'}} + \frac{\partial^2 f_{t,k}}{\partial b^2}\frac{\boldsymbol{db}}{\boldsymbol{d\theta'}} + \frac{\partial^2 f_{t,k}}{\partial b \partial w_t^k}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta'}} \\
\frac{\boldsymbol{d^2 b}}{\boldsymbol{d\theta d\theta'}} &= \frac{\boldsymbol{\partial^2 b}}{\boldsymbol{\partial\theta\partial\theta'}} + \frac{\boldsymbol{\partial^2 b}}{\boldsymbol{\partial\theta\partial w_t^k}}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta'}} + \frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta}}\frac{\boldsymbol{\partial^2 b}}{\boldsymbol{\partial w_t^k \partial\theta'}} + \frac{\partial^2 b}{(\partial w_t^k)^2}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta}}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta'}} + \frac{\partial b}{\partial w_t^k}\frac{\boldsymbol{d^2 w_t^k}}{\boldsymbol{d\theta d\theta'}} \\
\frac{\boldsymbol{d}}{\boldsymbol{d\theta}}\left(\frac{\partial f_{t,k}}{\partial w_t^k}\right) &= \frac{\partial^2 f_{t,k}}{\partial a \partial w_t^k}\frac{\boldsymbol{da}}{\boldsymbol{d\theta'}} + \frac{\partial^2 f_{t,k}}{\partial b \partial w_t^k}\frac{\boldsymbol{db}}{\boldsymbol{d\theta'}} + \frac{\partial^2 f_{t,k}}{(\partial w_t^k)^2}\frac{\boldsymbol{dw_t^k}}{\boldsymbol{d\theta'}}.
\end{aligned}
$$

Part III

Extensions

EXTENSIONS

9 The Milstein scheme and higher order approximations **170**

9.1 Introduction 170

9.2 The Milstein density 171

9.3 The likelihood based on the Milstein density 174

9.4 Bayesian approach to diffusions using Milstein densities 176

9.5 An example 178

9.6 Conclusion and extensions 179

9.7 Appendix 180

10 A continuous-time bivariate model with unobserved stochastic volatility **186**

10.1 A review of stochastic volatility models 186

10.2 Estimating a bivariate diffusion 191

10.3 Examples of continuous-time processes with latent volatility 197

10.4 Conclusion and further research 204

10.5 Appendix 204

11 Multivariate stochastic differential equations **208**

11.1 Models 208

11.2 Discretization and augmentation issues 211

11.3 MCMC inference 212

11.4 Conclusion 219

11.5 Appendix 219

12 Concluding remarks and further topics **221**

Part III considers the extensions to the material discussed in Part II. Chapter 9 considers the use of the Milstein and Taylor 1.5 approximations, as an alternative to the Euler scheme. The Milstein scheme can be shown to imply a closed-form expression for the one-step ahead conditional transition density. The method can be extended to consider a continuous-time bivariate model with unobserved stochastic volatility, which is the subject of Chapter 10. The theory for a general multivariate stochastic differential equation system is then discussed in Chapter 11. The thesis concludes with Chapter 12.

CHAPTER 9

THE MILSTEIN SCHEME AND HIGHER ORDER APPROXIMATIONS

In previous Chapters, we have considered the Euler approximation to obtain a discretised version of the SDE. Existing methods based on approximating the diffusion with a stochastic Taylor scheme use the Euler-Maruyama approximation in obtaining an expression for the one-step ahead transition density, due to its analytical tractability. The discretisation implied by the Euler scheme was given in (2.11). A normal transition density for $f(y_{t+1}|y_t;\theta)$ is obtained by taking the increments of the Wiener process to be normal, with mean zero and variance equal to the time step, $\Delta^\dagger$, as in (2.12). Higher order approximations can also be used to discretise the continuous-time process. The use of a more sophisticated approximation implies a lower discretisation error, but these are in general more difficult to implement; MCMC methods rely on being able to evaluate the one-step ahead conditional transition density implied by the scheme, which until now has only been possible using the Euler scheme.

We will consider some higher order stochastic Taylor schemes. Specifically, the Milstein and Taylor 1.5 schemes will be discussed and a closed form expression for the density implied by the Milstein scheme will be shown to exist. Section 9.2 will define the Milstein scheme, giving a closed form expression for the density. The order of convergence is discussed and compared to the Euler approximation. Section 9.3 presents the likelihood based on the Milstein density and examines the bias and inconsistency in the log-likelihood, inherent without the introduction of auxiliary points. Section 9.3.2 analyses use of the Milstein density using a simulated likelihood approach, as in the case of Pedersen (1995b). This is followed by an outline of the Bayesian approach to diffusions in Section 9.4. An example using simulated data comparing the Euler and Milstein schemes is presented in Section 9.5. The Chapter concludes with Section 9.6.

9.1 INTRODUCTION

The approximation proposed by Milstein (1978), takes the form

$$y_{t+1} = y_t + a(y_t,t;\theta)\Delta^\dagger + b(y_t,t;\theta)(W_{t+1}-W_t) + \frac{1}{2}b(y_t,t;\theta)\frac{db(y_t,t;\theta)}{dy_t}\{(W_{t+1}-W_t)^2 - \Delta^\dagger\}. \quad (9.1)$$

On comparing (9.1) with (2.11), the Milstein is thus simply the Euler plus an additional quadratic term. In Section 9.2, this will be shown to imply a non-central chi-squared distribution for the approximate transition density. The one-step ahead conditional transition density will be seen to have a form which

can be calculated explicitly. Use of the Milstein provides an order of improvement on earlier results obtained from maximising the likelihood using the Euler scheme.

The Milstein density can be used to perform likelihood analysis with or without the introduction of auxiliary points and is therefore applicable to methods discussed in Section 2.5. These include estimating functions, filtering methods, indirect inference and efficient method of moments based estimation methods, the simulated likelihood methods of Pedersen and Geyer and non-parametric methods. However, though the discretisation error is less under Milstein than under Euler, the bias still depends on $\Delta^\dagger \to 0$, unless additional auxiliary points are introduced between the observed data.

If we consider introducing $M_t = M$ auxiliary points between each pair of observations, then given $y^*_{t,j}$, we can obtain the next point in the sequence through

$$\begin{aligned} y^*_{t,j+1} &= y^*_{t,j} + a(y^*_{t,j}, t; \theta)\Delta^\dagger + b(y^*_{t,j}, t; \theta)(W_{t,j+1} - W_{t,j}) \\ &\quad + \frac{1}{2} b(y^*_{t,j}, t; \theta) \frac{db(y^*_{t,j}, t; \theta)}{dy^*_{t,j}} \{(W_{t,j+1} - W_{t,j})^2 - \Delta^\dagger\}, \end{aligned}$$

where $y^*_{t,j}$ represents one of the auxiliary points introduced between the observations $y_{t,0} = y_t$ and $y_{t,M+1} = y_{t+1}$, for $j = 0, \ldots, M$. We have already seen that the introduction of auxiliary points can reduce bias considerably. The one-step ahead conditional transition density can then be defined as

$$f^M(y_{t+1}|y_t; \theta) = \int f(y_{t+1}|y^*_t, y_t; \theta) f(y^*_t|y_t; \theta) dy^*_t$$

where we use the notation $y^*_t = \{y^*_{t,1}, \ldots, y^*_{t,M}\}$.

The Milstein densities are thus integrated out with respect to the auxiliary points. We can therefore perform approximate likelihood inference based upon $f^M(y_{t+1}|y_t; \theta)$, either by using maximum likelihood principles or Bayesian methods. In the former case, the likelihood will be estimated by simulation; the logarithm of the result is then maximised with respect to θ. This type of method is usually called a *maximum simulated likelihood* technique. The Bayesian method will be based on MCMC methods, simulating the augmented state space and the parameters as with the Euler approach. Though both methods currently reduce the bias in parameter estimates, as compared to methods that do not introduce auxiliary points, they rely on the Euler approximation. Use of the Milstein scheme in this context therefore provides a better approximation to the true transition density and hence the parameter estimates.

9.2 The Milstein density

Definitions and the conditions under which the SDE exists were outlined in Section 2.1. Issues concerning the convergence rate of the approximate transition density implied by the Milstein scheme is discussed in 9.2.1. The Euler density is available in closed form and an analytical expression for the Milstein will be provided in 9.2.2.

9.2.1 Orders of Convergence

The addition of one more term to the Euler scheme yields the Milstein scheme which is an order 1.0 strong Taylor scheme. Hence the strong convergence order is increased from $\gamma = 0.5$ to $\gamma = 1$, as outlined

in Section 2.3.3. The Milstein in general, attains a weak order of convergence of $\beta = 2$, if in addition to the assumptions in Appendix 2.7.1, the following conditions hold: the conditional moments of $y_{t+1}|y_t$ closely approximate those of the Itô process; the drift and volatility functions are sufficiently smooth; the time interval is regular, see Kloeden and Platen, (1992, pp. 374-375) for more detail. Suppose we impose the assumptions (A.2.2), (A.2.3), (A.2.9), (A.2.12) and (A.2.14). Then for the Milstein approximation x^δ, the estimate

$$\mathrm{E}\{|y_T - x^\delta|\} \leq K\delta$$

where the constant K does not depend on δ. A proof can again be found in Kloeden and Platen (1992, Section 10.3 and 10.6).

9.2.2 Approximate transition densities

Both the Euler and the Milstein schemes can trivially be used to simulate diffusions. Furthermore, both imply an approximate transition density $f(y_{t+1}|y_t;\theta)$ which can be used for estimation. Convergence of the likelihoods was addressed in Section 2.4.2. For the Euler approximation, we can express conditional distribution of y_{t+1} given y_t and θ as

$$y_{t+1}|y_t;\theta \sim N\{y_t + a(y_t,t;\theta)\Delta^\dagger, b^2(y_t,t;\theta)\Delta^\dagger\}.$$

Use of a higher order scheme reduces the inherent discretisation bias. The Milstein approximation is one alternative that can be used and as far as we know, the density has not appeared in the literature before.

Let $a = a\{y(t),t;\theta\}$ and $b = b\{y(t),t;\theta\}$ for notational convenience. The transition density implied by the Milstein is given by the next theorem

Theorem 9.1. *Given the SDE specified by (2.1), if the usual conditions under which the SDE exists and is unique are satisfied and further, that assumptions (A.2.2), (A.2.3) and (A.2.9)-(A.2.14) hold, then the approximate one-step ahead transition density implied by (9.1) can be expressed as*

$$f(y_{t+1}|y_t;\theta) = \frac{\exp(-\lambda/2)}{|A|\sqrt{2\pi}} z_{t+1}^{-1/2} \exp(-\frac{z_{t+1}}{2}) \cosh(\sqrt{\lambda z_{t+1}}) \tag{9.2}$$

where

$$\begin{aligned} z_{t+1} &= \frac{y_{t+1} - B}{A} \\ \lambda &= \frac{1}{\Delta^\dagger (b')^2} \\ A &= \frac{bb'\Delta^\dagger}{2} \\ B &= -\frac{b}{2b'} + y_t + a\Delta^\dagger - \frac{bb'\Delta^\dagger}{2}. \end{aligned}$$

The density is defined for $z_{t+1} \in \mathbb{R}^+$ *and* $b' \neq 0$.

Proof. Denote y_{t+1} by w, given a fixed value of y_t. From (9.1) it follows that

$$\begin{aligned} w &= y_t + a\Delta^\dagger + b\sqrt{\Delta^\dagger}U + \frac{1}{2}bb'\Delta^\dagger(U^2 - 1) \\ &= A(U+\delta)^2 + B \\ &= Az + B \end{aligned}$$

where $U \sim N(0,1)$, $\delta = 1/(b'\sqrt{\Delta^\dagger})$ and $z = (U+\delta)^2$. Thus z follows a non-central chi-squared distribution with $\nu = 1$ degrees of freedom and non-centrality parameter δ^2. (For a more extensive coverage see Johnson, Kotz, and Balakrishnan, 1995, chapter 29). Letting $\lambda = \delta^2$, the density of z is given by

$$f_z(z) = \frac{1}{2}\exp\left\{-\frac{(\lambda+z)}{2}\right\}\left(\frac{z}{\lambda}\right)^{-1/4} I_{-\frac{1}{2}}(\sqrt{\lambda z})$$

(see Fisher 1928) where

$$I_{-\frac{1}{2}}(x) = \sqrt{\frac{2}{x}}\sum_{j=0}^{\infty}\frac{(x/2)^{2j}}{j!\,\Gamma(j+\frac{1}{2})} = \sqrt{\frac{2}{\pi x}}\cosh(x)$$

is a modified Bessel function of the first kind of order $-\frac{1}{2}$, (see Abramowitz and Stegun, 1970 Ch. 10; the expression in terms of cosh is given in equation 10.2.14). A justification of why this follows naturally is provided in Appendix 9.7.3. So the density of w is given by

$$\begin{aligned} f(w) &= \frac{1}{|A|}f_z\left(\frac{w-B}{A}\right) \\ &= \frac{1}{2|A|}\exp\left\{-\frac{(A\lambda+w-B)}{2A}\right\}\left(\frac{w-B}{\lambda A}\right)^{-\frac{1}{4}}\sqrt{\frac{2}{\pi}}\left\{\frac{\lambda(w-B)}{A}\right\}^{-\frac{1}{4}}\cosh\left\{\sqrt{\frac{\lambda(w-B)}{A}}\right\} \\ &= \frac{1}{|A|}\frac{1}{\sqrt{2\pi}}\left(\frac{w-B}{A}\right)^{-\frac{1}{2}}\exp\left\{-\frac{(A\lambda+w-B)}{2A}\right\}\cosh\left\{\sqrt{\frac{\lambda(w-B)}{A}}\right\}. \end{aligned}$$

This concludes the proof. □

The mean and variance of z are given by $\mathrm{E}(z) = v + \lambda$ and $\mathrm{var}(z) = 2(v+\lambda)$, which give $E(y_{t+1}|y_t;\theta) = y_t + a\Delta^\dagger$ and $var(y_{t+1}|y_t;\theta) = b^2\Delta^\dagger + (bb'\Delta^\dagger)^2/2$. These results can easily be verified from (9.1). This density is non-symmetric, unlike the density in the Euler case. It is also worth noting that the Milstein reduces to the Euler when b' is close to zero, see Appendix 9.7.2.[1] For numerical reasons, it is often easier to work with the exponential function when evaluating the hyperbolic cosine function and so we use $\cosh(x) = (1/2)\{\exp(x) + \exp(-x)\}$.

[1]This means that for small b', the Euler attains an order of convergence which is close to that under the Milstein scheme.

9.3 THE LIKELIHOOD BASED ON THE MILSTEIN DENSITY

We now focus on the approximate likelihood implied by the Milstein scheme. Estimating the parameter of interest involves maximising the likelihood, based on the approximate transition density implied by the stochastic Taylor scheme. We will also present further applications of the Milstein density. These involve introducing missing paths between the observed values, integrating the auxiliary points out of the joint density and then computing the implied conditional transition density. Hence one could employ a simulated likelihood method, see for example, Pedersen (1995b), or a Bayesian Markov Chain Monte Carlo (MCMC) method, conditioning on both neighbouring observations as considered in earlier parts of the thesis.

9.3.1 The likelihood based on the observed points only

When the transition densities are known, the log-likelihood for the data, $l_T(\theta)$ in terms of the conditional distributions of y_{t+1} given y_t and θ, is given by $l_T(\theta) = \sum_{t=1}^{T-1} \log g(y_{t+1}|y_t;\theta)$, where $g(y_{t+1}|y_t;\theta)$ is used to denote the true transition density as defined previously. We work with this representation for reasons outlined in Section 6.1. If the strong solution of the underlying process is available, $g(y_{t+1}|y_t;\theta)$ will be known, and so the maximum likelihood estimator will be consistent. In general, no such strong solution exists and so we approximate $g(y_{t+1}|y_t;\theta)$ with $f(y_{t+1}|y_t;\theta)$ which is implied by the Milstein scheme.

For discretely observed diffusions, the approximate MLE has, under relatively weak conditions, the usual properties of consistency, asymptotic normality and efficiency, see for example Billingsley (1961) and Dacunha-Castelle and Florens-Zmirou (1986). The approximate log-likelihood function takes the form

$$\hat{l}_T(\theta) = \sum_{t=1}^{T-1} \log\{f(y_{t+1}|y_t;\theta)\},$$

where we assume that we can evaluate and sample from the transition density, $f(y_{t+1}|y_t;\theta)$. The time-point T denotes the time period under consideration so that the interval contains all the observed time points. A necessary condition for the existence of a consistent estimator of θ is that $T\Delta \to \infty$, that is, $\hat{\theta}_T \to \theta$ as $T \to \infty$.

As outlined in Section 2.5, the approximation based on discrete observations, (without the introduction of auxiliary points), is biased and inconsistent, see Florens-Zmirou (1993). The direct implication is that the approximate MLE will be inconsistent when the true transition density is unknown. An example of this is shown in Table 9.1. The approximate log-likelihood is computed for two simulated data sets, corresponding to the cases $\Delta^\dagger = 2$ and $\Delta^\dagger = 5$ for the CIR process.[2] The Monte Carlo results of a small simulation study are shown in Table 9.1. The results show that for the drift parameters, the bias and MSE is lower under the Milstein scheme. For the volatility coefficient, there is negligible difference between the two schemes. Results are similar when the parameters are re-parameterized.

[2]Data was obtained using the strong solution of the process.

Monte Carlo Results (CIR process), $\Delta^{\dagger} = 2$					
	true value	mean	bias	s.d.	MSE
Euler					
α	0.5	0.25976	-0.24024	0.01358	0.05790
β	0.2	0.10449	-0.09551	0.00592	0.00916
σ^2	0.05	0.03700	-0.01300	0.00219	0.00017
Milstein					
α	0.5	0.34381	-0.15619	0.03315	0.02549
β	0.2	0.13845	-0.06155	0.01319	0.00396
σ^2	0.05	0.03600	-0.01400	0.00207	0.00020
Monte Carlo Results (CIR process), $\Delta^{\dagger} = 5$					
	true value	mean	bias	s.d.	MSE
Euler					
α	0.5	0.17871	-0.32129	0.00835	0.10329
β	0.2	0.07271	-0.12729	0.00340	0.01621
σ^2	0.05	0.02499	-0.02501	0.00173	0.00063
Milstein					
α	0.5	0.26631	-0.23369	0.01324	0.05478
β	0.2	0.10834	-0.09166	0.00535	0.00843
σ^2	0.05	0.02292	-0.02708	0.00126	0.00074

Table 9.1: *The results from a small simulation study for the parameters of the CIR process are shown using the Euler and Milstein scheme. One hundred sets of data with $T = 500$, $\Delta^{\dagger} = 2$ and $\Delta^{\dagger} = 5$ were generated using the strong solution and used to evaluate the approximate log-likelihoods using a numerical optimisation scheme. The true parameter value, mean, bias, standard deviation and mean square errors are shown.*

9.3.2 A Pedersen type approach to diffusions using the Milstein scheme

The method proposed by Pedersen (1995b), outlined in Sections 2.5.5 and 8.3 can be extended to implement the Milstein scheme as the discretisation scheme.

In practice, denoting the iteration number by k, (for $k = 1, \ldots, N$), each point $y_{t,i+1}^{(k)*}$ is simulated forwards, where $y_{t,i+1}^{(k)*}|y_{t,i}^{(k)*}$ is given by a non-central chi-squared distribution under the Milstein approximation, as

$$y_{t,i+1}^{*(k)} = y_{t,i}^{*(k)} + a\{y_{t,i}^{*(k)};\theta\}\Delta + b\{y_{t,i}^{*(k)};\theta\}\sqrt{\Delta}U_{i+1}^{(k)} + \frac{1}{2}b\{y_{t,i}^{*(k)};\theta\}\frac{\partial b\{y_{t,i}^{*(k)};\theta\}}{\partial y_{t,i}^{*(k)}}\Delta\left[\left\{U_{i+1}^{(k)}\right\}^2 - 1\right] \quad (9.3)$$

for $i = 0, \ldots, M-1$ with U_i denoting the standard Normal. An estimate of $f^M(y_{t+1}|y_t;\theta)$, as defined in (8.3), is then given by

$$\hat{f}^M(y_{t+1}|y_t;\theta) = \frac{1}{N}\sum_{k=1}^{N} f(y_{t+1}|y_{t,M}^{*(k)};\theta) \quad (9.4)$$

so that an estimate of the approximate log-likelihood is given by,

$$\hat{l}_T^{(M)}(\theta) = \sum_{t=1}^{T-1} \log \hat{f}^M(y_{t+1}|y_t;\theta). \tag{9.5}$$

9.4 Bayesian approach to diffusions using Milstein densities

The MCMC method allows us to produce variates from the posterior density $f(\theta|Y,Y^*)$ by repeatedly sampling a Markov chain whose invariant distribution is the target density of interest. The conditional density $f(y_t^*|y_t,y_{t+1};\theta)$ is sampled using the M-H algorithm, except that the Euler discretisation is now replaced by the Milstein. Since these samples typically form a (high dimensional) correlated sample from the target density of interest, we will sample y_t^* in blocks and monitor the correlograms and the inefficiency factors, [3] to gauge how long one should run the sampler for and how accurate the degree of the posterior estimates are. It is again important to construct a proposal that is easy to sample from and to ensure that the algorithm has good mixing properties, as was emphasised in Section 5.

Section 9.4.1 deals with the sampling of Y^* from $Y^*|Y;\theta$, so that given the data Y and denoting $Y^* = (y_1^{*\prime},\ldots,y_{T-1}^{*\prime})'$, we generate M values of $y_t^* = (y_{t,1},\ldots,y_{t,M})$ between each pair of successive observed values y_t and y_{t+1}. Section 9.4.2 then outlines the method for generating the parameters from their posterior distributions, conditional on Y and Y^*, which for these models will often be relatively easy to sample from.

9.4.1 Obtaining the proposal using the Milstein scheme

The aim is to sample Y^* from $f(Y^*|Y;\theta)$, where we will make use of the conditional independence of y_t^* given the observations y_t and y_{t+1}. Suppose we focus on a pair of observations y_t and y_{t+1}; we wish to simulate y_t^* from $f(y_t^*|y_t,y_{t+1}) \propto \Pi_{k=0}^{M} f(y_{t,k+1}^*|y_{t,k}^*)$. Each of these conditional distributions is given by the Milstein approximation defined in equation (9.1). Working with the log-density, this can be expressed as

$$\begin{aligned}\log f(y_{t,k+1}^*|y_{t,k}^*) \quad &= \quad l_1 = c - \log\left|\frac{b_0 b_0' \Delta}{2}\right| - \frac{1}{2}\log d_1 - \frac{d_1}{2} - \frac{1}{2\Delta(b_0')^2} \\ &\quad + \log\left\{\cosh\left(\frac{d_1}{\Delta(b_0')^2}\right)^{\frac{1}{2}}\right\}\end{aligned}$$

where

$$a_1 = a(y_{t,k+1}), \quad b_1 = b(y_{t,k+1}), \quad b_1' = \frac{\partial b}{\partial y_{t,k+1}}, \quad b_1'' = \frac{\partial^2 b}{\partial y_{t,k+1}^2}$$

and

$$d_1 = \left(y_{t,k+1} - y_{t,k} - a_0\Delta + \frac{b_0 b_0'\Delta}{2} + \frac{b_0}{2b_0'}\right) / \left(\frac{b_0 b_0'\Delta}{2}\right).$$

[3] Recall that the inefficiency factors are defined as the variance of the sample mean of the MCMC sampling scheme divided by the variance of the sample mean from a hypothetical sampler drawing independent and identically distributed random variables from the posterior.

Suppose we update blocks of size m, where $(1 \leq m \leq M)$, starting with observation $y^*_{t,k}$, for $(1 \leq k \leq M)$. We thus have a block $y^*_{t(k,m)} = (y^*_{t,k}, y^*_{t,k+1}, \ldots, y^*_{t,k+m-1})$ with neighbours $y^*_{t,k-1}$ and $y^*_{t,k+m}$. For clarity we will denote the block to be updated by $w = (w_1, \ldots, w_m)$, with neighbours w_0 and w_{m+1}. The approximate log density for w given its neighbours implied by the Milstein approximation is

$$\begin{aligned} l(w) &= \log f(w|w_0, w_{m+1}; \theta) \\ &= c + \sum_{j=0}^{m} \log f(w_{j+1}|w_j; \theta) = c + \sum_{j=0}^{m} l_j \end{aligned} \tag{9.6}$$

The (analogous) expression for l_{j+1} is shown in Appendix 9.7.3. To deal with the non-Gaussianity in the likelihood, we perform a second order Taylor expansion of $l(w)$ around $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_m)$, which for simplicity we choose as the linear interpolation between the two fixed neighbours. A candidate value z is sampled from

$$q(w) = q(w|w_0, w_{m+1}; \theta) \sim N(\tilde{\mu}, \tilde{\Sigma}), \tag{9.7}$$

where the neighbours w_0 and w_{m+1} are always the most recently sampled values. Candidate values are then accepted with probability

$$\alpha\{z, w^{(j)}\} = \min\left[1, \frac{f\{z|w_0^{(j+1)}, w_{m+1}^{(j)}; \theta\} q\{w^{(j)}\}}{f\{w^{(j)}|w_0^{(j+1);\theta}, w_{m+1}^{(j)}\} q(z)}\right],$$

where $w^{(j)}$ is the current value of w at iteration j of the MCMC sampling. If the proposal value is rejected, the next item in the chain is taken to be $w^{(j)}$. In theory, analytical forms of the derivatives are required and are shown in Appendix 9.7.3. However, as these expressions can be rather involved, it is easier to work with the numerical derivatives in the simulations.

9.4.2 Sampling the parameters

We now focus on the parameters θ to be sampled from the density $f(\theta|Y, Y^*)$. Denote by $z = (y_1, y^*_{1,1}, \ldots, y^*_{1,M}, y_2, y^*_{2,1}, \ldots, y^*_{T-1,M}, y_T)' = (z_1, \ldots z_n)'$ where $n = (T-1)(M+1)+1$ is the length of z. Here both observed states Y and simulated auxiliary states Y^* are treated as the observed data. Consider the following representation of the conditional likelihood,

$$f(z|\theta) = f(z_2, \ldots z_n|\theta) = \prod_{t=2}^{n} f(z_t|z_{t-1}; \theta), \tag{9.8}$$

where each $f(z_t|z_{t-1}; \theta)$ is given by the conditional density implied by the Milstein approximation. Given this specification we can estimate the parameters within a Bayesian framework. We will use generic methods such as the Adaptive Rejection Metropolis Sampling (ARMS) procedure (see Gilks, Best, and Tan 1995) sampling one parameter at a time (since this method can only be applied univariately). For certain cases, it may be advisable to tailor the algorithms to the model.

9.5 An example

9.5.1 The Cox, Ingersoll and Ross model

The CIR model was defined in Section 5.2.4. For our purposes, we are interested in finding the Milstein approximation and taking a second order Taylor expansion of the log- likelihood. Suppose we take the transformation $x_t = \log y_t$, due to the restriction that $y_t > 0$. Then on applying Itô's lemma, we obtain

$$dx_t = \left\{ \frac{\alpha}{\exp(x_t)} - \beta - \frac{\sigma^2}{2\exp(x_t)} \right\} dt + \frac{\sigma}{\sqrt{\exp(x_t)}} dW_t.$$

The one-step ahead specification under the Milstein scheme given in equation (9.1) is given by,

$$\begin{aligned} x_{t,k+j} &= x_{t,k+j-1} + \left\{ \frac{\alpha}{exp(x_{t,k+j-1})} - \beta - \frac{\sigma^2}{2\exp(x_{t,k+j-1})} + \frac{\sigma^2}{4\exp(2x_{t,k+j-1})} \right\} \Delta \\ &\quad + \frac{\sigma}{\exp\frac{x_{t,k+j-1}}{2}} \epsilon_{t,k+j} - \frac{\sigma^2}{4\exp(2x_{t,k+j-1})} \epsilon^2_{t,k+j} \end{aligned}$$

where $\epsilon_{t,k+j} \sim N(0, \Delta)$ and $j = 0, \ldots, m-1$ for a block size m, $(1 \leq m \leq M)$. The conditional mean and variance are obtained as

$$\begin{aligned} \mathrm{E}(x_{t,k+j}|x_{t,k+j-1};\theta) &= x_{t,k+j-1} + \left\{ \frac{\alpha}{\exp(x_{t,k+j-1})} - \beta - \frac{\sigma^2}{2\exp(x_{t,k+j-1})} \right\} \Delta \\ \mathrm{var}(x_{t,k+j}|x_{t,k+j-1};\theta) &= \frac{\sigma^2\Delta}{\exp(x_{t,k+j-1})} + \frac{\sigma^2\Delta^2}{8\exp(4x_{t,k+j-1})}, \end{aligned}$$

so that the Milstein density for $f(y_{t+1}|y_t;\theta)$ is given by (9.2), with

$$\begin{aligned} a_{t,} &= \frac{\alpha}{exp(x_{t,k+j-1})} - \beta - \frac{\sigma^2}{2\exp(x_{t,k+j-1})} + \frac{\sigma^2}{4\exp(2x_{t,k+j-1})} \\ b &= \frac{\sigma}{\exp\frac{x_{t,k+j-1}}{2}} \\ b' &= -\frac{\sigma}{2\exp\frac{x_{t,k+j-1}}{2}} \end{aligned}$$

The posterior for the parameters is then simply obtained by employing the ARMS procedure proposed by Gilks, Best, and Tan (1995).

9.5.2 Comparison of methods using simulated data

In this Section we apply the methods developed above to the CIR process using the simulated data, in which we will compare the method proposed by Pedersen (1995b) under both the Euler and Milstein scheme and the MCMC algorithm using the Milstein scheme.

The data for our illustration is simulated from the above model with $\alpha = 0.5$, $\beta = 0.2$ and $\sigma^2 = 0.01$. Using an initial value of $y_0 = 0.08$, ten sets of $T = 500$ observations are obtained using the strong solution, with $\Delta^\dagger = 2$. Here $\Delta^\dagger$ can be thought of as the time interval in the observed data and determines the bias in the discretised time gap.

9.5.3 Monte Carlo experiment, MCMC algorithm

The results from a small simulation study for the parameters of the CIR process are shown using the Euler and Milstein scheme in Table 9.2 and 9.3. The MCMC algorithm was run for 1000 iterations, with a burn-in of 200 for $M = 2$ (random block size 1) and $M = 5$ (random block size 2) under both schemes and $M = 0, 3$ in Table 9.3. Results show that the bias and MSE are lower under the Milstein scheme for both of the drift parameters. There is negligible difference between the mean and between the MSE of the estimates under both schemes.

Monte Carlo Results (CIR process, MCMC), $\Delta^{\dagger}$=1				
	mean	bias	s.d.	MSE
Euler				
$M = 2$				
α (0.5)	0.42865	-0.07135	0.00385	0.00657
β (0.2)	0.17812	-0.02188	0.01312	0.00065
σ^2 (0.05)	0.04459	-0.00541	0.00313	3.9×10^{-5}
$M = 5$				
α (0.5)	0.46405	-0.03595	0.00555	0.00132
β (0.2)	0.18979	-0.01021	0.00202	0.00011
σ^2 (0.05)	0.06403	0.01403	0.00035	1.3×10^{-5}
Milstein				
$M = 2$				
α (0.5)	0.43102	-0.06898	0.03631	0.00608
β (0.2)	0.17634	-0.02366	0.01295	0.00073
σ^2 (0.05)	0.04368	-0.00632	0.00283	4.8×10^{-5}
$M = 5$				
α (0.5)	0.47104	-0.02896	0.00624	0.00088
β (0.2)	0.18950	-0.01051	0.00227	0.00012
σ^2 (0.05)	0.04594	-0.00406	0.00039	1.66×10^{-5}

Table 9.2: *The results from a small simulation study for the parameters of the CIR process are shown using the Euler and Milstein scheme. Ten sets of data with $T = 500$ and $\Delta^{\dagger} = 1$ were generated, using the strong solution. The MCMC algorithm was run for 1000 iterations, with a burn-in of 200 for $M = 2$ (random block size 1) and $M = 5$ (random block size 2) under both schemes.*

9.6 Conclusion and extensions

A closed form solution for the one-step ahead conditional transition density using the Milstein scheme has been shown to exist. This higher order Taylor approximation enables us to obtain an order of improvement in accuracy in estimating the parameters in a non-linear diffusion, as compared to the use of the Euler-Maruyama discretisation scheme. The method can easily be extended to estimate the parameters using either the method proposed by Pedersen (1995b), or using a MCMC scheme.

The Milstein scheme has obvious advantages for more non-linear models. The Euler should be used for models which are relatively close to being linear as it is computationally less expensive than the Milstein. In addition, efficiency can be increased by varying the number of auxiliary points. M_t could be endogenised and the spacing made dependent on the conditional variance of the y_t^*, given the fixed neighbours and the current value of the parameters. The idea would then be to increase M_t for larger

Monte Carlo Results (CIR process), $\Delta^{\dagger} = 1$ Milstein MCMC				
	mean	bias	s.d.	MSE
M=0				
α (0.5)	0.42159	-0.07841	0.05786	0.00950
β (0.2)	0.16916	-0.030838	0.02080	0.00138
σ^2 (0.05)	0.04073	-0.00927	0.00301	9.5×10^{-5}
M=3				
α (0.5)	0.47673	-0.02327	0.06289	0.00450
β (0.2)	0.19075	-0.0092543	0.022868	0.00061
σ^2 (0.05)	0.04663	-0.00337	0.00373	2.5×10^{-5}

Table 9.3: *The results from a small simulation study for the parameters of the CIR process are shown using the Milstein scheme. Ten sets of data with $T = 500$ and $\Delta^{\dagger} = 1$ were generated, using the strong solution. The MCMC scheme was run for 1000 iterations with a burn-in of 200. True values of the parameters are given in parentheses. A Comparison over different values of M is shown.*

variances. The Milstein scheme can be used in more complicated multivariate set-ups which is discussed in the subsequent two Chapters.

9.7 Appendix

9.7.1 Notes on proof of Milstein density

This follows naturally by expressing $(2j)!! = (2j)(2j-2)\dots(4)(2)$ and $(2j-1)! = (2j-1)(2j-3)\dots(3)(1)$ which is the j'th moment of a chi-squared random variable with one degree of freedom. Hence $(2j)! = (2j)!!(2j-1)!!$ and $2^j(j!) = (2j)!!$ implying that

$$2^j\,\Gamma(j+\frac{1}{2}) = 2^j(j-\frac{1}{2})(j-\frac{3}{2})\dots(\frac{1}{2})\Gamma(\frac{1}{2}) = (2j-1)!!\,\Gamma(\frac{1}{2}),$$

so that

$$4^j\,j!\,\Gamma(j+\frac{1}{2}) = 2^j\,j!\,(2j-1)!!\,\Gamma(\frac{1}{2}) = (2j)!\,\Gamma(\frac{1}{2}).$$

Hence, going back to the infinite sum,

$$\sum_{j=0}^{\infty}\left\{\frac{\left(\frac{x^2}{4}\right)^j}{j!\,\Gamma(\frac{1}{2})}\right\} = \frac{1}{\Gamma(\frac{1}{2})}\sum_{j=0}^{\infty}\frac{x^{2j}}{(2j)!!} = \frac{1}{\sqrt{\pi}}\cosh(x),$$

using $\mathcal{F}_{-\frac{1}{2}(x)} = \sqrt{\frac{2}{\pi x}}\cos x$; see Gradshteyn and Ryzhik 1965, section 8.464. Here, $\mathcal{F}$ is the real argument component of $I_{\nu(x)} = \exp\left(-\frac{\pi}{2}\nu\, i\right)\,\mathcal{F}_{\nu}\left\{\exp\left(\frac{\pi}{2i}\right)x\right\}$. For an additional reference, see Watson 1944.

In general, Bessel functions with 'half' parameters ($I_{\frac{1}{2}+\nu}$, $\mathcal{F}_{\frac{1}{2}+\nu}$ for $\nu \in \mathbb{Z}$) can be written in terms of trigonometric functions. Other Bessel functions cannot be expressed in such explicit forms.

9.7.2 Milstein reduces to Euler for small b'

Suppose we denote $x = y_{t+1} - y_t - a\Delta$. Then for the Milstein scheme:

$$\begin{aligned} y_{t+1} - B &= x + \frac{b}{2b'} + \frac{bb'\Delta}{2} \\ z_{t+1} &= \frac{y_{t+1} - B}{A} = \frac{x + \frac{b}{2b'} + \frac{bb'\Delta}{2}}{\frac{bb'\Delta}{2}} = 1 + \frac{2x}{bb'\Delta} + \frac{1}{(b')^2\Delta}. \end{aligned}$$

For small b':

$$\begin{aligned} \sqrt{z} &= \sqrt{1 + \frac{2x}{bb'\Delta} + \frac{1}{(b')^2\Delta}} = \sqrt{\frac{1}{(b')^2\Delta}\left\{1 + \frac{2xb'}{b} + (b')^2\Delta\right\}} \\ &\approx \frac{1}{b'\sqrt{\Delta}}\left[1 + \frac{1}{2}\left\{\frac{2xb'}{b} + (b')^2\Delta\right\} + \frac{1}{8}\left\{\frac{2xb'}{b} + (b')^2\Delta\right\}^2 \ldots\right] \\ &\approx \frac{1}{b'\sqrt{\Delta}}\left\{1 + \frac{xb'}{b} + \frac{(b')^2\Delta}{2} - \frac{x^2(b')^2}{2b^2} + \ldots\right\} \end{aligned}$$

so that

$$\begin{aligned} \sqrt{z\lambda} &\approx \frac{1}{(b')^2\Delta}\left\{1 + \frac{xb'}{b} + \frac{(b')^2\Delta}{2} - \frac{x^2(b')^2}{2b^2} + \ldots\right\} \\ \cosh\sqrt{z\lambda} &\approx \frac{1}{2}\exp(\sqrt{z\lambda}). \end{aligned}$$

This implies that

$$\begin{aligned} 2\exp\left(-\frac{\lambda + z}{2}\right)\cosh\sqrt{z\lambda} &\approx \exp\left(-\frac{1}{2} - \frac{x}{bb'\Delta} - \frac{1}{(b')^2\Delta}\right)\exp\left\{\frac{1}{(b')^2\Delta} + \frac{x}{bb'\Delta} + \frac{1}{2} + \frac{x^2}{2b^2}\right\} \\ &= \exp\left(\frac{-x^2}{2b^2}\right). \end{aligned}$$

So, assuming $b > 0$,

$$2|A|\sqrt{z} \to b\sqrt{\Delta}$$

proving that for small b', the Milstein density reduces to that of the Euler.

9.7.3 Obtaining the proposal under the Milstein scheme

In Section 9.4.1, an expression for l_j was obtained using the closed-form one-step ahead transition density implied by the Milstein. Similarly, $\log f(y_{t,k+j+1}|y_{t,k+j})$ can be expressed as

$$\begin{aligned}\log f(y_{t,k+j+1}|y_{t,k+j}) &= l_{j+1} \\ &= c - \log\left|\frac{b_j b_j' \Delta}{2}\right| - \frac{1}{2}\log d_{j+1} - \frac{d_{j+1}}{2} - \frac{1}{2\Delta(b_j')^2} \\ &\quad + \log\left\{\cosh\left(\frac{d_{j+1}}{\Delta(b_j')^2}\right)^{\frac{1}{2}}\right\}.\end{aligned}$$

We are interested in the approximate log density for w given the neighbours given by equation (5.3). To deal with the non-Gaussian element of the likelihood, we propose to perform a second order Taylor expansion of $l(w)$ around a fixed value of w, denoted by $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_m)$. This gives

$$l(w) \approx c + x'u - \frac{1}{2}x'Vx = c' - \frac{1}{2}(x - V^{-1}u)'V(x - V^{-1}u)$$

where $x = (w - \hat{w})$ and

$$\begin{aligned}u &= \frac{\partial l(\hat{w})}{\partial w} = \{u_j\} \quad \text{for } j = 1, \ldots m \\ V &= -\frac{\partial^2 l(\hat{w})}{\partial w \partial w'} = \{V_{ij}\} \quad \text{for } i, j = 1, \ldots, m.\end{aligned}$$

This shows that $l(w)$ is approximately the log of a multivariate Gaussian density with mean $\tilde{\mu} = \hat{w} + V^{-1}u$ and variance matrix $\tilde{\Sigma} = V^{-1}$. Note that for many models, it may be advisable to replace the $N(\tilde{\mu}, \tilde{\Sigma})$ distribution by a multivariate$-t$ density, or some other heavy tailed multivariate distribution.

To derive the derivatives of l_j and l_{j+1} with respect to w_{j+1}, we first use the following notation

$$\begin{aligned}Q_{l_j} &= \sinh\left\{\frac{d_j}{\Delta(b_{j-1}')^2}\right\}^{\frac{1}{2}} \frac{1}{2}\left\{\frac{d_j}{\Delta(b_{j-1}')^2}\right\}^{-\frac{1}{2}} \frac{d_j'}{\Delta(b_{j-1}')^2} \\ Q_{l_{j+1}} &= \sinh\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{\frac{1}{2}} \frac{1}{2}\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{-\frac{1}{2}} \left\{\frac{(b_j')^2 \frac{\partial d_{j+1}}{\partial w_{j+1}} - 2d_{j+1}(b_j')(b_j'')}{\Delta(b_j')^4}\right\}\end{aligned}$$

where

$$\begin{aligned}d_j' &= \frac{\partial d_j}{\partial w_{j+1}} = \frac{b_{j-1}b_{j-1}'\Delta^{-1}}{2} \\ d_{j+1}' &= \frac{\partial d_{j+1}}{\partial w_{j+1}} \\ &= \frac{\frac{b_j(b_j')\Delta}{2}\left[-1 - a_j'\Delta + \frac{\Delta}{2}\{b_j(b_j'') + (b_j')^2\} + \frac{1}{2}\left\{\frac{(b_j')^2 - b_j(b_j'')}{(b_j')^2}\right\}\right] - \frac{d_{j+1}\Delta}{2}\{b_j(b_j'') - (b_j')^2\}}{\left\{\frac{b_j(b_j')\Delta}{2}\right\}^2}.\end{aligned}$$

Then

$$\begin{aligned}
\frac{\partial l_j}{\partial w_{j+1}} &= -\frac{d'_j}{2d_j} - \frac{d_j}{2} + \frac{Q_{l_j}}{\cosh\left\{\frac{d_j}{\Delta(b'_{j-1})^2}\right\}^{\frac{1}{2}}} \\
\frac{\partial l_{j+1}}{\partial w_{j+1}} &= -\left\{\frac{b_j(b''_j)+(b'_j)^2}{b_j b'_j}\right\} + -\frac{d'_{j+1}}{2} + \frac{b''_j}{\Delta(b'_j)^3} + \frac{Q_{l_{j+1}}}{\cosh\left\{\frac{d_{j+1}}{\Delta(b'_j)^2}\right\}^{\frac{1}{2}}}.
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{\partial l}{\partial w_{j+1}} &= -\frac{d'_j}{2}\left(\frac{1}{d_j}+1\right) - \frac{d'_{j+1}}{2}\left(\frac{1}{d_{j+1}}+1\right) + \frac{b''_j}{\Delta(b'_j)^3} - \left\{\frac{b_j b''_j + (b'_j)^2}{b_j b'_j}\right\} \\
&\quad + \frac{1}{2}\tanh\left\{\frac{d_j}{\Delta(b'_{j-1})^2}\right\}^{\frac{1}{2}}\left\{\frac{d_j}{\Delta(b'_{j-1})^2}\right\}^{-\frac{1}{2}}\frac{d'_j}{\Delta(b'_{j-1})^2} \\
&\quad + \frac{1}{2}\tanh\left\{\frac{d_{j+1}}{\Delta(b'_j)^2}\right\}^{\frac{1}{2}}\left\{\frac{d_{j+1}}{\Delta(b'_j)^2}\right\}^{-\frac{1}{2}}\left\{\frac{(b'_j)^2 d'_{j+1} - 2d_{j+1}b'_j b''_j}{\Delta(b'_j)^4}\right\}
\end{aligned}$$

The second derivatives are expressed as follows

$$\begin{aligned}
\frac{\partial^2 l_j}{\partial w_{j+1}^2} &= -\frac{1}{2}\left\{\frac{d_j d_j'' - (d_j')^2}{d_j^2} + d_j''\right\} \\
&\quad + \frac{1}{4}\operatorname{sech}\left\{\frac{d_j}{\Delta(b_{j-1}')^2}\right\}^{\frac{1}{2}}\left\{\frac{d_j}{\Delta(b_{j-1}')^2}\right\}^{-1}\left\{\frac{d_j'}{2\Delta(b_{j-1}')^2}\right\}^2 \\
&\quad - \frac{1}{4}\tanh\left\{\frac{d_j}{\Delta(b_{j-1}')^2}\right\}^{\frac{1}{2}}\left\{\frac{d_j}{\Delta(b_{j-1}')^2}\right\}^{-\frac{3}{2}}\left\{\frac{d_j'}{2\Delta(b_{j-1}')^2}\right\}^2 \\
&\quad + \tanh\left\{\frac{d_j}{\Delta(b_{j-1}')^2}\right\}^{\frac{1}{2}}\left\{\frac{d_j}{\Delta(b_{j-1}')^2}\right\}^{-\frac{1}{2}}\left\{\frac{d_j''}{2\Delta(b_{j-1}')^2}\right\} \\
\frac{\partial^2 l_{j+1}}{\partial w_{j+1}^2} &= -\left[\frac{b_j b_j'(b_j' b_j'' + b_j b_j''' + 2b_j' b_j'') - \{b_j(b_j'' + (b_j')^2\}\{b_j b_j'' + (b_j')^2\}}{(b_j b_j')^2}\right] \\
&\quad - \frac{d_{j+1}''}{2} - \left\{\frac{d_{j+1} d_{j+1}'' - (d_{j+1}')^2}{d_{j+1}^2}\right\} + \frac{1}{\Delta}\left\{\frac{b_j' b_j''' - 3(b_j'')^2}{(b_j')^4}\right\} \\
&\quad + \frac{1}{4}\operatorname{sech}\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{\frac{1}{2}}\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{-1}\left\{\frac{d_{j+1}' - 2d_{j+1}b_j'b_j''}{\Delta(b_j')^2}\right\}^2 \\
&\quad - \frac{1}{4}\tanh\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{\frac{1}{2}}\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{-\frac{3}{2}}\left\{\frac{d_{j+1}' - 2d_{j+1}b_j'b_j''}{\Delta(b_j')^2}\right\}^2 \\
&\quad + \tanh\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{\frac{1}{2}} \\
&\quad \times \left[\frac{(b_j')^2\{d_{j+1}'' - 2d_{j+1}'b_j'b_j'' - 2d_{j+1}(b_j'')^2 - 2d_{j+1}b_j'b_j'''\} - 2b_j'b_j''(d_{j+1}' - 2d_{j+1}b_j'b_j'')}{\Delta(b_j')^4}\right] \\
&\quad \times \left\{\frac{d_{j+1}' - 2d_{j+1}b_j'b_j''}{\Delta(b_j')^2}\right\}.
\end{aligned}$$

The expression for the total second derivative is therefore given by

$$\frac{\partial^2 l}{\partial w_{j+1}^2} = \frac{\partial^2 l_j}{\partial w_{j+1}^2} + \frac{\partial^2 l_{j+1}}{\partial w_{j+1}^2}$$

The only term involving w_{j+1} and w_{j+2} is l_{j+1}, so,

$$\begin{aligned}
\frac{\partial^2 l_{j+1}}{\partial w_{j+1}\partial w_{j+2}} &= \frac{\partial^2 l_{j+1}}{\partial w_{j+2}\partial w_{j+1}} \\
&= -(b_j b_j'\Delta)^{-1} + \frac{(b_j b_j'\Delta)^{-2}}{8d_{j+1}^2} \\
&\quad + \frac{1}{4}\operatorname{sech}\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{-\frac{1}{2}}\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{-1}\left\{\frac{1}{b_j(b_j')^3\Delta^2}\right\}\left\{\frac{(b_j')^2 d_{j+1}' - 2d_{j+1}b_j'b_j''}{\Delta(b_j')^4}\right\} \\
&\quad + \frac{1}{2}\tanh\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{\frac{1}{2}}\left\{\frac{d_{j+1}}{\Delta(b_j')^2}\right\}^{-\frac{1}{2}}\left[\frac{4\{b_j b_j'' - (b_j')^2\}}{(b_j)^3(b_j')^5\Delta^2} - \frac{4b_j''}{\Delta}\right]
\end{aligned}$$

Due to the complicated nature of the derivatives, we have used the numerical derivatives in the algorithms.

CHAPTER 10

A CONTINUOUS-TIME BIVARIATE MODEL WITH UNOBSERVED STOCHASTIC VOLATILITY

Many non-linear models have been proposed to model financial time series that allow for changes in the variance over time. Suppose the return on holding a particular asset at time t, is denoted by r_t, so that the mean adjusted series is y_t. As the level of new information between $t-1$ and t varies, one can expect changes in the variances. The volatility functions, $b^2(r_t, \tau_t; \theta)$, in the univariate models considered in earlier parts of the thesis, specify constant volatility coefficients, which is unrealistic given the empirical evidence. The Vasicek, CIR, Affine CIR and general parametric models applied to the Eurodollar deposit spot-rate, in Section 7, were shown to provide an inadequate fit to the data. Volatility clustering in the residuals suggest that these models should be extended to incorporate an underlying model for the volatility, such as a heavy tailed latent stochastic volatility component, consistent with the empirical evidence in the literature. We now extend the methods developed in Section 5 to consider a bivariate system of SDE's, where one of the processes is unobserved. A review of discrete and continuous-time stochastic volatility models, existing estimation methods and some of the empirical findings of financial times series are presented in 10.1. A method for estimating a bivariate diffusion is presented in 10.2. Dealing with the end points when one of the processes is latent is covered in 10.2.1 and alternative algorithms for estimating a continuous bivariate model with stochastic volatility are given in 10.2.2. The geometric Brownian motion and Generalised CIR models, both incorporating a latent stochastic volatility structure are considered in 10.3. A general formulation, representing either multi-factor models, a system of more than 2 series series, or a combination of series, each with a multi-factor model associated with it, is deferred to Section 11. It will allow us to cover systems of SDE's which may arise in practical situations and extend the dependence structure between processes. The Chapter concludes with Section 10.4.

10.1 A REVIEW OF STOCHASTIC VOLATILITY MODELS

A review of some discrete models with a stochastic variance component is first presented in 10.1.1. The class of autoregressive conditional heteroscedasticity, (ARCH), models, where the conditional variance depends on past values of the series, is briefly reviewed and contrasted with stochastic volatility models, where the conditional variance is additionally allowed to depend on some latent process. The stochastic volatility framework has the advantage that it is closed under temporal aggregation, unlike the ARCH

class. General models of volatility are outlined in 10.1.1. Existing estimation methods for stochastic volatility models in discrete and continuous-time are then reviewed in 10.1.2 and 10.1.3, respectively.

10.1.1 General models of volatility

In a discrete-time setting, the return can be thought of as being represented by some overall mean plus the sum of intra-day movements. An outline of various discrete time models is given in Taylor (1986). Tauchen and Pitts (1983) consider the case for the mean being zero in their model of individual traders reacting to different levels of information. Clark (1973) considers the case where the volatility is a deterministic function and Tauchen and Pitts (1983) assume the volatility process to follow strict white noise. One way of modelling variance changes is to let the return depend on a stochastic process σ_t, (e.g. AR) and a sequence of i.i.d. random variables ξ_t, so that $y_t = \sigma_t \xi_t$ is a stochastic variance model. Modelling the conditional variance as a function of squared past returns and volatilities, corresponds to ARCH models, initially proposed by Engle (1982), subsequently developed in Bollerslev, Chow, and Kroner (1992), Bollerslev and Engle (1993), Bera and Higgins (1993) and reviewed in Bollerslev, Engle, and Nelson (1994).[1] ARCH models can be expressed as discrete-time stochastic difference equations, with the advantage that the likelihood function is analytically tractable and easy to maximise. Variance changes tend to be serially correlated while the series exhibit little dynamic behaviour in the mean, see Bollerslev, Chow, and Kroner (1992). In addition, many financial time series typically exhibit volatility clustering, (groups of highly volatile observations occur together), a feature which is incorporated by generalised ARCH models, or GARCH.[2] For a basic review of stochastic variance models and variations on the ARCH model, see (Harvey 1993, Ch. 8).

A criticism of the ARCH class is that it violates some of the necessary conditions that are required for economic structural interpretation. Specifically, these include temporal aggregation, (whether the model is robust to changes in the sampling frequency),[3] contemporaneous aggregation, (taking linear combinations of univariate models), marginalisation, (relating the multivariate model of conditional heteroscedasticity to the implied univariate models for the components) and obtaining the stationary densities. In cases where the stationary density exists, it is unknown, (though in recent work, Pitt and Walker (1998) use the t-ARCH(1) to construct a stationary density using the Gibbs sampler, with Student-t marginals). Drost and Nijman (1993) show that even the simplest ARCH model is not closed under temporal aggregation, having to restrict attention to symmetric weak GARCH models to achieve this property. Nelson, (1990, 1995) examines conditions under which ARCH systems converge in distribution to Itô processes, as the time interval between observations goes to zero. An example is the GARCH(1,1), where the stationary distribution of the conditional variance is not known, but shown to be an inverse Gamma distribution in continuous-time. This implies, that for short time intervals, the marginal distribution for the volatility is approximately distributed as a Student-t. In addition, ARCH diffusions are often criticised for not fitting the data. Though the model produces characteristic autocorrelation functions for y_t^2, empirical plausibility often requires the model to additionally incorporate a jump diffusion term. This has been noted in Barndorff-Nielsen and Shephard (1999), where the lack of

[1] These include the variations of ARCH, in particular, the original ARCH(p) specification of Engle (1982), the GARCH model of Bollerslev (1986), the GARCH-M model of Engle and Bollerslev (1986), the log-GARCH model of Pantula (1986) and Geweke (1986) and the exponential ARCH model of Nelson (1991).

[2] Clustering is directly related to ARMA structures of the squared residuals, a defining feature of GARCH models.

[3] If closed under aggregation, it would allow use of a discrete ARCH model to estimate a diffusion model, when an analytical solution is not available for the latter. Conversely, if distributional properties of the diffusion limit of an ARCH model are known, but unavailable for the ARCH model, the diffusion could be used as an approximation.

flexibility, due to only having three parameters - the average level, persistence and magnitude in persistence of volatility, is discussed. The marginal distribution should be left to vary freely, but under ARCH, is implied by the relationship between the parameters. Imposing symmetry in the model rules out incorporation of a leverage effect, as this creates skewness at low frequencies.[4] Meddahi and Renault (1996), (see also Meddahi and Renault, 1999a, 1999b), discuss the square-root stochastic autoregressive volatility (SR-SARV) model. The model captures the ARMA dynamics for squared innovations, (representative of a GARCH process), is consistent with respect to aggregation and can be generalised multivariately. Starting with a continuous-time multi-factor model, they obtain a class of discrete-time SR-SARV(p), (with ARMA(p,p) dynamics for the squared process), and hence a relation between the discrete and continuous-time representations. Treating the model in state-space form, conditional moment restrictions are derived, enabling estimation using method of moments based estimators.[5] Extending to the class of SR-SARV-*in the mean*, allows the mean and variance to share common factors in discrete and continuous-time.

10.1.2 The latent stochastic volatility model

The stochastic volatility (SV) model, unlike the ARCH class, is closed under aggregation allowing structural economic interpretation. Suppose y_t denotes the mean-adjusted return at time τ_t, then the standard SV model can be expressed as

$$\begin{aligned} y_t &= \beta \exp\left(\frac{h_t}{2}\right)\xi_t \\ h_t &= \mu_h + \phi(h_{t-1} - \mu_h) + \sigma_\nu \nu_t \quad \text{for } t \geq 2 \\ h_1 &\sim N\left(\mu_h, \frac{\sigma_\nu^2}{1-\phi^2}\right), \quad \begin{matrix} \xi_t \sim NID(0,1) \\ \nu_t \sim NID(0,1) \end{matrix} \end{aligned} \tag{10.1}$$

The log volatility is denoted by h_t and the constant scaling parameter β or $\exp(\mu_h/2)$ is the average volatility level. The parameter σ_ν^2 is the volatility of the volatility and the persistence in the volatility can be interpreted from ϕ. Kim, Shephard, and Chib (1998) discuss restrictions on μ_h and β, so that the model is identified, stating that either $\beta = 1$, or $\mu_h = 0$. Typically, β is set to unity so that μ_h is left unrestricted, but for reasons of economic interpretation, $\beta = \exp(\mu_h/2)$ is reported.[6] Ghysels, Harvey, and Renault (1996) note that the basic SV model seems to capture most of the empirical patterns found in financial time series. Reviews can be found in Shephard (1996), Taylor (1994) and Jacquier, Polson, and Rossi (1994). The asymmetric behaviour typically reported in stock prices, (see, for example, Nelson 1991, Schwert 1989 and Campbell and Kyle 1993), can be incorporated by allowing ξ_t and ν_t to be dependent. Empirical findings of leptokurtosis and skewness, (right tails are found to be more extreme), of the unconditional distribution of sample variances are most often reported for financial data. To incorporate this stylised feature, the model should give a right skewed, positive valued distribution. The log normal has been proposed by Clark (1973) and Tauchen and Pitts (1983). Squared returns typically exhibit pronounced serial correlation with little or no serial dependence. In addition, autocorrelations of the volatility and squared volatility (as with absolute and squared returns) have been

[4] In stock data, a rise in volatility associated with a fall in the stock price is termed a leverage effect.

[5] The continuous-time representation of the SR-SARV(p) model specifies a multi-factor model for the variance and is considered by Gallant, Hsu, and Tauchen (1998).

[6] Setting $\beta = 1$ allows direct interpretation between the continuous and discrete models, see (10.1.3).

shown to be positive for several lags. Gallant and Tauchen (1997a) have observed that long financial time series can sometimes exhibit structural regime switches.

The model can be linearised as

$$\log y_t^2 = \log \beta^2 + h_t + \epsilon_t$$

where $\epsilon_t = \log \xi_t^2$ with mean -1.27 and variance $\pi^2/2$. The errors ϵ_t and ν_t are contemporaneously uncorrelated. If the initial model is extended to allow correlation between ξ_t and ν_t, then the correlation coefficient can be recovered by conditioning on the signs of the original returns, see Harvey and Shephard (1996). The Quasi-maximum likelihood method (QMLE) based on the Kalman filter (KF) was proposed by Harvey, Ruiz, and Shephard (1994) in which (ν_t, ϵ_t) are jointly normal on assuming $\epsilon_t \sim N(-1.27, \pi^2/2)$. However, ϵ_t is non-Gaussian, since $\xi^2 \sim \chi_1^2$. This implies that

$$\begin{aligned} f(\xi^2) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\xi} \exp\left(-\frac{\xi^2}{2}\right) \\ f(\epsilon) &= \frac{1}{\sqrt{2\pi}} \exp\left\{\frac{\epsilon - \exp(\epsilon)}{2}\right\} \end{aligned}$$

where the transformation $\epsilon = \log \xi^2$ is applied. Sandmann and Koopman (1998) show how far the normal approximation can differ to the actual distribution of ϵ_t which is right-skewed and has a long left tail spanning the negative axis. It is also inefficient in the sense that the information regarding the form of the true likelihood is not being used. Shephard (1994) suggests a Bayesian approach which is more precise.

Alternative methods have been proposed which include estimators based on method of moments (MM) and MCMC methods. MM based estimators have been proposed by Taylor (1986), Melino and Turnbull (1990), Duffie and Singleton (1993), Hansen and Scheinkman (1995) and Andersen and Sørensen (1996). However, they can generally be inefficient and sub-optimal to a likelihood based method, see Andersen and Sørensen (1996) and Jacquier, Polson, and Rossi (1994). Gallant and Tauchen (1996) propose to match the score of the auxiliary model through simulation, but are unable to obtain an estimate of the latent volatility. Extensions have been proposed by Andersen (1994) and Ghysels and Jasicek (1996) to allow estimates of the volatility to be obtained via the KF. Kim, Shephard, and Chib (1998) approximate the distribution of ϵ_t by a mixture of normal distributions. Bayesian methods based on MCMC methods have been proposed by Jacquier, Polson, and Rossi (1994), (though convergence of the chain can be slow); see also Geweke (1994) and Shephard and Kim (1994). Shephard and Pitt (1997) demonstrate that a block MCMC algorithm outperforms the JPR method on computational efficiency. A simulated likelihood method has also been proposed by Danielsson (1994). Shephard and Pitt (1997), Durbin and Koopman (1997) and Durbin and Koopman (1998) construct a likelihood function for the general state space model using simulation techniques.

10.1.3 Stochastic volatility models in continuous-time

Many techniques for estimating continuous-time SV models have been proposed; good reviews are provided in Engle and Lee (1996), Shephard (1996), Ghysels, Harvey, and Renault (1996) and Rossi, (1996, Parts II and III).

A class of continuous-time SV models is considered in Barndorff-Nielsen and Shephard (1999), which can be simulated without inducing any discretisation error. They examine the use of non-Gaussian OU processes in finance. Hull and White (1987) and Chesney and Scott (1989) approximate the continuous version with the discrete-time model defined in (10.1). The indirect inference method proposed by Gourieroux, Monfort, and Renault (1993) and Gallant and Tauchen (1996), (which is reviewed in Section 2.5.4), has also been used to estimate the continuous-time SV model.

Suppose we consider a continuous-time model for returns, with realisations $r(t)$, whose conditional variance depends upon past observations and a continuous-time latent SV process, $\{\sigma^2(t)\}$ for $t \in [0, T]$. The Black and Scholes (1973) model is often extended to allow volatility to follow a diffusion process, that is,

$$dr(t) = \mu r(t)dt + \sigma(t)r(t)dW(t)_1 \tag{10.2}$$

$$d\log\sigma^2(t) = \kappa_2\{\alpha - \log\sigma^2(t)\}dt + \xi dW(t)_2 \tag{10.3}$$

where α is the instantaneous rate of return and $\{W(t)_1, W(t)_2\}$ are Wiener processes, possibly correlated with coefficient ρ, so that $dW(t)_1\, dW(t)_2 = \rho\, dt$, with $|\rho| < 1$. The process to be augmented is given by $\{\log\sigma^2(t)\}_{t=1}^T$. The parameter vector of interest is $\theta = (\mu, \alpha, \kappa_2, \xi)'$, consisting of the expected return μ and the hyper-parameters (α, κ_2, ξ). The specification (10.2–10.3) has received considerable attention in the literature. It has recently been considered by Gallant, Hsu, and Tauchen (1998) as a special case of a two-factor model encompassing various specifications that have been estimated and tested in empirical work.[7] Wiggins (1987), Scott, (1987, 1991), Hull and White, (1987, 1988), Stein and Stein (1991) and Heston (1993) all considered models of the form (10.2), with $\sigma(t)$ following a continuous-time process.[8] If $r(t)$ is not assumed to cause the underlying volatility in the model, then it is natural to assume that $W(t)_1$ and $W(t)_2$ are stochastically independent. That is, intra-day price changes are independent of how many price changes there are and $r(t)$ can be thought of as being driven by the level of market activity and the reaction of the price to that level of activity. There is strong empirical evidence in the literature that $\rho \neq 0$, implying that the model should realistically allow for a negative correlation between $\log\sigma^2(t)$ and $r(t)$. Extending the framework to allow the process in (10.3) to depend on $r(t)$ is deferred to Section 11.

An alternative class of models is the constant elasticity of volatility, (CEV), model proposed by Cox and Ross (1976), extended to the generalised CIR, (Cox, Ingersoll, and Ross 1985, Hull and White 1987 and

[7]The general two factor model in Gallant, Hsu, and Tauchen (1998) is given by

$$dU_1(t) = \mu_1 dt + \mu_2 \exp\{U_2(t)\}\, dW(t)_1$$
$$dU_2(t) = \mu_3 U_2(t)dt + \{\mu_4 + \mu_5 U_2(t)\}\, dW(t)_2.$$

[8]If $q(t)$ denotes a function of $\sigma(t)$, then examples of models appearing in the literature include,

$$dq(t) = q(t)\left[\{\alpha_1 - \alpha_2 q(t)\}\, dt + \xi dW(t)_2\right] \tag{10.4}$$

$$dq(t) = \alpha_2\{\alpha_1 - q(t)\}dt + \xi dW(t)_2 \tag{10.5}$$

$$dq(t) = \left\{\frac{\alpha_1}{q(t)} - \alpha_2 q(t)\right\} dt + \xi dW(t)_2. \tag{10.6}$$

(10.4) was considered by Scott (1987), Wiggins (1987), with $\alpha_2 = 0$ and Hull and White (1987), with $\alpha_2 = \rho = 0$; both used $q(t) = \log\sigma(t)$. Note that taking $\sigma(t)$ to be the positive root of $\sigma^2(t)$ ensures positivity. Scott (1987) proposed (10.5), which was examined further by Stein and Stein (1991), both took $\rho = 0$ and $q(t) = \sigma^2(t)$. (10.6) was proposed by Hull and White (1988) and Heston (1993), with $q(t) = \sigma(t)$.

Taylor 1982), where

$$dr(t) = \kappa_1\{\mu - r(t)\}dt + \sigma(t)r(t)^\gamma dW(t)$$

where $\gamma > 0, (\neq 0.5)$ and κ_1 and μ are non-negative. This is considered by Johnson and Shanno (1987), with a CEV specification for $\sigma^2(t)$ and Melino and Turnbull (1990), where the log-volatility is assumed to follow an OU process. Both the generalised CIR and GBM models will be considered in 10.3.1 and 10.3.2.

In our MCMC framework, discretising both SDE's and introducing $h_t = \log \sigma_t^2$ for ease of notation, we know that, $h_{t+1}|h_t; \kappa_2, \alpha, \xi \overset{\text{approx}}{\sim} N\{\kappa_2\alpha\Delta^\dagger + (1-\kappa_2\Delta^\dagger)h_t,\ \xi^2\Delta^\dagger\}$. As an example, suppose r_t is defined as the exchange rate at time t and y_t is the mean-adjusted return, so that in practice,

$$y_t = 100 \times \left\{(\log r_t - \log r_{t-1}) - \frac{1}{T}\sum_{i=1}^{T}(\log r_i - \log r_{i-1})\right\},$$

see Kim, Shephard, and Chib (1998). Introducing additional auxiliary points between each time point, the states to be augmented are $r_t^* = \{r_{t,j}^*\}_{j=1}^M$ and $\{h_{t,j}\}_{j=0}^{M+1}$, for $t = 1, \ldots, T$. In terms of the discrete SV model defined in (10.1), $\phi = 1 - \kappa_2\Delta$, $\mu_h = \alpha$ and $\sigma_\nu = \xi\sqrt{\Delta}$.

10.2 Estimating a bivariate diffusion

Suppose we consider a bivariate system of SDE's given by

$$\begin{aligned} dr(t) &= a\{r(t), h(t), t; \theta\}dt + b\{r(t), h(t), t; \theta\}dW(t)_1 \\ dh(t) &= e\{h(t), t; \theta\}dt + n\{h(t), t; \theta\}dW(t)_2 \end{aligned} \tag{10.7}$$

where $W(t)_1$ and $W(t)_2$ are assumed to be independent and the drift and volatility functions of the $h(t)$ process are not assumed to depend on $r(t)$. Conditions on the drift and volatility functions under which the SDE can be solved as a diffusion $X(t) = \{r(t), h(t)\}'$ are assumed to be satisfied, (see, for example, Strook and Varadhan, 1979, Ch. 5). On discretising the process and applying the Euler scheme,

$$\begin{aligned} r_{t+1} &= r_t + a(r_t, h_t, t, \theta)\Delta + b(r_t, h_t, t; \theta)\sqrt{\Delta}\epsilon_{t+1} \\ h_{t+1} &= h_t + e(h_t, t, \theta)\Delta + n(h_t, t; \theta)\sqrt{\Delta}\xi_{t+1}. \end{aligned}$$

Suppose we define the measurements and auxiliary states as,

$$\begin{array}{ll} h = h_1, \ldots, h_T & r = r_1, \ldots, r_T \\ h_t^* = \{h_{t,k}\}_{k=1}^M & r_t^* = \{r_{t,k}\}_{k=1}^M \\ h^* = \{h_t^*\}_{t=1}^T & r^* = \{r_t^*\}_{t=1}^T \\ h_{t,M+1} = h_{t+1,0} & r_{t,M+1}^* = r_{t+1,0}^* \end{array}$$

and in addition define

$$\begin{aligned} h_{\backslash t} &= \{h_1, \ldots, h_{t-1}, h_{t+1}, \ldots, h_T\} \\ r_{\backslash t} &= \{r_1, \ldots, r_{t-1}, r_{t+1}, \ldots, r_T\}, \end{aligned}$$

so that there are $n = (M+1)T - M$ distinct time points for each process. Between each pair of observations r_t and r_{t+1}, the states $r_t^*|r_t, r_{t+1}, h, h^*; \theta$ can be updated as outlined in Section 5 and no new issues arise. Augmentation of h_t^* requires a modification of existing methods, since,

$$f(h_t|r, h_{\backslash t}; \theta) \propto f(r_{t+1}|r_t, h_t; \theta)\, f(h_t|h_{t-1}, h_{t+1}; \theta). \tag{10.8}$$

Consider a block of size $m, (1 \leq m \leq M)$, starting with observation $h_{t,k}, (1 \leq k \leq m)$ and consider augmentation between the time points t and $t+1$. We then have a block $(h_{t,k}, \ldots, h_{t,k+m-1}) = (w_1, \ldots, w_m) = w$, with neighbours $h_{t,k-1} = w_0$ and $h_{t,k+m} = w_{m+1}$. On letting $r_{t,k}, \ldots, r_{t,k+m-1} = v = v_1, \ldots v_m$ we can then decompose the block into

$$\begin{aligned} f(w|w_0, w_{m+1}, v; \theta) \quad \propto \quad & f(w_1|w_0; \theta)\, f(v_2|v_1; w_1; \theta) \\ & f(w_2|w_1; \theta)\, f(v_3|v_2; w_2; \theta) \\ & \vdots \\ & f(w_m|w_{m-1}; \theta)\, f(v_{m+1}|v_m; w_m; \theta) \\ & f(w_{m+1}|w_m; \theta) \\ \text{so that } l(w) \quad = \quad & \log f(w|w_0, w_{m+1}, v; \theta) \\ = \quad & c + \sum_{j=0}^{m-1} \{\log f(w_{j+1}|w_j; \theta) + \log f(v_{j+2}|v_{j+1}, w_{j+1}; \theta)\} + \log f(w_{m+1}|w_m; \theta) \\ = \quad & c + \sum_{j=0}^{m} l_{t,j}^{(1)} + \sum_{j=0}^{m-1} l_{t,j+1}^{(2)} \end{aligned}$$

where $l_{t,j}^{(1)} = \log f(w_{j+1}|w_j; \theta)$ and $l_{t,j+1}^{(2)} = \log f(v_{j+2}|v_{j+1}, w_{j+1}; \theta)$. We perform a second order Taylor expansion $l(w)$ around $\hat{w} = (\hat{w}_1, \ldots, \hat{w}_m)$, so that

$$l(w) \approx c + x' u_t - \frac{1}{2} x' V_t x = c' - \frac{1}{2}(x - V_t^{-1} u_t)' V_t (x - V_t^{-1} u_t),$$

where $x = w - \hat{w}$ and

$$u_t = \frac{\partial l(\hat{w})}{\partial w} \quad \text{and} \quad V_t = -\frac{\partial^2 l(\hat{w})}{\partial w \partial w'}.$$

Then we know that $l(w) \approx \log N(\tilde{\mu}, \tilde{\Sigma})$, where

$$\begin{aligned} \tilde{\mu} &= \begin{cases} \hat{w} + V_t^{-1} u_t & \text{if } V_t \text{ is positive definite} \\ \hat{w} + V_t^{*\,-1} u_t & \text{otherwise} \end{cases} \\ \tilde{\Sigma} &= \begin{cases} V_t^{-1} & \text{if } V_t \text{ is positive definite} \\ V_t^{*\,-1} & \text{otherwise} \end{cases} \end{aligned}$$

Details are given in Appendix 10.5.1.[9]

Given a set of observations, the process is then updated conditional on the observed values. Regarding notation, we will differentiate between $a_{t,j}$, $b^2_{t,j}$, as defined in Section 5, which denotes the drift and volatility as functions of $r_{t,k+j}$, and $a(h_{t,j})$, $b^2(h_{t,j})$, to denote the drift and volatility as functions of $h_{t,k+j}$, so that,

$$\begin{aligned} a(h_{t,j}) &= a(r_{t,k+j}, h_{t,k+j}) = a(r_{t,k+j}, w_{j+1}) \\ b(h_{t,j}) &= b(r_{t,k+j}, h_{t,k+j}) = b(r_{t,k+j}, w_{j+1}) \\ e_{t,j} &= e(h_{t,k+j}) = e(w_{j+1}) \\ n_{t,j} &= n(h_{t,k+j}) = n(w_{j+1}) \end{aligned}$$

for $j = -1, 0, \ldots, m-1$. Let $e'_{t,j}$ and $e''_{t,j}$ denote the first and second derivatives of $e(w_{j+1})$ with respect to w_{j+1} and define $a(h_{t,j})'$, $a(h_{t,j})''$, $b(h_{t,j})'$, $b(h_{t,j})''$ and $n'_{t,j}$, $n''_{t,j}$ similarly. When the structure is completely latent, as with a stochastic volatility component, a slight modification is required for the end points.

10.2.1 Dealing with the end points for a latent volatility structure

In the case where h_t is latent for all t, the neighbours in each block can be updated by including the endpoint in the next block to be updated. However, this will not be true of $h_{1,0}$ and $h_T = h_{T-1,M+1}$. These are therefore treated separately, by updating $h_{1,0}$ and h_T at the start and end of each sweep. Now

$$f(h_{1,0}|h_{1,1}, r, r^*; \theta) \propto f(h_{1,1}|h_{1,0}; \theta)\, f(r^*_{1,1}|r^*_{1,0}, h_{1,0}; \theta)$$

so that $l(h_{1,0}) = c + l^{(1)}_{1,1} + l^{(2)}_{1,1}$, where

$$\begin{aligned} l^{(1)}_{1,1} &= \log f(h_{1,1}|h_{1,0}; \theta) = c - \frac{1}{2}\log(n^2_{1,0}\Delta) - \frac{1}{2n^2_{1,0}\Delta}(h_{1,1} - h_{1,0} - e_{1,0}\Delta)^2 \\ l^{(2)}_{1,1} &= \log f(r_{1,1}|r_{1,0}, h_{1,0}; \theta) = c - \frac{1}{2}\log\{b(h_{1,0})^2\Delta\} - \frac{1}{2b(h_{1,0})^2\Delta}\{r_{1,1} - r_{1,0} - a(h_{1,0})\Delta\}^2. \end{aligned}$$

[9]Note that forms of V_t differ depending on whether or not it is negative definite. The general procedure will be to compute V_t based on $V^t_{j+1,j+1}$ and examine whether or not it is p.d. In the case where it is not, V_t is re-computed using $V^{*t}_{j+1,j+1}$, where

$$\begin{aligned} V_t &= \{V^t_{ij}\}^m_{i,j=1} \\ V^t_{j+1,j+1} &= -\frac{\partial^2 l^{(1)}_{t,j}}{(\partial w_{j+1})^2} - \frac{\partial^2 l^{(1)}_{t,j+1}}{(\partial w_{j+1})^2} - \frac{\partial^2 l^{(2)}_{t,j+1}}{(\partial w_{j+1})^2} \\ V^{*t}_{j+1,j+1} &= -\frac{\partial^2 l^{(1)}_{t,j}}{(\partial w_{j+1})^2} - \frac{\partial^2 l^{(1)}_{t,j+1}}{(\partial w_{j+1})^2} \end{aligned}$$

for $j = 0, \ldots, m-1$. Note that V^t_{ij}, for $i \neq j$, is the same in both cases as it doesn't depend on $l^{(2)}_{t,j+1}$. Details are given in Appendix 10.5.1.

The first and second derivatives with respect to $h_{1,0}$ are given in Appendix 10.5.1 and

$$\begin{aligned} u_0 &= \frac{\partial l(w)}{\partial h_{1,0}} = \frac{\partial l_{1,1}^{(1)}}{\partial h_{1,0}} + \frac{\partial l_{1,1}^{(2)}}{\partial h_{1,0}} \\ V_0 &= -\frac{\partial^2 l(w)}{(\partial h_{1,0})^2} = \begin{cases} -\frac{\partial^2 l_{1,1}^{(1)}}{(\partial h_{1,0})^2} - \frac{\partial^2 l_{1,1}^{(2)}}{(\partial h_{1,0})^2} & \text{if } V_0 \text{ is positive definite} \\ -\frac{\partial^2 l_{1,1}^{(1)}}{(\partial h_{1,0})^2} & \text{otherwise} \end{cases} \end{aligned}$$

which are evaluated at $\hat{h}_{1,0}$, so that the proposal is $q(h_{1,0}|h_{1,1}, r; \theta) \sim N(\hat{h}_{1,0} + V_0^{-1}u_0, V_0^{-1})$.

For the remaining endpoint h_T,

$$\begin{aligned} f(h_T|h_{T-1,M}, r, r^*; \theta) &= f(h_T|h_{T-1,M}; \theta) \\ \text{so that } \log f(h_T|h_{T-1,M}, r, r^*; \theta) &= l(h_T) = c + l_{T-1,M}^{(1)}, \end{aligned}$$

where $l_{T-1,M}^{(1)}$ is given in Appendix 10.5.1. Hence,

$$\begin{aligned} u_T &= \frac{\partial l(w)}{\partial h_T} = \frac{\partial l_{T-1,M}^{(1)}}{\partial h_T} \\ V_T &= -\frac{\partial^2 l(w)}{(\partial h_T)^2} = -\frac{\partial^2 l_{T-1,M}^{(1)}}{(\partial h_T)^2} \end{aligned}$$

and are evaluated at $\hat{h}_T$, so that the proposal is $q(h_T|h_{T-1,M}, r; \theta) \sim N(\hat{h}_T + V_T^{-1}u_T, V_T^{-1})$.

10.2.2 Alternative algorithms

When both models are linear, the MH procedure defined in Section 10.2 collapses to a Gibbs sampler. In some cases where the second derivative of the log-likelihood with respect to the latent states, $\partial^2 l(w)/(\partial w_{j+1})^2$, is negative definite, we can use a simpler procedure. This takes the form of an Accept/Reject, (A/R), algorithm. The crucial difference between the A/R algorithm and the MH scheme proposed above, is that a point which is rejected is re-sampled. Though restricted to specific models, the advantage of A/R is its simplicity - since there is no autocorreltaion, (unlike either of the MH or Gibbs samplers), there is no need to compute either the acceptance rates, or the inefficiency factor estimates. In this case, points are also obtained from the exact posterior distribution. However, for most non-linear SDE's we consider, (e.g. the geometric Brownian motion and generalised CIR models), the second derivative is not negative definite, preventing us from applying this procedure.

In some cases, procedures outlined in Kim, Shephard, and Chib (1998) and Shephard and Pitt (1997) can be followed. First, from (10.8) we know that,

$$f(h_{t,k}|h_{\backslash t,k}, r; \theta) \propto f(r_{t,k+1}|r_{t,k}, h_{t,k}; \theta)\ f(h_{t,k}|h_{t,k-1}, h_{t,k+1}; \theta)$$

where $f(h_{t,k}|h_{t,k-1}, h_{t,k+1}; \theta)$ can be thought of as a conditional prior density that dominates $f(r_{t,k+1}|r_{t,k}, h_{t,k}; \theta)$. Note that, $f(h_{t,k}|h_{t,k-1}, h_{t,k+1}; \theta)$ is not strictly a prior, since $h_{t,k}$ is updated using information contained

in the observations. Additionally,

$$\begin{aligned} f(h_{t,k}|h_{t,k-1}, h_{t,k+1}; \theta) &\propto f(h_{t,k}|h_{t,k-1}; \theta)\, f(h_{t,k+1}|h_{t,k}; \theta) \\ &= \frac{1}{\sqrt{2\pi\xi^2\Delta}} \exp\left[-\frac{1}{2\xi^2\Delta}\{h_{t,k+1} - h_{t,k} - \kappa_2(\alpha - h_{t,k})\Delta\}^2\right] \\ &\quad \times \frac{1}{\sqrt{2\pi\xi^2\Delta}} \exp\left[-\frac{1}{2\xi^2\Delta}\{h_{t,k} - h_{t,k-1} - \kappa_2(\alpha - h_{t,k-1})\Delta\}^2\right], \end{aligned}$$

which can be expressed as

$$\log f(h_{t,k}|h_{t,k-1}, h_{t,k+1}; \theta) = c - \frac{\{1 + (1 - \kappa_2\Delta)^2\}}{2\xi^2\Delta}\left[h_{t,k} - \left\{\frac{(h_{t,k-1} + h_{t,k+1})(1 - \kappa_2\Delta) + \rho\kappa_2\Delta^2}{1 + (1 - \kappa_2\Delta)^2}\right\}\right]^2$$

so that $h_{t,k}|h_{\backslash t,k} \sim N(\mu_{t,k}, \Sigma_{t,k})$, where

$$\begin{aligned} \Sigma_{t,k} &= \frac{\xi^2\Delta}{1 + (1 - \kappa_2\Delta)^2} \\ \mu_t &= \frac{(h_{t,k-1} + h_{t,k+1})(1 - \kappa_2\Delta) + \rho\kappa_2\Delta^2}{1 + (1 - \kappa_2\Delta)^2}. \end{aligned}$$

Let $\log f(r_{t,k+1}|h_{t,k}, r_{t,k}; \theta) = l(r_{t,k}, h_{t,k})$ and define

$$\frac{\partial l(r_{t,k}, h_{t,k})}{\partial h_{t,k}} = l'(r_{t,k}, h_{t,k}) \quad \text{and} \quad \frac{\partial^2 l(r_{t,k}, h_{t,k})}{\partial h_{t,k}\partial h_{t,k}'} = l''(r_{t,k}, h_{t,k})$$

and assume $\Sigma_{t,k}$ to be non-singular. The prior for the latent process could be set to

$$h_{1,0} \sim N\left\{\alpha, \frac{\xi^2}{\kappa_2(2 - \kappa_2\Delta)}\right\},$$

which is the unconditional mean and variance for a stationary process, (in which $|1 - \kappa_2\Delta| < 1$). If $l''(r_{t,k}, h_{t,k})$ is negative semi-definite for all $h_{t,k}$, then we can use the following theorem, taken from Shephard and Pitt (1997).

Theorem 10.1. *We can sample from $h_{t,k}|h_{t,k-1}, h_{t,k+1}, r_{t,k}; \theta$ by making suggestions $h_{t,k}$ from*

$$N\{\mu_{t,k} + \Sigma_{t,k} l'(r_{t,k}, \hat{h}_{t,k}),\ \Sigma_{t,k}\},$$

which are accepted with probability

$$\exp\left\{l(r_{t,k}, h_{t,k}) - l(r_{t,k}, \hat{h}_{t,k}) - l'(r_{t,k}, \hat{h}_{t,k})'(h_{t,k} - \hat{h}_{t,k})\right\} \ \forall\, m_{t,k}.$$

The probability of rejection is minimised by choosing $\hat{h}_{t,k}$ as the mode of $f(h_{t,k}|h_{t,k-1}, h_{t,k+1}, r_{t,k}; \theta)$.

$\hat{h}_{t,k}$ can be recursively computed within each sweep as

$$\hat{h}_{t,k}^{(i+1)} = \mu_{t,k} + \Sigma_{t,k} l' \left\{ r_{t,k}, \hat{h}_{t,k}^{(i)} \right\} \quad \textit{where } \hat{h}_{t,k}^{(0)} = \mu_{t,k}.$$

In terms of the model in (10.7), we know that

$$r_{t,k+1} | r_{t,k}, h_{t,k} \overset{\text{approx}}{\sim} N \left\{ r_{t,k} + a(r_{t,k}, h_{t,k}, \tau_{t,k}; \theta)\Delta,\ b^2(r_{t,k}, h_{t,k}, \tau_{t,k}; \theta)\Delta \right\}.$$

Using the notation defined in Section 10.2,

$$\begin{aligned}
l(r_{t,k}, h_{t,k}) &= c - \frac{1}{2}\log b^2(h_{t,k})\Delta - \frac{1}{2b^2(h_{t,k})\Delta} v_{t,k}^2 \\
l'(r_{t,k}, h_{t,k}) &= -\frac{b'(h_{t,k})}{b(h_{t,k})} + \frac{1}{\Delta b^2(h_{t,k})} \left\{ v_t a'_{h_t} \Delta + \frac{v_t^2 b'(h_{t,k})}{b(h_{t,k})} \right\} \\
l''(r_{t,k}, h_{t,k}) &= -\left[\frac{b(h_{t,k}) b''(h_{t,k}) - \{b'(h_{t,k})\}^2}{b^2(h_{t,k})} \right] + \frac{v_{t,k} a''(h_{t,k})}{b^2(h_{t,k})} - \frac{\{a'(h_{t,k})\}^2 \Delta}{b^2(h_{t,k})} \\
&\quad - \frac{4 v_{t,k} a'(h_{t,k}) b'(h_{t,k})}{b^3(h_{t,k})} + \frac{v_{t,k}^2 b''(h_{t,k})}{\Delta b^3(h_{t,k})} - \frac{3 v_{t,k}^2 \{b'(h_{t,k})\}^2}{b^4(h_{t,k})},
\end{aligned}$$

where $v_{t,k} = r_{t,k+1} - r_{t,k} - a(h_{t,k})\Delta$ and $c = -(1/2)\log 2\pi$. Specific forms for $l(r_{t,k}, h_{t,k})$ will depend on the $r_{t,k}$ process and will be considered later. The procedure for the single-move algorithm, (using the procedure in Section 5 to update $r_{t,k}$, for all t and $k = 1, \ldots, M$), is given below.

1. Initialise $\{r_{t,k}\}_{k=1}^{M}$, $\{h_{t,k}\}_{k=0}^{M+1}$ for all $t = 1 \ldots, T$ and initialise θ.
2. For each t, $(t = 1, \ldots, T)$, and for each k, $(k = 1, \ldots, M)$,
   - sample $r_{t,k}$ from $r_{t,k} | r_{t,k-1}, r_{t,k+1}, h_{t,k-1}, h_{t,k}; \theta$
   - sample $h_{t,k}$ from $h_{t,k} | h_{t,k-1}, h_{t,k+1}, r_{t,k}, r_{t,k+1}; \theta$.
3. Sample ξ^2 from $\xi^2 | r_{\text{obs}}, \left[\{r_{t,k}\}_{k=1}^{M}\right]_{t=1}^{T}, \left[\{h_{t,k}\}_{k=0}^{M+1}\right]_{t=1}^{T}; \kappa_2, \alpha$.
4. Sample κ_2 from $\kappa_2 | r_{\text{obs}}, \left[\{r_{t,k}\}_{k=1}^{M}\right]_{t=1}^{T}, \left[\{h_{t,k}\}_{k=0}^{M+1}\right]_{t=1}^{T}; \xi^2, \alpha$.
5. Sample α from $\alpha | r_{\text{obs}}, \left[\{r_{t,k}\}_{k=1}^{M}\right]_{t=1}^{T}, \left[\{h_{t,k}\}_{k=0}^{M+1}\right]_{t=1}^{T}; \kappa_2, \xi^2$.

The algorithm is repeated from step 2, for N sweeps. Since $l''(r_{t,k}, h_{t,k})$ is not, in general, negative semi-definite for the types of non-linear SDE's we consider, $l(r_{t,k}, h_{t,k})$ will not be log-concave in $h_{t,k}$. Instead, we could use a pseudo dominating Metropolis sampler, (see Tierney 1994, Shephard and Pitt 1997). Notice that this single-move metropolis sampler is a special case of the procedure outlined in 10.2. A quadratic approximation of the log density is used to obtain a Gaussian bounding density. The acceptance rates can be increased by iterating the second order expansion around the mode of the density, to give a Laplace approximation density as the proposal. The quadratic expansion is performed around $\hat{h}_{t,k}$ to give

$$l(r_{t,k}, h_{t,k}) = l(r_{t,k}, \mu_t) + (h_{t,k} - \mu_{t,k}) l'(r_{t,k}, \mu_{t,k}) + \frac{1}{2}(h_{t,k} - \mu_{t,k})^2 l''(r_{t,k}, \mu_{t,k})$$

so that on denoting $\log f(h_{t,k}|h_{t,k-1}, h_{t,k+1}, r_{t,k}, r_{t,k+1}; \theta) = \log f(h_{t,k})$

$$\begin{aligned}
\log f(h_{t,k}) &= c + \log f(h_{t,k}|h_{t,k-1}, h_{t,k+1}; \theta) + l(r_{t,k}, h_{t,k}) \\
&= c' - \frac{1}{2\Sigma_{t,k}}(h_{t,k} - \mu_t)^2 + l(r_{t,k}, h_{t,k}) \\
&\leq c' - \frac{1}{2\Sigma_{t,k}}(h_{t,k} - \mu_{t,k})^2 + l(r_{t,k}, \mu_{t,k}) + (h_{t,k} - \mu_{t,k})l'(r_{t,k}, h_{t,k}) \\
&\quad + \frac{1}{2}(h_{t,k} - \mu_{t,k})^2 l''(r_{t,k}, \mu_t) \\
&= c'' - \left\{ \frac{1 + l''(r_{t,k}, \mu_{t,k})\Sigma_{t,k}}{2\Sigma_{t,k}} \right\} \left[h_{t,k} - \left\{ \frac{\mu_{t,k} - l'(\mu_{t,k})\Sigma_{t,k} + \mu_{t,k} l''(\mu_{t,k})\Sigma_{t,k}}{1 + l''(\mu_{t,k})\Sigma_{t,k}} \right\} \right]^2 \\
&= \log g(h_{t,k}).
\end{aligned}$$

Draws z of $h_{t,k}|h_{t,k-1}, h_{t,k+1}, r_{t,k}, r_{t,k+1}; \theta$ are then suggested from $N(\mu^*_{t,k}, \Sigma^*_{t,k})$ where,

$$\begin{aligned}
\mu^*_{t,k} &= \left\{ \frac{\mu_{t,k} - l'(\mu_{t,k})\Sigma_{t,k} + \mu_{t,k} l''(\mu_{t,k})\Sigma_{t,k}}{\Sigma^*_{t,k}} \right\} \\
\Sigma^*_{t,k} &= 1 + l''(\mu_{t,k})\Sigma_{t,k}.
\end{aligned}$$

The Metropolis probability of accepting z is

$$\min \left(\frac{f\left\{z|h^{(j)}_{t,k-1}, h^{(j)}_{t,k+1}, r_{t,k}, r_{t,k+1}; \theta\right\} \min\left[f\left\{h^{(j)}_{t,k}|h^{(j)}_{t,k-1}, h^{(j)}_{t,k+1}, r_{t,k}, r_{t,k+1}\right\}, g\left\{r_{t,k}, h^{(j)}_{t,k}\right\}\right]}{f\left\{h^{(j)}_{t,k}|h^{(j)}_{t,k-1}, h^{(j)}_{t,k+1}, r_{t,k}, r_{t,k}\right\} \min\left[f\left\{z|h^{(j)}_{t,k-1}, h^{(j)}_{t,k+1}, r_{t,k}, r_{t,k}\right\}, g(r_{t,k}, z)\right]}, 1 \right).$$

The danger here, is that if the expansion is performed around the wrong point, the variance may be negative and so it may be simpler to consider a first order Taylor expansion.

10.3 Examples of continuous-time processes with latent volatility

10.3.1 Geometric Brownian motion

If the coefficients in the drift and volatility function only depend on time, so that

$$dr(t) = a(t)r(t)dt + b(t)r(t)dW(t)$$

then

$$r(t) = r(0)\exp\left[\int_0^t \left\{a(s) - \frac{b^2(s)}{2}\right\} ds + \int_0^t b(s)dW(s)\right].$$

see Kloeden and Platen, (Ch. 4, 1992). Suppose

$$dr(t) = \mu r(t)dt + \sigma(t)r(t)dW(t)$$

then on applying Itô's rule to the transformation $\log r(t)$, we obtain

$$d\log r(t) = \left\{\mu - \frac{\sigma^2(t)}{2}\right\} dt + \sigma(t) dW(t). \tag{10.9}$$

If $\sigma(t) = \sigma$, independent of t, then the strong solution of the process is given by

$$\log r(t+1) = \log r(t) + \left(\mu - \frac{\sigma^2}{2}\right) \Delta_t^\dagger + \sigma \int_t^{t+1} dW(s)$$

where $\Delta_t^\dagger$ is the time step between t and $t+1$. This gives the conditional transition density of $\log r(t+1) | \log r(t)$ as

$$\log r(t+1) | \log r(t) \sim N\left\{\log r(t) + \left(\mu - \frac{\sigma^2}{2}\right) \Delta_t^\dagger,\ \sigma^2 \Delta_t^\dagger\right\}. \tag{10.10}$$

μ is the expected return earned by an investor in a short period of time. Typically, μ is annualised and expressed as a proportion, so that its value depends on the risk of a stock's return and the interest rate; increased risk entails a higher expected return. The typical expected return of 8% on a stock, (above the return on the risk-free investment), is approximately 16% per year. This corresponds to $\mu = 0.16$ in the model.[10] Typical values of σ are $[\sqrt{0.2}, \sqrt{0.4}]$, see Hull, (Hull 1989, p. 219). Note that the Euler scheme applied to equation (10.9), when $a(t)$ and $b(t)$ are homogeneous constant coefficients also gives us the strong solution. One could alternatively re-parameterize (μ, σ^2) as (γ, σ^2), where $\gamma = \mu - \sigma^2/2$. If $\sigma(t)$ varies with t, then the strong solution is given by

$$\log r(t+1) = \log r(t) + \mu \Delta_t^\dagger - \frac{1}{2} \int_t^{t+1} \sigma^2(s) ds + \int_t^{t+1} \sigma(s) dW(s)$$

which now depends on the integrated volatility. If we discretise the process using the Euler scheme and let $\Delta_t = \Delta$ for all t, then

$$\log r_{t+1} = \log r_t + \left\{\mu - \frac{\exp(h_t)}{2}\right\} \Delta + \exp(\frac{h_t}{2}) \sqrt{\Delta} \epsilon_{t+1}$$

where $\epsilon_t \sim NID(0,1)$. This gives the conditional specification

$$\log r_{t+1} | \log r_t \overset{\text{approx}}{\sim} N\left[\log r_t + \left\{\mu - \frac{\exp(h_t)}{2}\right\} \Delta, \exp(h_t)\Delta\right]. \tag{10.11}$$

The log-likelihood for the data and auxiliary points, based on (10.11), is given by

$$\log f(\log r, \log r^*; \theta) = -\frac{(n-2)}{2} \log 2\pi\Delta - \frac{1}{2} \sum_{t=1}^{n-1} h_t - \frac{1}{2\Delta} \sum_{t=1}^{n-1} \frac{\left\{\log r_{t+1} - \log r_t - \mu\Delta + \frac{\exp(h_t)\Delta}{2}\right\}^2}{\exp(h_t)}.$$

[10] Since $dr/r \sim N(\mu\Delta, \sigma^2\Delta)$, so that the change is $\mu/365$, for a change Δ in the stock prices.

The posterior of the drift for the GBM and derivatives for the MH algorithm

For the posterior density of μ, suppose we consider the conjugate prior density

$$f(\mu) \propto \exp\left\{-\frac{1}{2}(\mu-\mu_0)'\Gamma(\mu-\mu_0)\right\}$$

implying that $\mu \sim N(\mu_0, \Gamma^{-1})$. Suppose that we express the log-likelihood of the observations and the auxiliary points as

$$\log f(\mu|\log r, \log r^*) \propto c - \frac{1}{2\Delta}(z - X\mu\Delta)'(z - X\mu\Delta),$$

where

$$z = \begin{bmatrix} \frac{\log r_2 - \log r_1 + \frac{\exp(h_1)\Delta}{2}}{\exp\left(\frac{h_1}{2}\right)} \\ \vdots \\ \frac{\log r_n - \log r_{n-1} + \frac{\exp(h_{n-1})\Delta}{2}}{\exp\left(\frac{h_{n-1}}{2}\right)} \end{bmatrix} \quad \text{and} \quad X = \begin{bmatrix} \frac{1}{\exp(h_1/2)} \\ \vdots \\ \frac{1}{\exp(h_{n-1}/2)} \end{bmatrix}.$$

Then the log posterior can be expressed as

$$\begin{aligned} \log f(\mu|\log r, \log r^*) &= c - \frac{1}{2\Delta}(z - X\mu\Delta)'(z - X\mu\Delta) - \frac{1}{2}(\mu-\mu_0)'\Gamma^{-1}(\mu-\mu_0) \\ &= c' - \frac{1}{2}(v - V^{-1}u)'V(v - V^{-1}u) \end{aligned}$$

where

$$\begin{aligned} u &= (X'z + \Gamma\mu_0) \\ V &= (X'X\Delta + \Gamma) \\ v &= \mu. \end{aligned}$$

Hence $\log f(\mu|\log r, \log r^*) \sim N(V^{-1}u,\, V^{-1})$.

For the MH algorithm defined in Section 5, we define the following

$$\begin{aligned} a_t &= \mu - \frac{\exp(h_t)}{2} \\ a_t' &= a_t'' = 0 \\ b_t &= \exp\left(\frac{h_t}{2}\right) \\ b_t' &= g_t' = g_t'' = 0 \\ g_t &= \frac{1}{\exp(h_t)} \end{aligned}$$

and an initial value of μ can be found from

$$\begin{aligned}\frac{\partial \log f}{\partial \mu} &= \sum_{t=1}^{n-1}\left\{\frac{\log r_{t+1}-\log r_t-\mu\Delta+\frac{\exp(h_t)\Delta}{2}}{\exp(h_t)}\right\}\\ \Rightarrow \hat{\mu} &= \frac{1}{\Delta\sum_{t=1}^{n-1}\exp(-h_t/2)}\sum_{t=1}^{n-1}\left\{\log r_{t+1}-\log r_t+\frac{\exp(h_t)\Delta}{2}\right\}.\end{aligned}$$

For the MH algorithm to update the latent process, we define

$$\begin{aligned}a(h_t) &= \mu-\frac{\exp(h_t)}{2}\\ a'(h_t) &= a''(h_t)=-\frac{\exp(h_t)}{2}\\ b(h_t) &= \exp(h_t/2)\\ b'(h_t) &= \frac{1}{2}\exp(h_t/2)\\ b''(h_t) &= \frac{1}{4}\exp(h_t/2)\end{aligned}$$

so that $h_{t,k}$ is drawn from $h_{t,k}|h_{t,k-1},h_{t,k+1},r_{t,k},r_{t,k+1};\theta$. As the second derivative is not negative definite,[11] the variance of the proposal depends only on the second derivative of $l(h_{t,k})$ and not $l''(r_{t,k},h_{t,k})$. We could use the single-move psuedo-dominating Metropolis sampler defined in 10.2.2, changing the mean, but not the variance. The latter gives more flexibility as it is not reduced by conditioning. If we choose $\exp(\hat{h_{t,k}})=(\log r_{t,k+1}-\log r_{t,k}-\mu\Delta)(2/\Delta)$, then $l''(r_{t,k},h_{t,k})<0$, since

$$l(r_{t,k},\hat{h}_{t,k})=d-\frac{\hat{h}_{t,k}}{2}-\frac{\left\{\log r_{t,k+1}-\log r_{t,k}-\mu\Delta+\frac{\exp(\hat{h}_{t,k})\Delta}{2}\right\}^2}{2\Delta\exp(\hat{h}_{t,k})}.$$

10.3.2 Generalised CIR model

Many models for interest rates are formulated using continuous time processes. Vasicek (1977) specifies a stationary model with mean reversion. The Cox, Ingersoll, and Ross (1985) model has a volatility specification that depends on the square-root of the interest rate. A *level effect* is incorporated into the model, that is, the volatility increases with the level of the interest rate. Papers by Dyvbig (1989), Longstaff and Schwartz (1992), Brenner, Harjes, and Kroner (1994), Koedijk, Nissen, Schotman, and Wolff (1994) and Gallant and Tauchen (1997a) all suggest extending the basic model to incorporate a stochastic volatility component. This can be captured by the canonical representation of the generalised

[11] Notice that the log-likelihood, $l(r_{t,k},h_{t,k})$ and its derivatives are given by

$$\begin{aligned}l(r_{t,k},h_{t,k}) &= c-\frac{1}{2}\log\Delta-\frac{h_{t,k}}{2}-\frac{v_{t,k}^2}{2\Delta\exp(h_{t,k})}\\ l'(r_{t,k},h_{t,k}) &= -\frac{1}{2}+\frac{v_{t,k}^2}{2\Delta\exp(h_{t,k})}-\frac{v_{t,k}}{2}\\ l''(r_{t,k},h_{t,k}) &= \frac{v_{t,k}}{2}-\exp(h_{t,k})\frac{\Delta}{4}+\frac{v_{t,k}^2}{4\Delta\exp(h_{t,k})}(4-3\Delta).\end{aligned}$$

CIR model as

$$dr(t) = \kappa_1\{\mu - r(t)\}dt + \sigma(t)r(t)^{\gamma}dW(t) \tag{10.12}$$

where $\gamma > 0$ and κ_1 and μ are non-negative. The parameter κ_1 measures the speed of reversion, while $r(t)$ mean reverts towards μ, the level. This model has been considered by Andersen and Lund (1997). They obtain a semi-nonparametric estimate of $r(t)$ given information up to $t-1$ based on QML methods. Estimates of the structural parameters are found by constructing an auxiliary model with the same characteristics as the diffusion. This can be tested by examining whether the QML score is close to zero. Effectively, Monte Carlo integration is used to calculate the expected score for each candidate θ.

If we impose the transformation $y(t) = \log r(t)$, then

$$dy(t) = \left[\frac{\kappa_1\mu}{\exp\{y(t)\}} - \kappa_1 - \frac{\exp\{h(t)\}}{2}\exp\{2(\gamma-1)y(t)\}\right]dt + \exp\{h(t)/2\}\exp\{(\gamma-1)y(t)\}dW(t)$$

so that on applying the Euler scheme, we obtain

$$y_{t+1} = y_t + \left[\frac{\kappa_1\mu}{\exp(y_t)} - \kappa_1 - \frac{\exp(h_t)}{2}\exp\{2(\gamma-1)y_t\}\right]\Delta + \exp(h_t/2)\exp\{(\gamma-1)y_t\}\sqrt{\Delta}\epsilon_{t+1}$$

where $\epsilon_t \sim N(0,1)$. With respect to the return,

$$d\log r_t = \left\{\frac{\kappa_1\mu}{r_t} - \kappa_1 - \frac{\exp(h_t)}{2}r_t^{2(\gamma-1)}\right\}dt + \exp(h_t/2)r_t^{(\gamma-1)}dW_t$$

with the discretised process specified as

$$\log r_{t+1} = \log r_t + \left\{\frac{\kappa_1\mu}{r_t} - \kappa_1 - \frac{\exp(h_t)}{2}r_t^{2(\gamma-1)}\right\}\Delta + \exp(h_t/2)r_t^{(\gamma-1)}\sqrt{\Delta}\epsilon_{t+1}.$$

Initial values for the parameters: Initial values can be found by observing that

$$\log f(\theta|\log r) = c - \frac{1}{2\Delta^{\dagger}}\sum_{t=1}^{T-1}\frac{\left[\log r(t+1) - \log r(t) - \frac{\exp\{h(t)\}}{2}r(t)^{2\gamma-2}\Delta^{\dagger} - \frac{\rho\Delta^{\dagger}}{r(t)} + \kappa_1\Delta^{\dagger}\right]^2}{\exp\{h(t)\}r(t)^{2(\gamma-1)}},$$

where $\rho = \kappa_1\mu$. By defining

$$\begin{aligned} x(t) &= \log r(t+1) - \log r(t) - \frac{\exp\{h(t)\}}{2}r(t)^{2\gamma-2}\Delta^{\dagger} \\ v(t)^2 &= \exp\{h(t)\}r(t)^{2(\gamma-1)} \end{aligned}$$

the log-likelihood becomes

$$\log f(\theta|\log r) = c - \frac{1}{2\Delta^{\dagger}}\sum_{t=1}^{T-1}\left\{\frac{x(t) - \frac{\rho\Delta^{\dagger}}{r(t)} + \kappa_1\Delta^{\dagger}}{v(t)}\right\}^2$$

so that the first order conditions are

$$\begin{aligned}
\frac{\partial \log f}{\partial \rho} &= \sum_{t=1}^{T-1} \frac{\left\{x(t) - \frac{\rho\Delta^\dagger}{r(t)} + \kappa_1\Delta^\dagger\right\}}{r(t)v(t)^2} \\
&\Rightarrow \quad \hat{\rho} = \frac{1}{\Delta^\dagger}\left[\frac{\sum_{t=1}^{T-1}\left\{\frac{x(t)+\kappa_1\Delta^\dagger}{r(t)v(t)^2}\right\}}{\sum_{t=1}^{T-1} r(t)^{-2}v(t)^{-2}}\right] \\
\frac{\partial \log f}{\partial \kappa_1} &= -\sum_{t=1}^{T-1} \frac{\left\{x(t) - \frac{\rho\Delta^\dagger}{r(t)} + \kappa_1\Delta^\dagger\right\}}{v(t)^2} \\
&\Rightarrow \quad \hat{\kappa_1} = \frac{1}{\Delta^\dagger}\left[\frac{\sum_{t=1}^{T-1}\left\{\frac{x(t)-\rho\Delta^\dagger/r(t)}{v(t)^2}\right\}}{\sum_{t=1}^{T-1} v(t)^{-2}}\right].
\end{aligned}$$

Given one of the parameters and initial values for the latent volatility process, the other is automatically set. Initial values for $h(t)$ can be set to $\log r^2(t)$.

The posterior of the parameters for the Generalised CIR and the derivatives for the MH algorithm To obtain the posterior distribution of the parameters, we know that

$$y_{t+1}|y_t \overset{\text{approx}}{\sim} N\left[y_t + \frac{\kappa_1\mu\Delta}{\exp(y_t)} - \kappa_1\Delta - \frac{\exp(h_t)}{2}\exp\left\{2(\gamma-1)y_t\right\}\Delta,\ \exp(h_t)\exp\left\{(\gamma-1)y_t\right\}\Delta\right]$$

so that the log-likelihood for the data and auxiliary points is

$$\begin{aligned}
\log f(Y, Y^*; \theta) &= -\frac{(n-2)}{2}\log 2\pi\Delta - \frac{1}{2}\sum_{t=1}^{n-1} h_t - \frac{(\gamma-1)}{2}\sum_{t=1}^{n-1} y_t \\
&\quad -\frac{1}{2\Delta}\sum_{t=1}^{n-1}\left[\frac{y_{t+1} - y_t - \frac{\kappa_1\mu\Delta}{\exp(y_t)} - \kappa_1\Delta - \frac{\exp(h_t)}{2}\exp\left\{2(\gamma-1)y_t\right\}\Delta}{\exp\left(\frac{h_t}{2}\right)\exp\left\{\frac{(\gamma-1)y_t}{2}\right\}}\right]^2.
\end{aligned}$$

Letting $\beta = \kappa_1\mu$ and

$$\begin{aligned}
z &= \begin{bmatrix} \frac{y_2 - y_1 - \{\exp(h_1)/2\}\exp\{2(\gamma-1)y_1\}\Delta}{\exp(h_1/2)\exp\{(\gamma-1)y_1/2\}} \\ \vdots \\ \frac{y_n - y_{n-1} - \{\exp(h_{n-1})/2\}\exp\{2(\gamma-1)y_{n-1}\}\Delta}{\exp(h_{n-1}/2)\exp\{(\gamma-1)y_{n-1}/2\}} \end{bmatrix} \\
X &= \begin{bmatrix} \frac{1}{\exp(h_1/2)\exp\{(\gamma-1)y_1/2\}} & \frac{1}{\exp(y_1)\exp(h_1/2)\exp\{(\gamma-1)y_1/2\}} \\ \vdots & \vdots \\ \frac{1}{\exp(h_{n-1}/2)\exp\{(\gamma-1)y_{n-1}/2\}} & \frac{1}{\exp(y_{n-1})\exp(h_{n-1}/2)\exp\{(\gamma-1)y_{n-1}/2\}} \end{bmatrix}.
\end{aligned}$$

On setting a conjugate prior for $\psi = (\beta, \kappa_1)'$ of

$$f(\psi) = \exp\left\{-\frac{1}{2}(\psi - \psi_0)'\Gamma(\psi - \psi_0)\right\}$$

we can express the log posterior as

$$\begin{aligned}\log f(\beta,\kappa_1|Y,Y^*;\gamma) &= c-\frac{1}{2\Delta}(z-X\psi\Delta)'(z-X\psi\Delta)-\frac{1}{2}(\psi-\psi_0)'\Gamma^{-1}(\psi-\psi_0)\\ &= c'-\frac{1}{2}(\psi-V^{-1}u)'V(\psi-V^{-1}u),\end{aligned}$$

where

$$\begin{aligned}u &= X'z+\Gamma\psi_0\\ V &= X'X\Delta+\Gamma\end{aligned}$$

so that $\log f(\psi|Y,Y^*)\sim N(V^{-1}u,V^{-1})$. The parameters μ can be retrieved through β/κ_1. For the MH algorithm, we define

$$\begin{aligned}a(t) &= \frac{\beta}{\exp(y_t)}-\kappa_1-\frac{\exp(h_t)}{2}\exp\{2(\gamma-1)y_t\}\\ a(t)' &= -\frac{\beta}{\exp(y_t)}-(\gamma-1)\exp(h_t)\exp\{2(\gamma-1)y_t\}\\ a(t)'' &= \frac{\beta}{\exp(y_t)}-2(\gamma-1)^2\exp(h_t)\exp\{2(\gamma-1)y_t\}\\ b(t) &= \exp\left(\frac{h_t}{2}\right)\exp\left\{\frac{(\gamma-1)}{2}y_t\right\}\\ b(t)' &= \left(\frac{\gamma-1}{2}\right)\exp\left(\frac{h_t}{2}\right)\exp\left\{\frac{(\gamma-1)}{2}y_t\right\}\\ g(t) &= \exp(-h_t)\exp\{-(\gamma-1)y_t\}\\ g(t)' &= -(\gamma-1)\exp(h_t)\exp\{-(\gamma-1)y_t\}\\ g(t)'' &= (\gamma-1)^2\exp(h_t)\exp\{-(\gamma-1)y_t\}.\end{aligned}$$

For the Metropolis sampler for the latent volatilities,[12]

$$\begin{aligned}
a_{h_t} &= \frac{\beta}{\exp(y_t)} - \kappa_1 - \frac{\exp(h_t)}{2}\exp\{2(\gamma-1)y_t\} \\
a'_{h_t} &= a''_{h_t} = -\frac{\exp(h_t)}{2}\exp\{2(\gamma-1)y_t\} \\
b_{h_t} &= \exp\left(\frac{h_t}{2}\right)\exp\left\{\frac{(\gamma-1)}{2}y_t\right\} \\
b'_{h_t} &= \frac{b_{h_t}}{2} \\
b''_{h_t} &= \frac{b_{h_t}}{4}.
\end{aligned}$$

10.4 Conclusion and further research

The methods developed in Section 5 have been extended to consider a bivariate system of SDE's, where one of the processes is unobserved. In particular, the geometric Brownian motion and Generalised CIR models, both incorporating a latent stochastic volatility structure were considered. The techniques can be illustrated in the context of a bivariate stochastic volatility model, extended to incorporate a Markov switching structure, which has not been considered in the thesis, but is part of current research work.

The general formulation is presented next, in Chapter 11, allowing a more complex study of SDE's which may arise in practical situations.

10.5 Appendix

10.5.1 Obtaining the proposal for the bivariate stochastic volatility model

In Section 10.2, a second order Taylor expansion of $l(w)$ is taken around $\hat{w}$, so that

$$l(w) \approx c' - \frac{1}{2}(x - V_t^{-1}u_t)'V_t(x - V_t^{-1}u_t) \approx \log N(\tilde{\mu}, \tilde{\Sigma}),$$

where $x = w - \hat{w}$, $\tilde{\mu} = \hat{w} + V_t^{-1}u_t$, $\tilde{\Sigma} = V_t^{-1}$ and

$$\begin{aligned}
u_t &= \frac{\partial l(\hat{w})}{\partial w} = \{u_{t,j}\} \text{ for } j = 1, \ldots, m \\
V_t &= -\frac{\partial^2 l(\hat{w})}{\partial w \partial w'} = \{V_{ij}^t\} \text{ for } i, j = 1 \ldots m.
\end{aligned}$$

[12] If we define $v_t = y_{t+1} - y_t - \frac{\rho\Delta}{\exp(y_t)} + \kappa_1\Delta + \frac{\exp(h_t)}{2}\exp\{2(\gamma-1)y_t\}\Delta$, then

$$\begin{aligned}
l(h_t) &= c - \frac{1}{2}\exp(h_t)\exp\{(\gamma-1)y_t\}\Delta - \frac{v_t^2}{2\Delta\exp(h_t)\exp\{(\gamma-1)y_t\}} \\
l'(h_t) &= -\frac{1}{2} + \frac{v_t^2}{2\Delta\exp(h_t/2)\exp\{(\gamma-1)y_t/2\}} - \frac{v_t}{2}\exp(h_t/2)\exp\{(3/2)(\gamma-1)y_t\} \\
l''(h_t) &= -\frac{\Delta}{2}\exp(h_t)\exp\{(\gamma-1)y_t\} + \frac{v_t^2(1-3\Delta)}{4\Delta\exp(h_t)\exp\{(\gamma-1)y_t\}}.
\end{aligned}$$

$$V^t_{j+1,j+1} = \begin{cases} -\frac{\partial^2 l^{(1)}_{t,j}}{(\partial w_{j+1})^2} - \frac{\partial^2 l^{(1)}_{t,j+1}}{(\partial w_{j+1})^2} - \frac{\partial^2 l^{(2)}_{t,j+1}}{(\partial w_{j+1})^2} & \text{if } \frac{\partial^2 l(\hat{w})}{\partial w \partial w'} \text{ is negative definite} \\ -\frac{\partial^2 l^{(1)}_{t,j}}{(\partial w_{j+1})^2} - \frac{\partial^2 l^{(1)}_{t,j+1}}{(\partial w_{j+1})^2} & \text{otherwise} \end{cases}$$

$$\text{and } V^t_{ij} = \frac{\partial^2 l^{(1)}_{t,j+1}}{\partial w_{j+1} \partial w_{j+2}} \quad \text{for } i \neq j,\ j = 0, \ldots, m-1.$$

We differentiate between the notation $a_{t,j}$, $b^2_{t,j}$, as defined in Section 5, (to denote the drift and volatility as functions of $r_{t,k+j}$), and the notation $a(h_{t,j})$, $b^2(h_{t,j})$, (to denote the drift and volatility as functions of $h_{t,k+j}$), so that,

$$\begin{aligned} a(h_{t,j}) &= a(r_{t,k+j}, h_{t,k+j}) = a(r_{t,k+j}, w_{j+1}) \\ b(h_{t,j}) &= b(r_{t,k+j}, h_{t,k+j}) = b(r_{t,k+j}, w_{j+1}) \\ e_{t,j} &= e(h_{t,k+j}) = e(w_{j+1}) \\ n_{t,j} &= n(h_{t,k+j}) = n(w_{j+1}) \\ o_{t,j} &= h_{t,k+j+1} - h_{t,k+j} - e_{t,j}\Delta = w_{j+2} - w_{j+1} - e_{t,j}\Delta \\ p_{t,j} &= 1 + e'_{t,j}\Delta \end{aligned}$$

for $j = -1, 0, \ldots, m-1$. Let $e'_{t,j}$ and $e''_{t,j}$ denote the first and second derivatives of $e(w_{j+1})$ with respect to w_{j+1} and define $n'_{t,j}$, $n''_{t,j}$ siniilarly. Consider a given element, w_{j+1}, (for $j = 0, \ldots m-1$), of the block. The only terms in $l(w)$ involving w_{j+1} are

$$\begin{aligned} l^{(1)}_{t,j} &= -\frac{1}{2}\log(n^2_{t,j-1}\Delta) - \frac{1}{2n^2_{t,j-1}\Delta}(w_{j+1} - w_j - e_{t,j-1}\Delta)^2 \\ l^{(1)}_{t,j+1} &= -\frac{1}{2}\log(n^2_{t,j}\Delta) - \frac{1}{2n^2_{t,j}\Delta}(w_{j+2} - w_{j+1} - e_{t,j}\Delta)^2 \\ l^{(2)}_{t,j+1} &= -\frac{1}{2}\log(b(h_{t,j})^2\Delta) - \frac{1}{2b(h_{t,j})^2\Delta}(r_{j+2} - r_{j+1} - a(h_{t,j})\Delta)^2. \end{aligned}$$

The first derivative of $l(w)$ with respect to w_{j+1} can now be defined as follows:

$$\begin{aligned} \frac{\partial l(w)}{\partial w_{j+1}} &= \frac{\partial l^{(1)}_{t,j}}{\partial w_{j+1}} + \frac{\partial l^{(1)}_{t,j+1}}{\partial w_{j+1}} + \frac{\partial l^{(2)}_{t,j+1}}{\partial w_{j+1}} \\ &= -\frac{1}{n^2_{t,j-1}\Delta}(w_{j+1} - w_j - e_{t,j-1}\Delta) \\ &\quad -\frac{n'_{t,j}}{n_{t,j}} + \frac{1}{2\Delta}\left(\frac{2n_{t,j}o_{t,j}p_{t,j} + 2o^2_{t,j}n'_{t,j}}{n^3_{t,j}}\right) - \frac{b'_{t,j}}{b(h_{t,j})} \\ &\quad + \frac{1}{2\Delta}\left\{\frac{2(r_{j+2} - r_{j+1} - a(h_{t,j})\Delta)a(h_{t,j})'b(h_{t,j})\Delta + (r_{j+2} - r_{j+1} - a(h_{t,j})\Delta)^2 b(h_{t,j})'}{b(h_{t,j})^3}\right\} \end{aligned}$$

where $a(h_{t,j})' = \partial a(h_{t,j})/\partial w_{j+1}$ and $b(h_{t,j})' = \partial b(h_{t,j})/\partial w_{j+1}$. The second derivatives are given by

$$\begin{aligned}
\frac{\partial^2 l}{\partial w_{j+1}^2} &= -\frac{1}{n_{t,j-1}^2\Delta} - \left\{ \frac{n_{t,j}n_{t,j}'' - \left(n_{t,j}'\right)^2}{n_{t,j}^2} \right\} \\
&+ \frac{1}{\Delta}\left(\frac{n_{t,j}'o_{t,j}p_{t,j} - n_{t,j}p_{t,j}^2 + n_{t,j}o_{t,j}e_{t,j}''\Delta + o_{t,j}^2 n_{t,j}'' - 2o_{t,j}n_{t,j}'p_{t,j}}{n_{t,j}^3} \right) \\
&- \frac{3n_{t,j}'}{\Delta}\left(\frac{n_{t,j}o_{t,j}p_{t,j} + o_{t,j}^2 n_{t,j}'}{n_{t,j}^4} \right) - \left\{ \frac{b(h_{t,j})b(h_{t,j})'' - \left(b(h_{t,j})'\right)^2}{b(h_{t,j})^2} \right\} \\
&+ \left\{ \frac{\left(a(h_{t,j})'\right)^2 b(h_{t,j})\Delta + (r_{j+2} - r_{j+1} - a(h_{t,j})\Delta)a(h_{t,j})''b(h_{t,j})}{b(h_{t,j})^3} \right\} \\
&+ \left\{ \frac{(r_{j+2} - r_{j+1} - a(h_{t,j})\Delta)a(h_{t,j})'b(h_{t,j})'}{b(h_{t,j})^3} \right\} \\
&+ \frac{1}{2\Delta}\left\{ \frac{-2(r_{j+2} - r_{j+1} - a(h_{t,j})\Delta)a(h_{t,j})'b(h_{t,j})'\Delta + (r_{j+2} - r_{j+1} - a(h_{t,j})\Delta)^2 b(h_{t,j})''}{b(h_{t,j})^3} \right\} \\
&- \frac{3b(h_{t,j})'(r_{j+2} - r_{j+1} - a(h_{t,j})\Delta)}{b(h_{t,j})^4}\left\{ 2a(h_{t,j})'b(h_{t,j})\Delta + (r_{j+2} - r_{j+1} - a(h_{t,j})\Delta)b(h_{t,j})' \right\}.
\end{aligned}$$

The only term involving w_{j+1} and w_{j+2} is $l_{t,j+1}^{(1)}$, implying that

$$\frac{\partial^2 l(w)}{\partial w_{j+1}\partial w_{j+2}} = \frac{\partial^2 l_{t,j+1}^{(1)}}{\partial w_{j+1}\partial w_{j+2}} = \frac{1}{\Delta}\left(\frac{n_{t,j}p_{t,j} + 2o_{t,j}n_{t,j}'}{n_{t,j}^3} \right),$$

since $\partial o_{t,j}/\partial w_{j+2} = 1$. In summary, the proposal density $q(w|w_0, w_{m+1}, r; \theta)$ obtained from the second order Taylor expansion of $l(w)$ is Gaussian with mean

$$\tilde{\mu} = \hat{w} + \tilde{\Sigma}u_t$$

and variance matrix

$$\tilde{\Sigma} = V_t^{-1}$$

where u_t is an $m \times 1$ vector with elements

$$u_{t,j+1} = \frac{\partial l(w)}{\partial w_{j+1}} \quad \text{for } j = 0, \ldots, m-1,$$

and V_t is an $m \times m$ matrix with elements

$$\begin{aligned} V^t_{j+1,j+1} &= -\frac{\partial^2 l(w)}{\partial w^2_{j+1}} \\ V^t_{j+1,j+2} &= V^t_{j+2,j+1} = -\frac{\partial^2 l(w)}{\partial w_{j+1}\partial w_{j+2}} \end{aligned}$$

and $V^t_{j+1,j+s} = 0$ for all j, when $s > 2$. Both u_t and V_t are evaluated at $w = \hat{w}$.

Derivatives for the end points in a latent volatility structure

The first and second derivatives with respect to $h_{1,0}$ in 10.2.1 are given by

$$\begin{aligned} \frac{\partial l^{(1)}_{1,1}}{\partial h_{1,0}} &= -\frac{1}{n_{1,0}} + \frac{1}{\Delta} o_{1,0} p_{1,0} + \frac{o^2_{1,0} n'_{1,0}}{n_{1,0}\Delta} \\ \frac{\partial l^{(2)}_{1,1}}{\partial h_{1,0}} &= -\frac{b(h_{1,0})'}{b(h_{1,0})} + (r_{1,1} - r_{1,0} - a(h_{1,0})\Delta) a(h_{1,0})' + \frac{(r_{1,1} - r_{1,0} - a(h_{1,0})\Delta)^2 b(h_{1,0})'}{b(h_{1,0})\Delta} \\ \frac{\partial^2 l^{(1)}_{1,1}}{(\partial h_{1,0})^2} &= \frac{n'_{1,0}}{n^2_{1,0}} + \frac{1}{\Delta}(o'_{1,0} p_{1,0} + o_{1,0} p'_{1,0}) + \left\{ \frac{(2 o_{1,0} o'_{1,0} n'_{1,0} + o^2_{1,0} n''_{1,0}) n_{1,0}\Delta - o^2_{1,0}\Delta (n'_{1,0})^2}{n^2_{1,0}\Delta^2} \right\} \\ \frac{\partial^2 l^{(2)}_{1,1}}{(\partial h_{1,0})^2} &= -\left\{ \frac{b(h_{1,0}) b(h_{1,0})'' - (b(h_{1,0})')^2}{b(h_{1,0})^2} \right\} + (r_{1,1} - r_{1,0} - a(h_{1,0})\Delta) a(h_{1,0})'' - (a(h_{1,0})')^2 \Delta \\ &\quad + \frac{1}{\Delta}\left\{ \frac{-2a(h_{1,0})' b(h_{1,0}) b(h_{1,0})' (r_{1,1} - r_{1,0} - a(h_{1,0})\Delta)\Delta}{b(h_{1,0})^2} \right\} \\ &\quad + \left\{ \frac{b(h_{1,0})(r_{1,1} - r_{1,0} - a(h_{1,0})\Delta)^2 b(h_{1,0})''}{b(h_{1,0})^2} \right\} \\ &\quad - \frac{1}{\Delta}\left\{ \frac{(r_{1,1} - r_{1,0} - a(h_{1,0})\Delta)^2 (b(h_{1,0})')^2}{b(h_{1,0})^2} \right\}. \end{aligned}$$

For the remaining endpoint h_T, observe that,

$$\begin{aligned} l^{(1)}_{T-1,M} &= c - \frac{1}{2}\log(n^2_{T-1,M}\Delta) - \frac{1}{2n^2_{T-1,M}\Delta}(h_T - h_{T-1,M} - e_{T-1,M}\Delta)^2, \\ \text{so that } u_T &= \frac{\partial l^{(1)}_{T-1,M}}{\partial h_T} = -\frac{1}{n^2_{T-1,M}\Delta}(h_T - h_{T-1,M} - e_{T-1,M}\Delta) \\ \text{and } V_T &= -\frac{\partial^2 l^{(1)}_{T-1,M}}{(\partial h_T)^2} = \frac{1}{n^2_{T-1,M}\Delta}. \end{aligned}$$

These are evaluated at $\hat{h}_T$.

CHAPTER 11

MULTIVARIATE STOCHASTIC DIFFERENTIAL EQUATIONS

The univariate analysis previously outlined is now extended to cover multivariate diffusion processes. The Chapter is organized as follows: the general model is outlined in Section 11.1, covering discretization and augmentation issues in Section 11.2. Section 11.3 presents the simulation of the auxiliary variables using a tuned Metropolis-Hastings (MH) step. We will focus on presenting an estimation procedure which is general and deals primarily with non-linear SDE's. The method may be simplified in situations where the SDE's are partially solvable, or where $a^j(X_{t,k}, \tau_{t,k}; \theta)$ and $b(X_{t,k}, \tau_{t,k}; \theta)$ depend linearly on, (or are independent of) $X_{t,k}[\backslash j]$. The MCMC algorithm is completed by a description of the methods required to sample the parameters of the model, conditioned on the augmented data. The Chapter concludes with Section 11.4.

11.1 MODELS

Consider an Itô stochastic process that satisfies a stochastic differential equation (SDE) of the form

$$dX(t) = a\{X(t), t; \theta\}dt + b\{X(t), t; \theta\}dW(t), \tag{11.1}$$

where $X(t) \in \mathbb{R}^p$, $a\{X(t), t; \theta\} : \mathbb{R}^p \times [0, \infty) \times \mathbb{R}^q \longmapsto \mathbb{R}^p$ and $b\{X(t), t; \theta\} : \mathbb{R}^p \times [0, \infty) \times \mathbb{R}^q \to \mathbb{R}^{p \times d}$ are the non-anticipative drift and volatility functions, respectively, depending on $X(t)$, time t, and an unknown parameter vector $\theta \subseteq \Theta \in \mathbb{R}^q$, and $W(t)$ is a vector in $\mathbb{R}^d$, which is a homogeneous Lévy process. A Lévy process is a stochastic process, with independent and stationary increments. For general properties of Lévy processes as Markov processes, see for example, Lévy (1937) Jacod and Shiryayev (1987); a brief overview can also be found in Bertoin (1996, Ch. 1). The analysis will be considered for the case where $W(t)$ represents the increments of a standard Wiener process and $p \geq d$. Assume that the conditions under which the SDE can be solved for a diffusion $X(t)$ are satisfied - for conditions on the existence and uniqueness of a solution to the SDE, (where 11.1 is viewed as a Wiener functional), see Watanabe (1984, Ch. 2.1, pp. 65-67) and Strook and Varadhan, 1979, Ch. 5.1). Essentially, if the drift and volatility functions are sufficiently smooth, a unique solution exists. If the functions have bounded derivatives, a global solution exists.

For a reference on stochastic calculus for multi-dimensional diffusions, see Strook and Varadhan, (1979, Ch. 4; for a simpler exposition on stochastic calculus, see Madsen, Nielsen, and Baadsgaard, (1998, Ch. 7) and for an introduction to stochastic calculus with application to finance, see Lamberton and Lapeyre

(1996). For a good exposition on Markov and diffusion processes, see Arnold, (1974, Ch. 2) and for Brownian motion, see Karatzas and Shreve (1991) and Borodin and Salminen (1996). An alternative definition of (11.1) is given by the Stratonovich SDE and can be estimated using bifurcation analysis, (see Baxendale 1994) or using Taylor expansions, (see Kloeden and Platen 1992, Ch. 5.6). As the Stratonovich SDE is not a Markov process or a martingale, due to the definition of the Stratonovich integral, we do not consider this interpretation of (11.1). The definition in (11.1) represents a large class of continuous time models and many are used extensively in economics and finance. For an overview, see for example, Campbell, Lo, and Mackinlay (1997, Ch. 6, 9 and 11) and Rossi (1996). [1] For an outline of numerical and simulation methods in finance, see Kloeden, Platen, and Schurz (1991, Ch. 6) Dempster and Pliska (1997, Part V) and Rogers and Talay (1997).

Suppose that one has measurements $X_{\text{obs}} = \{X^i_{\text{obs}}\}_{i=1}^p$, an array of observed data points, where $X^i_{\text{obs}} = (X^i_{\text{obs1}}, \ldots, X^i_{\text{obs}T_i})$, at times $\tau^i_1, \ldots, \tau^i_{T_i}$, where X^i_{obs}, and $\tau^i = (\tau^i_1, \ldots, \tau^i_{T_i})$ are both $(1 \times T_i)$ vectors and $i = 1, \ldots, p$. Some series in $X(t)$ can represent unobserved processes, such as some latent volatility structure and in the extreme, is represented by $T_i = 0$ for the i'th series. We then define τ to be the ordered set of all distinct time points represented in the observed data, with T being the length of τ. To simplify the exposition, suppose we now define $\Delta^{\dagger i}_t$ to be the time step for the most frequently observed process between times t and $t+1$. That is, $\Delta^{\dagger i}_t = \tau^i_{t+1} - \tau^i_t$, for $i = 1, \ldots, p$, $t = 1, \ldots, T$ and $X_t = X(\tau_t)$, where $\tau_t \in \tau$ so that X_t contains the observed data points at time t contained in X_{obs} and possibly missing data points that may have to be imputed. In the analysis set out below, these will be updated along with any auxiliary points we introduce between X_t and X_{t+1} and define $\Delta^\dagger = \min_{t=1\tau} \min_{t=1}^T \Delta^{\dagger i}_t$. In doing so, we are implicitly assuming that the time steps for each process, $\Delta^{\dagger i}_t$, for $i = 1, \ldots, p$, is an integer multiple of $\Delta^\dagger_t$. We will let $X = \{X_t\}_{t=1}^T$ represent the $p \times T$ matrix of observed (allowing for missing observations) data. Situations where the number of auxiliary points imputed between series and/or each pair of time points is varied, can be naturally extended from the methods presented below.

In the likelihood context, (assuming we have T equally spaced observations for each process), estimation of the parameters is based on the likelihood function $\log L(X, X_0; \theta) = \sum_{t=0}^{T-1} \log g(X_{t+1}|X_t, \theta)$, where $g(X_{t+1}|X_t; \theta)$ are the Markovian transition densities. If a strong solution of the underlying SDE process is available, that is, the stochastic differential equation,

$$X_t = X_0 + \int_0^t a(X_s, \tau_s; \theta) ds + \int_0^t b(X_s, \tau_s; \theta) dW_s$$

can be solved analytically in Itô form, for $t \in (0, T]$, then $g(X_t|X_0; \theta)$ is available in closed form and likelihood inference is straightforward. For a comprehensive treatment of results on the uniqueness and existence of weak and strong solutions to stochastic differential equations, see Yamada and Watanabe (1971). Analytic solutions of SDE's are rarely available and only known in a limited number of cases, (see Kloeden and Platen, 1992, Section 4.4, for examples of explicitly solvable SDE's). This has led to growing interest in methods for estimating SDE's on the basis of discrete measurements.

The system defined in (11.1) can represent either multi-factor models, a system of p series, (for example in a term structure model, where each series represents a different maturity level, or a series of exchange rates across different currencies), or a combination of series, each with a multi-factor model associated with it. Multi-factor term structure models have been considered by Brennan and Schwartz (1979), Chen

[1] For an overview of univariate models used in economics and finance, see for example, Merton (1990) and Dixit (1993).

and Scott (1993) and Duffie and Kan (1996); Andersen and Lund (1997) estimate a multi-factor model of the short-term interest rate involving the stochastic volatility model. Geyer and Pichler (1995) estimate multi-factor Cox, Ingersoll and Ross processes using state-space form representation. Term structure models over different countries has been considered by Nielsen and Saá-Requejo (1993) and Saá-Requejo (1993); Ball and Torous (1999) apply a square-root volatility process over different currencies. A well known jump diffusion model is the model by Merton (1976) for option pricing. Jump diffusion processes for asset prices has been considered by Ahn and Thompson (1988) and Babbs and Webber, (1999, 1996), who use the monetary system to signal jumps. The papers by Das and Foresi (1996), Das (1996) and Björk, Di Masi, Kabanov, and Runggaldier (1997) do not use any institutional features to explain the jump structure. Jump diffusions have also been considered by Honoré (1998), who tries to incorporate both institutional and empirical features into the model. Various estimation techniques for estimating continuous-time multivariate models have been proposed in the recent literature. Important developments include the indirect inference method (Smith 1993, Gourieroux, Monfort, and Renault 1993 and Broze, Scaillet, and Zakaïan 1998) and the efficient method of moments approach of Gallant and Tauchen (1996). Other simulation methods have also been proposed by Aït-Sahalia (1998), Duffie and Singleton (1993) and Eraker (1998). Discretely observed diffusions have also been fit using estimating functions (see Bibby and Sørensen 1995, Kessler and Sørensen 1999 and Sørensen 1997) and filtering methods have been proposed by Nielsen, Madsen, Melgaard, and Baadsgaard (1998), Nielsen (1998), Nielsen, Vestergaard, and Madsen (1998) and Baadsgaard, Nielsen, Spliid, Madsen, and Preisel (1997).

We consider alternative methods for dealing with the estimation problem of multivariate stochastic differential equations as we are interested in developing an efficient general method that is not model specific, is likelihood based, can handle non-stationarity, and is not dependent on finding an appropriate auxiliary model. Our idea is simply to treat the values of the diffusion between any two discrete measurements as missing data and then to apply tuned Markov chain Monte Carlo methods to learn about the missing data and the parameters. Though the non-parametric approaches avoid restricting the functional form of the model, (which discretization imposes), the Markov condition can often be violated, for example, if one of the structures is latent.

The general specification given by (11.1) is chosen because it allows us to cover systems of SDE's which may arise in practical situations, where initial information content differs from that provided in the theoretical setting - an example of this is given below. In addition, dependence between processes can come from two sources, $b(X_t, \tau_t; \theta)$ and dW_t. The square-root of the volatility function $b(X_t, \tau_t; \theta)$, a $(p \times d)$ matrix determines the amount of correlation between the processes that depend on X_t. The $(d \times 1)$ vector, dW_t captures the correlation between the driving forces of each process and is independent of X_t. On applying a discretization scheme, the one-step ahead conditional transition density of X_{t+1}, given X_t and θ is obtained from the distribution of $\xi_t = W_{t+1} - W_t \sim N(0, \Delta^\dagger Q)$. The d-dimensional variance matrix $Q = PP'$ is assumed to be non-singular, finite and positive definite, (note we rule out the reduced rank case $p < d$, as this can always be re-formulated). Hence, the variance matrix of $X_{t+1}|X_t; \theta$ is given by $\Delta^\dagger \Sigma$, where $\Sigma = BB'$, $B = b(X_t, \tau_t; \theta)P$ and is assumed to have the same properties as Q. The introduction of auxiliary points with discretization is covered formally in the next sub-section. In practice however, we may be presented with an alternative representation of the multivariate system. As an example, consider a situation in which a specific form is proposed for each process, so that,

$$dX_t^i = a^i(X_t, \tau_t; \theta)dt + b^i(X_t, \tau_t; \theta\}dW_t^i$$

where $X_t^i \in \mathbb{R}^1$, $a^i(X_t, \tau_t; \theta) : \mathbb{R}^p \times [0, \infty) \times \mathbb{R}^q \longmapsto \mathbb{R}^1$ and $b^i(X_t, \tau_t; \theta) : \mathbb{R}^p \times [0, \infty) \times \mathbb{R}^q \to \mathbb{R}^1$ and

W_t^i is a one-dimensional Wiener process with increment variance $\Delta_t^{\dagger i}$, (the time step between t and $t+1$ for the i'th process), and $\theta \in \mathbb{R}^p$. Vectorizing dX_t^i over i corresponds to (11.1) where $d = 1$. The term $\{b^i(X_t, \tau_t; \theta)\}^2$ specifies the variance for each process. With no additional information on the correlation structure, any dependence between processes, (assuming dependence does not depend on X_t), is captured through the dW_t vector. On discretizing the process, using for example, the Euler scheme, (and assuming $\Delta_t^{\dagger} = \Delta^{\dagger}$), we have

$$X_{t+1}^i = X_t^i + a^i(X_t, \tau_t; \theta)\Delta^{\dagger} + b^i(X_t, \tau_t; \theta)\sqrt{\Delta^{\dagger}}\xi_t^i \quad \text{where} \quad \begin{cases} \xi_t^i = W_{t+1}^i - W_t^i \\ \xi_t = (\xi_t^1, \ldots, \xi_t^p)' \end{cases}$$

and the random component $\xi_t \sim N(0, \Delta^{\dagger}Q)$. The variance matrix for the process $X_{t+1}|X_t; \theta$ is then given by $\Delta^{\dagger}BB'$, where $B = DP$ and D is a diagonal matrix with the form $D = \text{diag}\{b^1(X_t, \tau_t; \theta), \ldots, b^p(X_t, \tau_t; \theta)\}$. In the case where the Wiener processes are independent, $Q = I_p$, so that $B = D$.

11.2 Discretization and augmentation issues

For sake of simplicity, we will consider the Euler-Maruyama scheme to carry out the discretization of the SDE, defined for the univariate case in (2.1). The Euler scheme can be thought of as the simplest discrete time approximation with respect to the strong convergence criterion, (see Kloeden and Platen, 1992, Section 10.2). Extensions to higher order schemes, such as the Milstein approximation considered in Section 9 can be pursued.

Suppose we consider the application of the Euler scheme to each process in (11.1) and for simplicity, introduce M auxiliary points between each pair X_t and X_{t+1}. We will denote the auxiliary points by $X_t^* = (X_{t,1}^*, \ldots X_{t,M}^*)$, imputed at times $\tau_{t,1} < \ldots < \tau_{t,M}$. The elements of X_t^* are $X_{t,j}^* = X^*(\tau_{t,j})$, each a p dimensional vector, for $j = 1, \ldots, M$. To distinguish between processes, we will denote the part of X_t^* applicable to the $j'th$ process as $X_t^*[j]$; All processes *except* for the j'th are denoted by $X_t^*[\backslash j]$. For notational convenience we have assumed that the points are evenly spaced, so that $\tau_{t,k+1} - \tau_{t,k} = \frac{\Delta^{\dagger}}{M+1} = \Delta$ for all t, k.[2] Applying the Euler scheme to (11.1) and letting $X_{t,M+1} = X_{t+1,0}$ yields

$$X_{t,k+1} = X_{t,k} + a(X_{t,k}, \tau_{t,k}; \theta)\Delta + b(X_{t,k}, \tau_{t,k}; \theta)(W_{t,k+1} - W_{t,k}), \tag{11.2}$$

for all t and $k = 0, \ldots, M+1$. The distribution of $W_{t,k+1} - W_{t,k}$ is Normal with zero mean and a d-dimensional covariance matrix given by ΔQ, which is independent of $X_{t,k}$ and $\tau_{t,k}$, using the fact that the increments of $W_{t,k}$ are independent over time. We assume there exists a matrix P, such that $Q = PP'$ and Q is assumed to be non-singular, finite and positive definite. We will only consider cases where $p \geq d$, since $p < d$ corresponds to a singular Q matrix. We assume that $\Sigma(X_{t,k}) = \Sigma(X_{t,k}, \tau_{t,k}; \theta) = b(X_{t,k}, \tau_{t,k}; \theta)PP'b(X_{t,k}, \tau_{t,k}; \theta)'$ is also finite, non-singular and positive definite and use the notation $B(X_t) = B(X_{t,k}, \tau_{t,k}; \theta) = b(X_{t,k}, \tau_{t,k}; \theta)P$ and $\Sigma(X_{t,k}) = B(X_{t,k})B(X_{t,k})'$for compactness. Under this scheme, the approximate one-step ahead conditional transition density is given by

$$X_{t,k+1}|X_{t,k}; \theta \overset{\text{approx.}}{\sim} N\{X_{t,k} + a(X_{t,k}, \tau_{t,k}; \theta)\Delta,\ \Delta\Sigma(X_{t,k})\}, \tag{11.3}$$

[2] $\Delta^{\dagger}$ is determined by the time interval of the process. The choice of units for $\Delta^{\dagger}$ has a bearing on the scale of θ, (and so implicitly scales the drift and volatility functions). Since the initial choice of units is merely convenient labelling, (chosen for ease of interpretation), the scaling is arbitrary and not reflected in the notation.

which we can write in density form as $f(X_{t,k+1}|X_{t,k};\theta)$, to distinguish it from the (unknown) true transition density $g(X_{t,k+1}|X_{t,k};\theta)$.

11.3 MCMC INFERENCE

The univariate case outlined in Section 5 is now extended to the multivariate setting. The estimation method again relies on Markov chain Monte Carlo methods. Let $X^* = (X_t^*)_{t=1}^T$ represent the $pM \times (T-1)$ matrix of auxiliary points, where $\text{vec}(X_t^*)$ denotes the pM-dimensional vector obtained from applying the vec operator to X_t^*. If X_t contains any missing data points, these can be sampled at the same time as the auxiliary data. In the case where none of $X^*[j]$ is observed, the end points are treated as in 10.2.1. As no new issues arise, we will assume without loss of generality, that the X_t vector contains no missing data. In brief, the aim is to sample the Bayesian posterior density $\pi(\theta, X^*|X)$, by first sampling $X_t^*|X_t, X_{t+1}, \theta$, $t = 1, ..., T$. We let $n = pT(M+1) - Mp$ represent both the number of (potentially) observable data and the number of auxiliary points imputed. We then sample $\theta|X, X^*$ and repeat this process until the chain has converged to its stationary distribution. The key step in this process is the sampling of the latent data $\{X_t^*\}$. Given the dimensionality of the distribution $X_t^*|X_t, X_{t+1}, \theta$, this sampling must be conducted with care in order to produce a Markov chain in which the serial correlations vanish rapidly. A method will be presented below which samples the matrix of auxiliary states, X_t^*, by considering each process in turn in and is outlined in 11.3.2.

11.3.1 Overview

Consider the question of sampling X^* from $f(X^*|X;\theta) = \prod_{t=1}^n f(X_t^*|X_t, X_{t+1}, \theta)$, where we have used the fact that the X_t^* are conditionally independent, given X_t and X_{t+1}. It therefore suffices to consider the simulation of $X_t^* = (X_{t,1}^*, \ldots, X_{t,M}^*)$ from $f(X_t^*|X_t, X_{t+1}, \theta)$. Suppose we now restrict attention to the block X_t^*, between the vectors X_t and X_{t+1}. The target density of interest can also be written as

$$f(X_t^*|X_t, X_{t+1};\theta) \propto \prod_{k=0}^{M} f(X_{t,k+1}^*|X_{t,k}^*;\theta), \tag{11.4}$$

where $X_t = X_{t,0}^*$ and $X_{t+1} = X_{t,M+1}^*$.[3]

[3]As an example, consider the case $p = 3$, so that

$$\underset{X_t}{\begin{pmatrix} X_t[1] \\ X_t[2] \\ X_t[3] \end{pmatrix}} \quad \underset{X_{t,1}^* \quad X_{t,2}^* \quad \cdots \quad X_{t,M}^*}{\begin{pmatrix} X_{t,1}^*[1] & X_{t,2}^*[1] & \cdots & X_{t,M}^*[1] \\ X_{t,1}^*[2] & X_{t,2}^*[2] & \cdots & X_{t,M}^*[2] \\ X_{t,1}^*[3] & X_{t,2}^*[3] & \cdots & X_{t,M}^*[3] \end{pmatrix}} \quad \underset{X_{t+1}}{\begin{pmatrix} X_{t+1}[1] \\ X_{t+1}[2] \\ X_{t+1}[3] \end{pmatrix}}$$

We could proceed by

- sampling all of the block $X_t^*|X_t, X_{t+1};\theta$ at once, or
- sampling individual or multiple columns at a time, for example, $X_{t,1}^* \mid X_{t,2}^*, X_{t,3}^*, X_t, X_{t+1};\theta$, or
- sampling individual or multiple rows at a time, for example, $X_t^*[1], X_t^*[2] \mid X_t^*[3], X_t, X_{t+1};\theta$.

Sampling $X_t^*|X_t, X_{t+1}; \theta$ in one block : A method for updating the whole block in one sweep can be carried out using the following procedure. The approximate conditional log density for X_t^* is then

$$l(X_t^*) = \log f(X_t^*|X_t, X_{t+1}; \theta) = c + \sum_{s=0}^{M} \log f(X_{t,s+1}^*|X_{t,s}^*; \theta)$$

where $\log f(X_{t,s+1}^*|X_{t,s}^*; \theta) \sim N\left\{X_{t,s}^* + a(X_{t,s}^*, \tau_{t,s}; \theta)\Delta,\ \Delta\Sigma(X_{t,s}^*)\right\}$. If $l(X_t^*)$ is Gaussian, a Gibbs sampler can be used to update the states. However, suppose that the drift and volatility functions of each process are such that either $X_{t,s}^*[\backslash j]$ enters linearly into $a^j(X_{t,s}^*, \tau_{t,s}; \theta)$ and $b^j(X_{t,s}^*, \tau_{t,s}; \theta)$ or each process is independent so that $\Sigma(X_{t,s}^*)$ is a diagonal matrix. Unless each individual SDE, (ignoring explicitly solvable cases), has the property that $X_{t,s}^*[j]$ appears linearly in $a^j(X_{t,s}^*[j], \tau_{t,s}; \theta)$ or $b^j(X_{t,s}^*[j], \tau_{t,s}; \theta)$ can be expressed as a function of $\tau_{t,s}$ and θ alone, so that the product over the transition densities is also Gaussian, a MH algorithm must be used.[4] In general, $X_{t,s}^*$ appears nonlinearly in the functions $a(X_{t,s}^*, \tau_{t,s}; \theta)$ and $b(X_{t,s}^*, \tau_{t,s}; \theta)$ and therefore it is not possible to sample $X_{t,s}^*$ directly from the target density.

If $l(X_t^*)$ is non-Gaussian, a proposal density, (chosen so that it is easy to sample from), can be constructed by taking a second order Taylor expansion around $\hat{X}_t^*$, (which can additionally be iterated a few steps towards the mode of the density using a Newton-Raphson scheme), as proposed in the univariate case. The expansion gives rise to a multivariate Gaussian proposal density, which in general notation can be written as

$$q(X_t^*) = q(X_t^*|X_t, X_{t+1}; \theta) \sim N(\tilde{\mu}, \tilde{\Sigma}), \tag{11.5}$$

where the pM-dimensional vector mean $\tilde{\mu}$ and $(pM \times pM)$ variance matrix $\tilde{\Sigma}$ are defined by the first and second derivatives of the log-likelihood function evaluated at $\hat{X}_t^*$. To conclude the MH step, a proposal value z, $(pM \times 1)$, is drawn from $q(X_t^*)$ and is accepted with probability

$$\alpha\left\{X_t^{*(g)}, z\right\} = \min\left[1, \frac{f\{z|X_t, X_{t+1}; \theta\} q\left\{X_t^{*(g)}\right\}}{f\left\{X_t^{*(g)}|X_t, X_{t+1}; \theta\right\} q(z)}\right],$$

where $X_t^{*(g)}$ is the current value of X_t^* at iteration g of the MCMC sampling. If the proposal value is rejected, the next item in the chain is taken to be $X_t^{*(g)}$.

There is a danger that for some very complicated problems, this MCMC algorithm will reject quite often due to the size of the block to be sampled. The same is true for simpler models, (though the MH algorithm may collapse to the Gibbs sampler), especially when M is large; the sampler may have poor mixing properties, getting stuck in regions of the support of the target density. In such cases it can be effective to split the block into pieces and perform MCMC based on the sub-blocks.

Sampling individual or multiple columns at a time Suppose we consider sampling columns of X_t^*, given its neighbors. It will generally not be advisable to sample columns of X_t^* one at a time, due

[4]Models where $X_t^* \,|\, X_t, X_{t+1}; \theta$ is Gaussian and $X_{t,s}^*[\backslash j]$ enters linearly into the drift and volatility functions for each j-th process are limited. As an example, consider a GBM with a latent volatility structure where the log volatility is specified by the OU process; both processes have known strong solutions, but the volatilities appear non-linearly in the GBM model.

to issues such as increased autocorrelation time, considered in earlier parts of the thesis. In the cases where, for example, $X_t^*|X_t, X_{t+1}$ is Gaussian, we can partition X_t^* as

$$X_t^* = \begin{pmatrix} X_t^{*1} \\ \dots \\ X_t^{*2} \end{pmatrix} \text{ such that } \begin{cases} \text{for a given series } j \\ \text{and fixed } l, (l = 1, 2) \end{cases} X_{t,k}^*[j] \in X_t^{*l} \ \forall k.$$

In this case, $X_t^{*i} \ |X_t, X_t^{*\backslash i}, X_{t+1}; \theta$ is Gaussian and sampling each partition in blocks of columns is straightforward. Details of this scheme for univariate models were given in Section 5 and no new conceptual issues arise at this point.

For more complicated models, structures such as conditional Gaussianity, given both neighbors do not hold. An alternative method of reducing the size of the block is by subsampling proposed individual processes, one at a time. In the bivariate case, $(p = 2)$, we may propose to move the first series $X_t^*[1]$, keeping $X_t^*[2]$ fixed at its current value. This reduces the dimension considerably and will be the procedure outlined in 11.3.2 as the method is general and covers complicated systems of SDE's. Treating each process in turn, the sampling algorithm proposed will be based on sub-dividing $X_t^*[j]$ further into blocks to increase the efficiency of the sampler.

11.3.2 Simulation of the auxiliary states: sub-sampling

A general method for sampling X_t^*, is presented, based on subsampling proposed individual processes, one at a time. To illustrate, suppose we consider imputing M auxiliary points between each pair of observations, for p SDE's. Using $X_{t,k}^* = (X_{t,k}^*[1], \dots X_{t,k}^*[p])'$, the one step ahead conditional transition density is given by

$$f\left(X_{t,k+1}^*|X_{t,k}^*;\theta\right) = \left(\frac{1}{2\pi}\right)^{p/2} \left|\Delta\Sigma\left(X_{t,k}^*\right)\right|^{-1/2} \tag{11.6}$$
$$\times \exp\left[-\frac{1}{2\Delta}\left\{X_{t,k+1}^* - a(X_{t,k}^*, t;\theta)\Delta - X_{t,k}^*\right\}' \Sigma\left(X_{t,k}^*\right)^{-1} \left\{X_{t,k+1}^* - a(X_{t,k}^*, t;\theta)\Delta - X_{t,k}^*\right\}\right],$$

where $\Sigma(X^*_{t,k}) = b(X^*_{t,k})Qb(X^*_{t,k})'$.[5] When $M > 1$, we know that

$$\begin{aligned} f(X^*_t|X_t, X_{t+1};\theta) &\propto f(X_{t+1}|X^*_t;\theta)\ f(X^*_t|X_t;\theta)\,, \\ \text{where } f(X_{t+1}|X^*_t;\theta) &= f\left(X_{t+1}|X^*_{t,M};\theta\right) \\ \text{and } f(X^*_t|X_t;\theta) &= f\left(X^*_{t,1},\ldots,X^*_{t,M}|X_t;\theta\right) = \prod_{k=0}^{M-1} f\left(X^*_{t,k+1}|X^*_{t,k};\theta\right) \end{aligned}$$

so that $f(X^*_t|X_t, X_{t+1};\theta)$ is the product of $(M+1)$ Normal conditional transition densities,

$$\begin{aligned} & f(X^*_t|X_t, X_{t+1}) \propto \prod_{k=0}^{M} f(X^*_{t,k+1}|X^*_{t,k};\theta) \\ \text{where} \quad & f\left(X^*_{t,k+1}|X^*_{t,k};\theta\right) \sim N\left\{a(X^*_{t,k},\tau_{t,k},\theta)\Delta + X^*_{t,k},\ \Sigma(X^*_{t,k})\Delta\right\}. \end{aligned} \tag{11.7}$$

Block sampling for each process: The log-likelihood for X^*_t is given by

$$l(X^*_t) = c + \sum_{k=0}^{M} \log f(X^*_{t,k+1}|X^*_{t,k};\theta) = c + \sum_{k=0}^{M} l_k$$

where c is a constant independent of X^*_t and we note that

$$\log f\left(X^*_t[j] \mid X_t, X_{t+1}, X^*_t[\backslash j];\theta\right) = c + \log f\left(X^*_t|X_t, X_{t+1};\theta\right).$$

The auxiliary points, $X^*_t[j]$, will be updated by augmenting in blocks of size m, (where $1 \le m \le M$), for $j = 1,\ldots p$, starting with observation $X^*_{t,s}[j]$, $(1 \le s \le M)$. The block to be updated for the j'th process is therefore $X^*_{t,(s,m)}[j] = (X^*_{t,s}[j],\ldots,X^*_{t,s+m-1}[j])$, $(1\times m)$. We will define the matrix of derivatives to be used in the Taylor expansion of the log-likelihood with respect to X^*_t for notational compactness. That is, the matrix of derivatives for $X^*_{t,(s,m)} = (X^*_{t,s},\ldots,X^*_{t,s+m-1})$ will be stored in one matrix, the relevant

[5] As an example, if $M = 1$ and $p = 2$, then

$$\begin{aligned} X_{t+1}|X^*_t;\theta &\sim N\left[\begin{Bmatrix} a^1(X^*_{t,1},\tau_{t,1};\theta) \\ a^2(X^*_{t,1},\tau_{t,1};\theta) \end{Bmatrix}\Delta + \begin{pmatrix} X^*_{t,1}[1] \\ X^*_{t,1}[2] \end{pmatrix}, \begin{pmatrix} \Sigma(X^*_{t,1})_{11} & \Sigma(X^*_{t,1})_{12} \\ \Sigma(X^*_{t,1})_{21} & \Sigma(X^*_{t,1})_{22} \end{pmatrix}\Delta\right] \\ X^*_{t,1}|X_t;\theta &\sim N\left[\begin{Bmatrix} a^1(X_t,\tau_t;\theta) \\ a^2(X_t,\tau_t;\theta) \end{Bmatrix}\Delta + \begin{pmatrix} X_t[1] \\ X_t[2] \end{pmatrix}, \begin{pmatrix} \Sigma(X_t)_{11} & \Sigma(X_t)_{12} \\ \Sigma(X_t)_{21} & \Sigma(X_t)_{22} \end{pmatrix}\Delta\right]. \end{aligned}$$

On applying the result on conditional Normal distributions

$$\begin{aligned} X^*_{t,1}[1] \mid X_t, X^*_{t,1}[2];\theta \sim N\Big[& a^1(X_t,\tau_t;\theta)\Delta + X_t[1] + \Sigma(X_t)_{12}\Sigma(X_t)_{22}^{-1}\left\{X^*_{t,1}[2] - a^2(X_t,\tau_t;\theta)\Delta - X_t[2]\right\}, \\ & (\Sigma(X_t)_{11} - \Sigma(X_t)_{12}\Sigma(X_t)_{22}^{-1}\Sigma(X_t)_{21})\Delta\Big] \end{aligned}$$

and since

$$\begin{aligned} f(X^*_{t,1}[1] \mid X_t, X_{t+1}, X^*_{t,1}[2];\theta) &\propto f(X_{t+1}|X^*_t, X_t;\theta)\ f(X^*_{t,1}[1] \mid X_t, X^*_{t,1}[2];\theta) \\ &= f(X_{t+1}|X^*_t;\theta)\ f(X^*_{t,1}[1] \mid X_t, X^*_{t,1}[2];\theta), \end{aligned}$$

we can construct the proposal for the MH algorithm by taking a Laplace approximation of the log-likelihood around $\hat{X}^*_{t,1}[1]$.

entries being picked out when we update the j'th series. We define, for $j = 1, \ldots, p$ and $i = 1, \ldots, m$,

$$w_i = \begin{pmatrix} w_i^1 \\ w_i^2 \\ \vdots \\ w_i^p \end{pmatrix} = X^*_{t,s+i} = \begin{pmatrix} X^*_{t,s+i}[1] \\ X^*_{t,s+i}[2] \\ \vdots \\ X^*_{t,s+i}[p] \end{pmatrix} \ (p \times 1), \quad \text{and} \quad w^j = \begin{pmatrix} X^*_{t,s}[j] \\ X^*_{t,s+1}[j] \\ \vdots \\ X^*_{t,s+m-1}[j] \end{pmatrix}' \ (1 \times m)$$

$$w = \begin{pmatrix} w_1' \\ w_2' \\ \vdots \\ w_m' \end{pmatrix}' = \begin{pmatrix} w^1 \\ w^2 \\ \vdots \\ w^p \end{pmatrix} = \begin{pmatrix} X^*_{t,s}[1] & X^*_{t,s+1}[1] & \ldots & X^*_{t,s+m-1}[1] \\ X^*_{t,s}[2] & X^*_{t,s+1}[2] & \ldots & X^*_{t,s+m-1}[2] \\ \vdots & \ddots & \ddots & \vdots \\ X^*_{t,s}[p] & X^*_{t,s+1}[p] & \ldots & X^*_{t,s+m-1}[p] \end{pmatrix} = \begin{pmatrix} w_1^1 & w_2^1 & \ldots & w_m^1 \\ w_1^2 & w_2^2 & \ldots & w_m^2 \\ \vdots & \ddots & \ddots & \vdots \\ w_1^p & w_2^p & \ldots & w_m^p \end{pmatrix} \ (p \times m)$$

so that each column, w_i, has neighbors w_{i-1} and w_{i+1}, for $i = 1, \ldots, m$ and the fixed neighbors are w_0, w_{m+1}. Now,[6]

$$\begin{aligned} l(w) &= \log f(w|w_0, w_{m+1}; \theta) = c + \sum_{i=0}^{m} \log f(w_{i+1}|w_i; \theta) = c + \sum_{i=0}^{m} l_i \\ l(w^j) &= \sum_{i=1}^{m} l_i \text{ a function of } (w_i^j, w_{i+1}^j) \text{ alone, keeping } w^{\setminus j} \text{ and } (w_1^j, \ldots, w_{i-1}^j, w_{i+2}^j, \ldots, w_m^j) \text{ fixed,} \end{aligned}$$

a second order Taylor approximation of $l(w^j)$ is taken around $\hat{w}^j = (\hat{w}_1^j, \ldots, \hat{w}_m^j)'$.

$$l(w^j) = \sum_{i=0}^{m} l_i \approx c + (w^j - \hat{w}^j)' \frac{\partial l(\hat{w}^j)}{\partial w^j} - \frac{1}{2}(w^j - \hat{w}^j)' \frac{\partial^2 l(\hat{w}^j)}{\partial w^j \partial w^{j'}} (w^j - \hat{w}^j).$$

The derivatives used in the Taylor expansion are quite involved and so only the first derivatives are provided. These are given by,

$$\begin{aligned} \frac{\partial l(w^j)}{\partial w_i^j} &= \sum_{k=0}^{m} \frac{\partial l_k}{\partial w_i^j} = \frac{\partial l_{i-1}}{\partial w_i^j} + \frac{\partial l_i}{\partial w_i^j} \\ \frac{\partial^2 l(w^j)}{\left(\partial w_i^j\right)^2} &= \frac{\partial^2 l_{i-1}}{\left(\partial w_i^j\right)^2} + \frac{\partial^2 l_i}{\left(\partial w_i^j\right)^2} \\ \frac{\partial^2 l(w^j)}{\partial w_i^j \partial w_{i+1}^j} &= \frac{\partial^2 l(w^j)}{\partial w_{i+1}^j \partial w_i^j} = \frac{\partial^2 l_i}{\partial w_i^j \partial w_{i+1}^j} \end{aligned}$$

since the only term involving w_i^j and w_{i+1}^j is l_i. The proposal density $q(w^j|w_0^j, w_{m+1}^j; \theta)$ is thus Gaussian with mean $\tilde{\mu}^j$ and variance $\tilde{\Sigma}^j$, where

$$\tilde{\mu}^j = \hat{w}^j + \tilde{\Sigma}^j u_t^j \tag{11.8}$$

$$\tilde{\Sigma}^j = (V_t^j)^{-1} \tag{11.9}$$

[6]Due to the Markov structure, it will only be necessary to consider the derivatives of l_i with respect to $w_i = X^*_{t,s+i}$ and $w_{i+1} = X^*_{t,s+i+1}$, for $i = 1, \ldots, m-1$.

where u_t^j is $(m \times 1)$, with elements

$$u_{t,i}^j = \frac{\partial l(w^j)}{\partial w_i^j}$$

and V_t^j is $(m \times m)$, with elements

$$\begin{aligned} V^j_{(t,i),(t,i)} &= -\frac{\partial^2 l(w^j)}{\left(\partial w_i^j\right)^2} \\ V^j_{(t,i),(t,i+1)} &= V^j_{(t,i+1),(t,i)} = \frac{\partial^2 l(w^j)}{\partial w_i^j \partial w_{i+1}^j} \quad \text{for } i = 1, \ldots, m-1 \\ V^j_{(t,i),(t,i+s)} &= 0 \ \forall\, i \text{ and } s > 1, \end{aligned}$$

for $i = 1, \ldots, m$, unless otherwise stated. All quantities are evaluated at $\hat{w}^j$.

Expanding out the square of l_i and defining $B_i = B(w_i) = B(X^*_{t,s+i})$, we obtain

$$\begin{aligned} l_i &= d - \frac{1}{2}\log|B_iB_i'| - \frac{1}{2\Delta}\left[w_{i+1}'(B_iB_i')^{-1}w_{i+1} - 2w_{i+1}'(B_iB_i')^{-1}\left\{w_i + a(w_i, \tau_{t,s+i}; \theta)\Delta\right\}\right. \\ &\quad \left. + \left\{w_i + a(w_i, \tau_{t,s+i}; \theta)\Delta\right\}'(B_iB_i')^{-1}\left\{w_i + a(w_i, \tau_{t,s+i}; \theta)\Delta\right\}\right]. \end{aligned} \tag{11.10}$$

We first note that the vec operator applied to a scalar leaves it unchanged. We will alternate between different representations of a given expression, (depending on which is most convenient), so that the scalar (ABC), $\text{vec}(ABC)$ and $\text{vec}(ABC) = (C' \otimes A)\text{vec(B)}$ all denote the same term. Also, since the dimension of w_i is $(p \times 1)$, $\text{vec}(w_i) = w_i$. Results from Magnus and Neudecker, (1988, Part III, Chapters 9, 10) and Lütkepohl (1996) are used to derive the relevant expressions, which can also be found in Appendix 11.5.1.

For the first expression $-(1/2)\log|B_iB_i'|$,

$$\frac{\partial \log|B_iB_i'|}{\partial w_i'} = \text{vec}'\left\{(B_iB_i')^{-1}\right\}\left\{\frac{\partial}{\partial w_i'}\text{vec}(B_iB_i')\right\} \quad \text{dimension } (1 \times p) \tag{11.11}$$

$$= \text{vec}'\left\{(B_iB_i')^{-1}\right\}(I_{p^2} + K_{pp})(B_i \otimes I_p)\frac{\partial \text{vec}(B_i)}{\partial w_i'} \tag{11.12}$$

$$\frac{\partial \log|B_iB_i'|}{\partial w_{i+1}'} = 0_{1\times p} \tag{11.13}$$

where we have used (C.11.1) in (11.11) and the fact that $\partial\text{vec}'(A)/\partial X = \{\partial\text{vec}(A)/\partial X\}'$. Equation (11.12) is obtained using (C.11.2) and (C.11.3).

For the expression $w_{i+1}'(B_iB_i')^{-1}w_{i+1}$, we first note that

$$\begin{aligned} \frac{\partial}{\partial w_i'}\left\{\text{vec}\,(B_iB_i')^{-1}\right\} &= -\left\{(B_iB_i')^{-1} \otimes (B_iB_i')^{-1}\right\}\frac{\partial}{\partial w_i'}\text{vec}(B_iB_i') \quad \text{dimension } (p^2 \times p) \\ &= -\left\{(B_iB_i')^{-1} \otimes (B_iB_i')^{-1}\right\}(I_{p^2} + K_{pp})(B_i \otimes I_p)\frac{\partial}{\partial w_i'}\text{vec}(B_i) \end{aligned} \tag{11.14}$$

using (C.11.2) and (C.11.5). So, using (C.11.7)

$$\frac{\partial}{\partial w_i'}\left\{w_{i+1}'(B_iB_i')^{-1}w_{i+1}\right\} = -(w_{i+1}'\otimes w_{i+1}')\left\{(B_iB_i')^{-1}\otimes(B_iB_i')^{-1}\right\}(I_{p^2}+K_{pp}) \times(B_i\otimes I_p)\frac{\partial}{\partial w_i'}\mathrm{vec}(B_i) \tag{11.15}$$

$$\frac{\partial}{\partial w_{i+1}'}\left\{w_{i+1}'(B_iB_i')^{-1}w_{i+1}\right\} = 2w_{i+1}'(B_iB_i')^{-1}. \tag{11.16}$$

where (C.11.4) is used for (11.16); both (11.15) and (11.16) have dimension $(1\times p)$.

For the term $\mathrm{vec}\left[w_{i+1}'(B_iB_i')^{-1}\left\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\right\}\right]$, we observe that

$$\frac{\partial}{\partial w_i'}\mathrm{vec}\left\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\right\} = \frac{\partial}{\partial w_i'}\left\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\right\} = I_p+\Delta\frac{\partial}{\partial w_i'}a(w_i,\tau_{t,s+i};\theta)$$

so that

$$\begin{aligned}
&\frac{\partial}{\partial w_i'}\mathrm{vec}\left[w_{i+1}'(B_iB_i')^{-1}\left\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\right\}\right]\\
&= \left\{1\otimes w_{i+1}'(B_iB_i')^{-1}\right\}\frac{\partial}{\partial w_i'}\mathrm{vec}\left\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\right\}\\
&\quad+\left[\left\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\right\}'\otimes 1\right]\frac{\partial}{\partial w_i'}\left\{w_{i+1}'\mathrm{vec}(B_iB_i')^{-1}\right\}\quad\text{using (C.11.6–C.11.7)}\\
&= \left\{1\otimes w_{i+1}'(B_iB_i')^{-1}\right\}\left[I_p+\Delta\frac{\partial}{\partial w_i'}\left\{a(w_i,\tau_{t,s+i};\theta)\right\}\right]-\left\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\right\}'\left(I_p\otimes w_{i+1}'\right)\\
&\quad\times\left\{(B_iB_i')^{-1}\otimes(B_iB_i')^{-1}\right\}(I_{p^2}+K_{pp})(B_i\otimes I_p)\frac{\partial}{\partial w_i'}\mathrm{vec}(B_i)
\end{aligned}\tag{11.17}$$

$$\begin{aligned}
&\frac{\partial}{\partial w_{i+1}'}\mathrm{vec}\left[w_{i+1}'(B_iB_i')^{-1}\left\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\right\}\right]\\
&\quad=\left\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\right\}'(B_iB_i')^{-1}\quad\text{using (C.11.7)}
\end{aligned}\tag{11.18}$$

The next term in (11.10) is $\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\}'(B_iB_i)^{-1}\{w_i+a(w_i,\tau_{t,s+i};\theta)\Delta\}$ is a function of w_i

only, so that using (C.11.6),

$$
\begin{aligned}
&\frac{\partial}{\partial w_i'}\left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'(B_iB_i)^{-1}\left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\} \\
= &\left[1 \otimes \left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'(B_iB_i')^{-1}\right] \times \frac{\partial}{\partial w_i'}\left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\} \\
&+ \left[\left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}' \otimes 1\right]\frac{\partial}{\partial w_i'}\operatorname{vec}\left[\left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'(B_iB_i')^{-1}\right] \\
= &\left[1 \otimes \left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'(B_iB_i)^{-1}\right]\left\{I_p + \Delta\frac{\partial}{\partial w_i'}a(w_i, \tau_{t,s+i};\theta)\right\} \\
&+ \left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'\left(\left[I_p \otimes \left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'\right]\frac{\partial}{\partial w_i'}\left[\operatorname{vec}\left\{(B_iB_i')^{-1}\right\}\right]\right. \\
&\left.+ \left\{(B_iB_i')^{-1} \otimes 1\right\}\frac{\partial}{\partial w_i'}\operatorname{vec}\left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'\right) \\
= &\left[1 \otimes \left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'(B_iB_i)^{-1}\right]\left\{I_p + \Delta\frac{\partial}{\partial w_i'}a(w_i, \tau_{t,s+i};\theta)\right\} \\
&+ \left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'\left(-\left[I_p \otimes \left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'\right]\left\{(B_iB_i')^{-1} \otimes (B_iB_i')^{-1}\right\}\right. \\
&\left.\times(I_{p^2} + K_{pp})(B_i \otimes I_p)\frac{\partial}{\partial w_i'}\operatorname{vec}(B_i) + (B_iB_i')^{-1}\left[I_p + \Delta\frac{\partial}{\partial w_i'}\left\{a(w_i, \tau_{t,s+i};\theta)\right\}\right]\right) \qquad (11.19)
\end{aligned}
$$

and

$$
\frac{\partial}{\partial w_{i+1}'}\left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\}'(B_iB_i)^{-1}\left\{w_i + a(w_i, \tau_{t,s+i};\theta)\Delta\right\} = 0_{1\times p}. \qquad (11.20)
$$

Similar expressions can be derived for the second derivatives.

11.4 Conclusion

The univariate analysis previously outlined has been extended to deal with general multivariate diffusion processes. The method is still based around use of a tuned Metropolis-Hastings (MH) step. The method may be simplified in situations where the SDE's are partially solvable or for specific structures of the drift and volatility functions. Simulation experiments have not been reported here and form part of ongoing research.

11.5 Appendix

11.5.1 Matrix differential calculus results

The following are used in the derivation of the first and second derivatives of the log-likelihood with respect to a vector of auxiliary states in Section 11.3.2. For further reading, see Lütkepohl (1996) and Magnus and Neudecker, (1988, Part III, Chapters 9, 10).

Suppose we define X to be a $p \times 1$ vector, $A = A(X)$ to be a $(p \times p)$ non-singular matrix, $B = B(X)$ to be a $(p \times q)$ matrix, with BB' finite, positive definite and non-singular and $C = C(X)$ to be a $(q \times l)$

matrix. Additionally define $D = D(X)$, $(l \times q)$, and constants E $(n \times p)$, F $(q \times l)$ and G $(l \times m)$. Then

$$\frac{\partial}{\partial X'} \ln |A| = \text{vec}' \left(A^{-1}\right) \left\{ \frac{\partial}{\partial X'} \text{vec}(A) \right\} . (1 \times p^2) \tag{C.11.1}$$

$$\frac{\partial}{\partial X'} \text{vec}(BB') = (I_{p^2} + K_{pp})(B \otimes I_p) \frac{\partial}{\partial X'} \text{vec}(B) \tag{C.11.2}$$

where (C.11.2) is obtained from Lütkepohl, (10.5.5, (2), p. 196, 1996) and for an $(m \times n)$ matrix H, K_{mn} is defined to be the $(mn \times mn)$ commutation matrix $\text{vec}(H') = K_{mn}\text{vec}(H)$, where

$$K_{mn} = \sum_{k=1}^{m} \sum_{l=1}^{n} H_{kl} \otimes H'_{kl} \quad \text{with } H_{kl} = [hij], (m \times n) \text{ such that } \begin{cases} h_{kl} = 1 \\ h_{ij} = 0 \begin{cases} i \neq k \\ or\ j \neq l \end{cases} \end{cases} \tag{C.11.3}$$

and $K_{p1} = K_{1p} = I_p$, $K'_{pp} = K_{pp}$, (see Lütkepohl, Section 9.2.5, 1996). For a constant vector c, $(n \times 1)$, $\partial c / \partial X' = 0_{n \times p}$, (see Lütkepohl, Section 10.4.4, (1), 1996) and using a variation of Lütkepohl (1996, p. 174, Section 10.2.1, (9)) for (C.11.4),

$$\frac{\partial}{\partial X'} X'AX = X'(A' + A) \tag{C.11.4}$$

$$\frac{\partial}{\partial X'} \text{vec}(A^{-1}) = -(A'^{-1} \otimes A^{-1}) \frac{\partial}{\partial X'} \text{vec}(A) \tag{C.11.5}$$

$$\frac{\partial}{\partial X'} \text{vec}(BC) = (I_l \otimes B) \frac{\partial}{\partial X'} \text{vec}(C) + (C' \otimes I_p) \frac{\partial}{\partial X'} \text{vec}(B) \tag{C.11.6}$$

$$\frac{\partial}{\partial X'} \text{vec}(EBF) = (F' \otimes E) \frac{\partial}{\partial X'} \text{vec}(B) \tag{C.11.7}$$

$$\frac{\partial}{\partial X'} \text{vec}(EBFDH) = (H' \otimes EBF) \frac{\partial}{\partial X'} \text{vec}(D) + (H'D'F' \otimes E) \frac{\partial}{\partial X'} \text{vec}(B) \tag{C.11.8}$$

See Lütkepohl (1996, p. 200, Section 10.6.3, (1)) for (C.11.5). For (C.11.6–C.11.8), see Lütkepohl (1996, p. 196, Section 10.5.5, (3) and (4)).

Chapter 12

Concluding remarks and further topics

New methods to estimate general continuous-time non-linear stochastic differential equation models which typically arise in financial econometrics have been considered. These methods enable us to obtain more accurate estimates of the underlying parameters present. As the solution to the true continuous process is rarely known for these applications, the methods developed have relied on discretization to obtain an approximate likelihood $f^M(y_{t+1}|y_t)$ and using simulation techniques to obtain the posterior distribution for the unknown quantities of interest. The likelihood based method is implemented within the Bayesian paradigm, as in the context of these models, Bayesian methods are often analytically easier. The approach has many advantages, in particular, in the context of diagnostic testing, the Bayesian framework allows us to compare the fit of a model across various alternatives including non-nested models.

The missing paths between the observed data points are simulated in order to reduce the bias from the discretized version. Higher order schemes to discretize the process, such as the Milstein scheme have been considered. Their implementation with simulation methods has until now not been possible as it requires us to simulate from the density and additionally obtain an estimate of the likelihood at any given point. Many methods have relied on the Euler-Maruyama scheme, which is known to induce considerable bias when using discrete data, even for the less non-linear models. The existence of a closed form solution for the higher order Milstein scheme means that it can be used within any estimation method that relies on using a discretized version of the true model. The method can be easily extended to cover irregularly spaced intervals or a system of many stochastic differential equations with general structures for the driving process and state space.

The methodology presented in this thesis has relied on simulation techniques, which are flexible and efficient in the context of models that are often intractable and computationally demanding. The methodology provides an efficient toolkit for estimating continuous-time models, whose performance can be assessed through the use of diagnostic tools, so that meaningful inference can be conducted. The approach generalizes to multivariate cases, state processes, Lèvy processes, and general non-stationary models.

The subject of future research will concentrate on further developing the ideas of simulation estimation and stochastic differential equation theory in the context of non-linear economic series. Specifically, I hope to be able to apply these methods in the multivariate setting, in conjunction with the use of higher order Taylor expansions. I hope to also explore additional issues, such as evaluating the likelihood using techniques such as particle filtering.

Part IV

References

BIBLIOGRAPHY

Abramowitz, M. and I. Stegun (1970). *Handbook of Mathematical Functions* (Ninth ed.). New York: Dover Publications Inc.

Ahn, C. M. and H. E. Thompson (1988). Jump-diffusion processes and term structure of interest rate. *The Journal of Finance XL111*, 155–174.

Aït-Sahalia, Y. (1996a). Non-parametric pricing of interest rate derivative securities. *Econometrica 64*, 527–560.

Aït-Sahalia, Y. (1996b). Testing continuous-time models of the spot interest rate. *Review of Financial Studies 9*, 385–426.

Aït-Sahalia, Y. (1998). Maximum likelihood estimation of discretely sampled diffusions: a closed-form approach. Pre-print.

Albert, J. H. and S. Chib (1993). Bayesian analysis of binary and polichotomous response data. *Journal of the American Statistical Association 88*, 669–679.

Andersen, T. G. (1994). Stochastic autoregressive volatility: a framework for volatility modelling. *Mathematical Finance 4*, 75–102.

Andersen, T. G. and J. Lund (1997). Estimating continuous-time stochastic volatility models of the short-term interest rate. *Journal of Econometrics 77*, 343–377.

Andersen, T. G. and B. Sørensen (1996). GMM estimation of a stochastic volatility model: A Monte Carlo study. *Journal of Business and Economic Statistics 14*, 328–352.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis* (Second ed.). New York: Wiley.

Arnold, L. (1974). *Stochastic Differential Equations: Theory and Applications.* New York: Wiley.

Baadsgaard, M., J. N. Nielsen, H. Spliid, H. Madsen, and M. Preisel (1997). Estimation in stochastic differential equations with a state dependent diffusion term. In Y. Sawaragi and S. Sagara (Eds.), *SYSID '97 - 11th IFAC Symposium on System Identification.*

Babbs, S. H. and N. Webber (1994/1999). A theory of the term structure with an official short rate. FORC preprint, University of Warwick.

Babbs, S. H. and N. Webber (1996). Term structure modelling under alternative regimes. In M. A. H. Dempster and S. R. Pliska (Eds.), *Mathematics of Derivative Securities*, Publications of the Isaac Newton Institute, pp. 394–422. Cambridge: Cambridge University Press.

Ball, C. A. and W. N. Torous (1999). The stochastic volatility of short-term interest rates. Forthcoming in Journal of Finance.

Bally, V. and D. Talay (1995). The law of the Euler scheme for stochastic differential equations: error analysis with Malliavin calculus. *Mathematics and Computers in Simulation 38*, 35–41.

Bally, V. and D. Talay (1996a). The law of the Euler scheme for stochastic differential equations. *Monte Carlo Methods and Applications 2*, 93–128.

Bally, V. and D. Talay (1996b). The law of the Euler scheme for stochastic differential equations (i): convergence rate of the distribution function. *Probability Theory and Related Fields 104*, 43–60.

Barndorff-Nielsen, O. E. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proc. Royal Society of London A 353*, 401–419.

Barndorff-Nielsen, O. E., P. Blæsild, J. L. Jensen, and M. Sørensen (1985). The fascination of sand. In A. C. Atkinson and S. E. Feinberg (Eds.), *A Celebration of Statistics: Centenary Volume of the ISI*, pp. 57–87. Heidelberg: Springer-Verlag.

Barndorff-Nielsen, O. E. and N. Shephard (1999). Non-Gaussian OU based models and some of their uses in financial economics. Economics discussion paper 1999-W9, Nuffield College, Oxford.

Basawa, I. V. and B. L. S. Rao (1980). *Statistical Inferences for Stochastic Processes.* Probability and Mathematical Statistics. London: Academic Press.

Baxendale, P. H. (1994). A stochastic Hopf bifurcation. *Probability Theory and Related Fields 9*, 581–616.

Bera, A. K. and M. L. Higgins (1993). ARCH models: properties, estimation and testing. *Journal of Economic Surveys 7*, 305–366.

Bertoin, J. (1996). *Lévy Processes.* Cambridge Tracts in Mathematics, 121. Cambridge: Cambridge University Press.

Besag, J., P. Green, D. Higdon, and K. Mengersen (1995). Bayesian computation and stochastic systems. *Statistical Science 10*, 3–66.

Bhattacharya, R. N. and E. C. Waymire (1990). *Stochastic Processes with Applications.* Probability and Mathematical Statistics. New York: John Wiley & Sons.

Biais, B., P. Hillion, and C. Spatt (1995). An empirical analysis of the limit order book and the order flow in the Paris Bourse. *Journal of Finance 50*, 1655–1689.

Bibby, B. M. (1994). Optimal combinations of martingale estimating functions for discretely observed diffusion processes. Technical report, Department of Theoretical Statistics, University of Aarhus. Research Report no. 298.

Bibby, B. M. and M. Sørensen (1995). Martingale estimating functions for discretely observed diffusion processes. *Bernoulli 1*, 17–39.

Bibby, B. M. and M. Sørensen (1997). A hyperbolic diffusion model for stock prices. *Finance and Stochastics 1*, 25–41.

Billingsley, P. (1961). *Statistical Inference for Markov Processes.* Chicago: Chicago University Press.

Billingsley, P. (1979). *Probability and Measure.* New York: Wiley.

Billio, M., A. Monfort, and C. P. Robert (1998). The simulated likelihood ratio method. Doc. Travail Crest, INSEE, Paris.

Biscay, R., J. C. Jimenez, J. J. Riera, and P. A. Valdes (1996). Local linearization method for the numerical solution of stochastic differential equations. Technical report, Cuban Neuroscience centre. Annals of Institute of Statistical Mathematics - in press.

Björk, T. (1996). Interest Rate Theory. CIME Lectures, Department of Finance, Stockholm School of Economics, Stockholm.

Björk, T., G. Di Masi, Y. Kabanov, and W. Runggaldier (1997). Towards a general theory of bond markets. *Finance and Stochastics 1*, 141–174.

Black, F. and M. Scholes (1973). The pricing of options and corporate liabilities. *Journal of Political Economy 81*, 637–654.

Bollerslev, T. (1986). Generalized ARCH. *Journal of Econometrics 31*, 307–327.

Bollerslev, T., R. Y. Chow, and K. F. Kroner (1992). ARCH modelling in finance: a review of the theory and empirical evidence. *Journal of Econometrics 52*, 5–59.

Bollerslev, T. and R. F. Engle (1993). Common persistence in conditional variances. *Econometrica 61*, 167–186.

Bollerslev, T., R. F. Engle, and D. B. Nelson (1994). ARCH models. *The Handbook of Econometrics 4*, 2959–3038.

Borodin, A. N. and P. Salminen (1996). *Handbook of Brownian Motion - Facts and Formulae.* Probability and its applications. Basel; Boston; Berlin: Birkhäuser Verlag.

Box, G. E. P. (1979). Strategy of scientific model building. In R. L. Launer and W. G. N (Eds.), *Robustness in Statistics.* New York: Academic Press.

Box, G. E. P. (1980). Sampling and Bayesian inference in scientific modeling and robustness (with discussion). *Journal of the Royal Statistical Society A 143*, 383–430.

Brennan, M. J. and E. S. Schwartz (1978). Finite-difference methods and jump processes arising in the pricing of contingent claims: a synthesis. *Journal of Financial and Quantitative Analysis 13*, 462–474.

Brennan, M. J. and E. S. Schwartz (1979). A continuous-time approach to the pricing of bonds. *Journal of Banking and Finance 3*, 133–155.

Brenner, R. J., R. H. Harjes, and K. F. Kroner (1994). Another look at alternative models of the short-term interest rate. Manuscript, Wells Fargo Nikko Investment Advisors, San Francisco, CA.

Broze, L., O. Scaillet, and J.-M. Zakaïan (1998). Quasi-indirect inference for diffusion processes. *Econometric Theory 14*, 161–186.

Campbell, J. Y. and A. S. Kyle (1993). Smart money, noise trading and stock price behaviour. *Review of Economic Studies 60*, 1–34.

Campbell, J. Y., A. W. Lo, and C. Mackinlay (1997). *The Econometrics of Financial Markets.* Princeton, New Jersey: Princeton University Press.

Carlin, P., A. Gelman, and R. M. Neal (1998). Markov chain Monte Carlo in practice: a roundtable discussion, with moderator: R. E. Kass. *The American Statistician, Statistical Practice 52*, 93–100.

Carter, C. K. and R. Kohn (1994). On Gibbs sampling for state space models. *Biometrika 81*, 541–553.

Chan, K., C. Karolyi, G. A. Longstaff, and A. B. Sanders (1992). An empirical comparison of alternative models of the short-term interest rate. *Journal of Finance 3*, 1209–1227.

Chan, K. S. (1993). Asymptotic behaviour of the Gibbs sampler. *Journal of the American Statistical Society 88*, 320–326.

Chan, K. S. and C. J. Geyer (1994). Discussion on Markov chains for exploring posterior distributions, (by L. Tierney). *Annals of Statistics 22*, 1747–1758.

Chen, R. R. and L. Scott (1993). Maximum likelihood estimation for a multi-factor general equilibrium model of the term structure of interest rates. *Journal of Fixed Income 3*, 14–31.

Chesney, M. and L. O. Scott (1989). Pricing European options: a comparison of the modified Black-Scholes model and a random variance model. *Journal of Financial and Qualitative Ananlysis 24*, 267–284.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association 90*, 1313–1321.

Chib, S. and E. Greenberg (1994). Bayes inference in regression models with ARMA (p, q) errors. *Journal of Econometrics 64*, 183–206.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis–Hastings algorithm. *The American Statistician 49*, 327–336.

Chib, S. and E. Greenberg (1996). Markov chain Monte Carlo simulation methods in econometrics. *Econometric Theory 12*, 409–431.

Chib, S., E. Greenberg, and R. Winkelmann (1998). Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics 86*, 33–54.

Chung, K. and R. Williams (1983). *An Introduction to Stochastic Integration*. Boston, MA: Birkhauser.

Ciesielski, Z. (1961). Hölder conditions for realizations of Gaussian processes. In *Translations of Mathematical Monographs*, Volume 99 of *American Mathematical Society*, pp. 4.3–413. ProvidenceRhode Island: AMS.

Clark, P. K. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica 41*, 135–155.

Cook, R. D. and S. Weisberg (1982). *Residuals and Influence in Regression*. Monographs on Statistics and Applied Probability. New York: Chapman and Hall.

Cootner, P. H. (1964). *The Random Character of Stock Market Prices*. Cambridge: MA: MIT Press.

Cox, D. R. and H. D. Miller (1968). *The Theory of Stochastic Processes*. New York: Wiley.

Cox, J. C., J. E. Ingersoll, and S. A. Ross (1985). A theory of the term structure of interest rates. *Econometrica 53*, 385–407.

Cox, J. C. and S. A. Ross (1976). The valuation of options for alternative stochastic processes. *Journal of Financial Econometrics 3*, 145–166.

Cramèr, H. (1946). *Mathematical Methods of Statistics*. New Jersey: Priceton University Press.

Crouch, E. A. C. and D. Spiegelman (1990). The evaluation of integrals of the form $\int f(t)\exp(-t^2)dt$: application to logistic normal models. *Journal of the American Statistical Association 85*, 464–469.

Dachunha-Castelle, D. and M. Duflo (1986). *Probability and Statistics*, Volume II. New York: Springer-Verlag.

Dacunha-Castelle, D. and D. Florens-Zmirou (1986). Estimation of the coefficients of a diffusion from discrete observations. *Stochastics 19*, 263–284.

Dagpunar, J. (1988). *Principles of Random Variate Generation*. Oxford: Clarendon Press.

Danielsson, J. (1994). Stochastic volatility in asset prices: estimation with simulated maximum likelihood. *Journal of Econometrics 61*, 375–400.

Das, S. R. (1996). Poisson-Gaussian processes and the bond markets. Manuscript, Harvard University.

Das, S. R. and S. Foresi (1996). Exact solutions for bond and option prices with systematic jump risk. *Review of Derivatives Research 1*, 7–24.

Dawid, A. P. (1982). The well-calibrated Bayesian (with discussion). *Journal of the American Statistical Association 77*, 605–613.

De Jong, P. and N. Shephard (1995). The simulation smoother for time series models. *Biometrika 82*(2), 339–350.

Dempster, M. A. H. and S. R. Pliska (Eds.) (1997). *Mathematics of Derivative Securities*. Publications of the Newton Institute. Cambridge: Cambridge University Press.

Dixit, A. (1993). *The Art of Smooth Pasting*. Switzerland: Harwood.

Dohnal, G. (1987). On estimating the diffusion coefficient. *Journal of Applied Probability 24*, 105–114.

Doob, J. L. (1953). *Stochastic Processes*. New York: John Wiley & Sons.

Doornik, J. A. (1996). *Ox: Object Oriented Matrix Programming, 1.10*. London: Chapman & Hall.

Doornik, J. A. and D. F. Hendry (1999). *Givewin: an Interface to Empirical Modelling*. Kent: Timberlake Consultants Ltd.

Dorogovtçev, A. Y. (1982). *The Theory of Estimates of the Parameters of Random Processes*. Kiev: Izd-vo pri. In Russian.

Drost, F. and T. H. E. Nijman (1993). Temporal aggregation of GARCH processes. *Econometrica 61*, 909–927.

Duffie, D. (1996). *Dynamic Asset Pricing Theory* (Second ed.). Princeton, New Jersey: Princeton University Press.

Duffie, D. and R. Kan (1996). A yield-factor model of interest rates. *Mathematical Finance 6*, 379–406.

Duffie, D. and K. J. Singleton (1993). Simulated moments estimation of Markov models of asset prices. *Econometrica 61*, 929–952.

Durbin, J. and S. J. Koopman (1997). Monte Carlo maximum likelihood approach for non-Gaussian state space models. *Biometrika 84*, 669–684.

Durbin, J. and S. J. Koopman (1998). Time series analysis for non-Gaussian observations based on state space models. Mimeo, London School of Economics.

Dyvbig, P. H. (1989). Bond and bond option pricing based on the current term structure. Manuscript, Washington University, St. Louis, M O.

Eberlein, E. and U. Keller (1995). Hyperbolic distributions in finance. *Bernoulli 1*(3), 281–299.

Elerian, O. (1998). A note on the existence of a closed-form conditional transition density for the Milstein scheme. Economics discussion paper 1998-W18, Nuffield College, Oxford.

Elerian, O., S. Chib, and N. Shephard (1998). Likelihood inference for discretely observed non-linear diffusions. Economics discussion paper 146, Nuffield College, Oxford.

Engle, R. F. (1982). ARCH with estimates of the variance of the United Kingdom inflation. *Econometrica 50*, 987–1007.

Engle, R. F. and T. Bollerslev (1986). Modelling the persistence of conditional variances. *Econometric Reviews 5*, 1–50.

Engle, R. F. and G. G. J. Lee (1996). Estimating diffusion models of stochastic volatility. In P. E. Rossi (Ed.), *Modelling Stock Market Volatility: Bridging the Gap to Continuous Time*, Chapter 11, pp. 333–356. California: Academic Press.

Engle, R. F. and J. R. Russell (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica 66*, 1127–1162.

Engle, R. F. and J. R. Russell (1999). Forecasting transaction rates: the autoregressive conditional duration model. National Bureau of Economic Research (NBER) Working paper: 4966, forthcoming in Econometrica.

Eraker, B. (1998). Markov chain Monte Carlo analysis of diffusion models with application to finance. HAE thesis, Norwegian School of Economics and Business Administration.

Fama, E. F. (1965). The behaviour of stock market prices. *Journal of Business 38*, 34–105.

Fama, E. F. (1970). Efficient capital markets: a review of theory and empirical work. *Journal of Finance 25*, 383–417.

Feller, W. (1951). Two singular diffusion problems. *Annals of Mathematics 54*, 173–182.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications* (Second ed.), Volume I. New York: John Wiley & Sons.

Feller, W. (1971). *An Introduction to Probability Theory and its Applications* (Second ed.), Volume II. New York: John Wiley & Sons.

Fisher, R. A. (Ed.) (1928). *The General Sampling Distribution of the Multiple Correlation Coefficient*, Volume 121(A), London. Royal Society of London. 654-673.

Florens-Zmirou, D. (1989). Approximate discrete-time schemes for statistics of diffusion processes. *Statistics 20*, 547–557.

Florens-Zmirou, D. (1993). On estimating the diffusion coefficient from discrete observations. *Journal of Applied Probability 30*, 790–804.

Frigessi, A., P. Di-Stefano, C. R. Hwang, and S. J. Sheu (1993). Convergence rates of the Gibbs sampler, the Metropolis algorithm and other single-state updating dynamics. *Journal of the Royal Statistical Society, B 55*(1), 205–219.

Gallant, A. R. and M. Davidien (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika 80*, 475–488.

Gallant, A. R. and J. R. Long (1997). Estimating stochastic differential equations efficiently by minimum chi-squared. *Biometrika 84*, 125–141.

Gallant, A. R. and G. Tauchen (1996). Which moments to match. *Econometric Theory 12*, 657–681.

Gallant, A. R. and G. Tauchen (1997a). Estimation of continuous models for stock returns and interest rates. Technical report, Duke University.

Gallant, A. R. and G. Tauchen (1997b, July). Reprojecting partially observed systems with application to interest rate diffusions. Technical report, Duke University.

Gallant, R., C. T. Hsu, and G. Tauchen (1998). Using high/low data to calibrate volatility diffusions and extract the forwrad integrated variance. .

Gamerman, D. and A. Smith (1996). Bayesian analysis of longitudinal data studies. *Bayesian Statistics 5*, 587–597.

Gelfand, A. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of American Statistical Association 85*, 398–409.

Gelman, A. (1995). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, and D. J. Spieghalter (Eds.), *In Markov Chain Monte Carlo in Practice*, pp. 131–143. London: Chapman & Hall.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (1995). *Bayesian Data Analysis* (First ed.). London: Chapman and Hall.

Gelman, A. and D. B. Rubin (1992a). Inference from iterative simulation using multiple sequences. *Statistical Science 7*(4), 457–472.

Gelman, A. and D. B. Rubin (1992b). A single series from the Gibbs sampler provides a false sense of security. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Simth (Eds.), *Bayesian Statistics*, Volume 4, pp. 625–632. Oxford: Oxford Universirt Press.

Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE transactions, PAMI 6*, 721–741.

Genon-Catalot, V. and J. Jacod (1993). On the estimation of the diffusion coefficient for multi-dimiensional diffusion processes. *Annales de l'Institut Henri Poincaré, Probabilités-Statistiques 29*, 119–151.

Gerlach, R., C. Carter, and R. Kohn (1999). Diagnostics for time series analysis. *Journal of Time Series Analysis 21*, 309–330. forthcoming.

Geweke, J. (1986). Modelling the persistence of conditional variances: a comment. *Econometric Reviews 5*, 57–61.

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica 57*, 1317–1339.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments (with discussion). In J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics*, Volume 4, pp. 169–193. Oxford: Oxford University Press.

Geweke, J. (1994). Comment on Bayesian analysis of stochastic volatility models by Jacquier, Polson and Rossi. *Journal of Business and Economic Statistics 12*, 397–399.

Geweke, J. (1996). Posterior simulators in econometrics. In D. M. Kreps and K. F. Wallis (Eds.), *Advances in Economics and Econometrics: Theory and Applications, 7th World Congress*, Volume 3, pp. 128–165. Cambridge: Cambridge University Press.

Geyer, A. L. J. and S. Pichler (1995). A state-space approach to estimate and test multi-factor Cox-Ingersoll-Ross models. Working paper, University of Economics, Vienna.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo. *Statistical Science 7*(4), 473–511.

Geyer, C. J. (1999). Likelihood inference for spatial point processes. In O. E. Barndorff-Nielsen, W. S. Kendall, and M. N. M. Lieshout (Eds.), *Current Trends in Stochastic Geometry and Applications*, pp. 141–172. London: Chapman & Hall.

Ghysels, E., A. C. Harvey, and E. Renault (1996). Stochastic volatility. In C. R. Rao and G. S. Maddala (Eds.), *Statistical Methods in Finance*. Amsterdam: North-Holland.

Ghysels, E. and J. Jasicek (1996). Stochastic volatility and time deformation: an application to trading volume and leverage effects. Mimeo, University of Montreal.

Gilks, W. R., N. G. Best, and K. K. C. Tan (1995). Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society C 44*, 155–173.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice.* Chapman & Hall.

Gilks, W. R. and G. O. Roberts (1995). Strategies for improving MCMC. In W. R. Gilks, S. Richardson, and D. J. Spieghalter (Eds.), *Markov chain Monte Carlo in Practice*, pp. 1–19. London: Chapman & Hall.

Gilks, W. R. and P. Wild (1992). Adaptive rejection sampling for Gibbs sampling. *Applied Statistics 41*, 337–348.

Godambe, V. P. and C. C. Heyde (1987). Quasi-likelihood and optimal estimation. *International Statistical Review 55*, 231–244.

Goldfeld, S., R. Quandt, and H. Trotter (1966). Maximization by quadratic hill climbing. *Econometrica 34*, 541–551.

Golub, G. H. and C. F. van Loan (1996). *Matrix Computations* (Third ed.). Baltimore: Johns Hopkins.

Gourieroux, C. and A. Monfort (1996). *Simulation-Based Econometric Methods.* Oxford University Press.

Gourieroux, C., A. Monfort, and E. Renault (1993). Indirect inference. *Journal of Applied Econometrics 3*, S85–S118.

Gradshteyn, I. S. and I. M. Ryzhik (1965). *Table of Integrals, Series and Products.* San Diego: Academic Press.

Greenstadt, J. (1967). On the relative efficiencies of gradient methods. *Mathematics of Computation* , 360–367.

Hansen, L. and J. A. Scheinkman (1995). Back to the future: generating moment implications for continuous-time Markov processes. *Econometrica 63*, 767–804.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*, 1029–1054.

Harvey, A. C. (1981). *The Econometric Analysis of Time Series* (Second ed.). New York: Philip Allan.

Harvey, A. C. (1993). *Time Series Models* (Second ed.). New York: Harvester Wheatsheaf.

Harvey, A. C., E. Ruiz, and N. Shephard (1994). Multivariate stochastic variance models. *Review of Economic Studies 61*, 247–264.

Harvey, A. C. and N. Shephard (1996). Estimation of an asymmetric stochastic volatility model for asset returns. *Journal of Business and Economic Statistics 14*, 429–434.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97–109.

Hendry, D. F. (1995). *Dynamic Econometrics* (First ed.). Oxford: Oxford University Press.

Heston, S. L. (1993). A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies 6*, 327–343.

Hills, S. E. and A. F. M. Smith (1992). Parameterization issues in Bayesian inference, (with discussion). In J. M. Bernado, J. Berger, A. P. Dawid, and S. A. F. M (Eds.), *Bayesian Statistics*, Volume 4, pp. 227–246. Oxford: Oxford University Press.

Holden, H., B. Øksendal, J. Ubøe, and T. Zhang (1996). *Stochastic Partial Differential Equations: A Modeling, White Noise Functional Approach.* Probability and its Applications. Berlin: Birkhäuser.

Honoré, P. (1998). Pitfalls in estimating jump-diffusion models. Working paper, Aarhus School of Business.

Hull, J. and A. White (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance 42*, 281–300.

Hull, J. and A. White (1988). An analysis of the bias in option pricing caused by a stochastic volatility. *Advances in Futures and Options Research 3*, 29–61.

Hull, J. and A. White (1990). Pricing interest-rate derivative securities. *Review of Financial Studies 3*, 573–592.

Hull, J. C. (1989). *Options, Futures and Other Derivatives* (third ed.). Prentice Hall.

Hutton, J. E. and P. I. Nelson (1986). Quasi-likelihood estimation for semi-martingales. *Stochastic Processes and Applications 22*, 245–257.

Ikeda, N. and S. Watanabe (1981). *Stochastic Differential Equations and Diffusion Processes*, Volume 24. North-Holland.

Itô, K. (1942). Differential equations determining Markov processes. *Zenkoku Shijō Sūgaku Danwakai 244*, 1352–1400. (in Japanese).

Itô, K. (1951). On stochastic differential equations. *Memoirs of American Mathematical Society 5*, 1–51.

Itô, K. (1987). *Kiyoshi Itô Selected Papers*. New York: Springer-Verlag.

Itô, K. and H. P. Mckean (1964). *Diffusion Processes and their Sample Paths*. New York: Academic Press.

Jacod, J. and A. N. Shiryayev (1987). *Limit Theorems for Stochastic Processes*. Berlin: Springer-Verlag.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994, October). Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics 12*(4), 11–26.

Johnson, H. and D. Shanno (1987). Option pricing when the variance is changing. *Journal of Financial and Quantitative Analysis 22*, 143–151.

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous Univariate Distributions* (Second ed.), Volume 2. New York: John Wiley.

Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*, Volume 9 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

Kalos, M. and P. Whitlock (1986). *Monte Carlo Methods*. Wiley.

Kanagawa, S. (1995). Error estimations for the Euler-Maruyama approximate solutions of stochastic differential equations. *Monte Carlo methods and applications 3*, 165–171.

Karatzas, I. and S. E. Shreve (1991). *Brownian Motion and Stochastic Calculus* (Second ed.). New York: Springer-Verlag.

Karlin, S. and H. M. Taylor (1975). *A First Course in Stochastic Processes* (Second ed.). San Diego, California: Academic Press.

Kessler, M. (1995). Martingale estimating functions for a markov chain. Preprint, Laboratoire de Probabilités, Université Paris VI.

Kessler, M. (1996). Simple and explicit estimating functions for a discretely observed diffusion process. Research Report No. 336, Department of Theoretical Statistics, University of Aarhus.

Kessler, M. and M. Sørensen (1999). Estimating equations based on eigenfunctions for a discretely observed diffusion process. *Bernoulli 5*, 299–314.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic volatility: likelihood inference and comparison with ARCH models. *Review of Economic Studies 65*, 361–393.

Kloeden, P. E. and E. Platen (1992). *Numerical Solutions to Stochastic Differential Equations*. New York: Springer.

Kloeden, P. E., E. Platen, and H. Schurz (1991). *Numerical Solution of SDE Through Computer Experiments*. Berlin: Springer-Verlag.

Kloek, T. and H. K. van Dijk (1978). Bayesian estimates of equation system parameters: an application on integration by Monte Carlo. *Econometrica 46*, 1–20.

Knorr-Held, L. (1996, July). Conditional prior proposals in dynamic models. Working paper.

Koedijk, K. G., F. G. J. A. Nissen, P. C. Schotman, and C. C. P. Wolff (1994). The dynamics of short-term interest rate volatility reconsidered. Manuscript, Limburg Institute of Financial Economics, Maastricht.

Kohatsu-Higa, A. and S. Ogawa (1997). Weak rate of convergence for an Euler scheme of nonlinear SDE's. *Monte Carlo Methods and Applications 3*, 327–345.

Kolmogorov, A. N. (1931). Über die analytischen methoden in der Wahrscheinlichkeitsrechnung. *Mathematische Annalen 104*, 415–458.

Koopman, S. J., N. Shephard, and J. A. Doornik (1999). Statistical algorithms for models in state space using SSF Pack 2.2. Forthcoming in Econometrics Journal.

Küchler, U., K. Neumann, M. Sørensen, and A. Streller (1994). Stock retuens and hyperbolic distributions. Discussion paper 23, Universität zu Berlin, Sonderforschungsbereich 373, Humboldt.

Lamberton, D. and B. Lapeyre (1996). *Introduction to Stochastic Calculus Applied to Finance*. London: Chapman & Hall.

Laroque, G. and B. Salanie (1989). Estimation of multi-market fix-price models: an application of psuedo maximum likelihood methods. *Econometrica 57*, 831–860.

Le Breton, A. (1976). On continuous and discrete sampling for parameter estimation in diffusion type processes. *Mathematical Programming Study 5*, 124–144.

Lévy, P. (1937). *Thèorie de l'Addition des Variables Alèatoires* (Second ed.). Paris: Gauthier-Villars.

Liptser, R. S. and A. N. Shiryayev (1977). *Statistics of Random Processes I: general theory*. New York: Springer-Verlag.

Liu, J. S., W. H. Wong, and A. Kong (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika 81*, 27–40.

Lo, A. (1988). Maximum likelihood estimation of generalized Itô processes with discretely sampled data. *Econometric Theory 4*, 231–247.

Loève, M. (1977). *Probability Theory* (Fourth ed.), Volume I. New York: Springer-Verlag.

Loève, M. (1978). *Probability Theory* (Fourth ed.), Volume II. New York: Springer-Verlag.

Longstaff, F. A. and E. S. Schwartz (1992). Interest rate volatility and the term structure. *Journal of Finance 47*, 1259–1282.

Lütkepohl (1996). *Handbook of Matrices*. Chichester: John Wiley and Sons.

Madsen, H., J. N. Nielsen, and M. Baadsgaard (1998). Statistics in Finance - Lecture Notes. Lyngby.

Magnus, J. R. and H. Neudecker (1988). *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Chichester: John Wiley & Sons.

Mandelbrot, B. (1963). The variation of certian speculative prices. *Journal of Business 36*, 394–419.

Maybeck, P. S. (1982). *Stochastic Models, Estimation and Control*, Volume II of *Mathematics in Science and Engineering, 141*. London: Academic Press.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica 57*, 995–1026.

Meddahi, N. and E. Renault (1996, November). Aggregations and marginalization of GARCH and stochastic volatility models. Discussion paper, GREMAQ, Université of Toulouse.

Meddahi, N. and E. Renault (1999a). Conditioning information in volatility models. Working paper, Université de Montreal; revision of a part of Meddahi and Renault (1996), "Aggregations and marginalization of GARCH and stochastic volatility models".

Meddahi, N. and E. Renault (1999b, May). Temporal aggregation of volatility models. Revision of a part of Meddahi and Renault (1996), "Aggregations and marginalization of GARCH and stochastic volatility models".

Melino, A. and S. M. Turnbull (1990). Pricing foreign currency options with stochastic volatility. *Journal of Econometrics 45*, 239–265.

Merton, R. (1976). Option pricing when the underlying stock returns are discontinuous. *Journal of Financial Economics 5*, 125–144.

Merton, R. (1990). *Continuous-Time Finance*. Cambridge: Blackwell.

Metropolis, N., A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087–1092.

Meyer, P. A. (1975). Sur la démonstration de prévisibilité de Chung et Walsh. In *Séminaire de Probabilités XV*, Volume 465 of *Lecture Notes in Mathematics*, pp. 530–533. Berlin: Springer.

Meyn, S. P. and R. L. Tweedie (1993). *Markoc Chains and Stochastic Stability*. London: Springer-Verlag.

Meyn, S. P. and R. L. Tweedie (1994). Computable bounds for convergence rates of Markov chains. *Annals of Applied Probability 4*, 124–148.

Milstein, G. N. (1974). Approximate integration of stochastic differential equations. *Theory of Probability and its Applications 19*, 557–562.

Milstein, G. N. (1978). A method of second-order accuracy integration of stochastic differential equations. *Theory of Probability and its Applications 23*, 396–401.

Mitchell, A. R. and D. F. Griffiths (1980). *The Finite Difference Method in Partial Differential Equations*. Chichester: Wiley & Sons.

Monfort, A. and H. K. van Dijk (1995). Simulation-based econometrics. In H. van Dijk, A. Monfort, and B. W. Brown (Eds.), *Econometric Inference Using Simulation Techniques*, Chichester, pp. 1–20. John Wiley and Sons.

Mood, A. M., F. A. Graybill, and D. C. Boes (1974). *Introduction to the Theory of Statistics* (Second ed.). New York: McGraw Hill.

Müller, P. (1993). A genetic appraoch to posterior integration and Gibbs sampling. Technical report, Institution of Statistics and Decision Sciences, Duke University.

Musiela, M. and M. Rutkowski (1997). *Martingale Methods in Financial Modelling* (First ed.). Number 36 in Applications of Mathematics, Stochastic Modelling and Applied Probability. Berlin: Springer-Verlag.

Neal, R. (1993). Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Nelson, D. B. (1990). ARCH models as diffusion approximations. *Journal of Econometrics 45*, 7–38.

Nelson, D. B. (1991). Conditional heteroscedasticity in asset pricing: a new approach. *Econometrica 59*, 347–370.

Nelson, D. B. (1995). ARCH models as diffusion approximations. In R. F. Engle (Ed.), *ARCH: Selected Readings*, Advanced texts in econometrics, Chapter 9, pp. 193–220. New York: Oxford University Press.

Nielsen, J. N. (1998, September). Estimation in continuous-time stochastic volatility models using non-linear filters. Presented at the Workshop on Statistical Inference for Stochastic Processes - Theoretical and Computational Aspects, MaPhySto, Aarhus.

Nielsen, J. N., H. Madsen, H. Melgaard, and M. Baadsgaard (1998). Estimation of embedded parameters in stochastic differential equations using discrete-time measurements. Submitted.

Nielsen, J. N., M. Vestergaard, and H. Madsen (1998, September). Non-linear filtering of univariate stochastic volatility models. Presented at the Workshop on Statistical Inference for Stochastic Processes - Theoretical and Computational Aspects, MaPhySto, Aarhus.

Nielsen, L. T. and J. Saá-Requejo (1993). Exchange rate and term structure dynamics and the pricing of derivative securities. Working paper, INSEAD.

Nummelin, E. (1984). *General Irreducible Makov chains and non-negative operators*. Cambridge Uinversity Press.

Øksendal, B. (1995). *Stochastic Differential Equations: An Introduction with Applications* (Fourth ed.). Berlin: Springer-Verlag.

Pantula, S. G. (1986). Modelling the persistence of conditional variances: a comment. *Econometric Reviews 5*, 71–73.

Pardoux, E. and D. Talay (1985). Discretization and simulation of stochastic differential equations. *Acta Applicandae Mathematicae 3*, 23–47.

Pedersen, A. R. (1994a). Quasi-likelihood inference for discretely observed diffusion processes. Research Report No. 295, Department of Theoretical Statistics, University of Aarhus.

Pedersen, A. R. (1994b). Uniform residuals for discretely observed diffusions. Unpublished paper: Department of Statistics, University of Aarhus.

Pedersen, A. R. (1995a). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion process. *Bernoulli 1*, 257–279.

Pedersen, A. R. (1995b). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics 22*, 55–71.

Peskun, P. H. (1973). Optimum Monte Carlo sampling using Markov chains. *Biometrika 60*, 607–612.

Pitt, M. K. and N. Shephard (1999). Analytic convergence rates and parameterisation issues for the Gibbs sampler applied to state space models. *Journal of Time Series Analysis 21*, 63–85.

Pitt, M. K. and S. G. Walker (1998, November). Marginal construction of stationary time series with application to volatility models. Working paper, Department of Mathematics, Imperial College, London.

Polson, N. G. (1996). Convergence of Markov chain Monte Carlo algorithms (with discussion). In J. M. e. a. Bernado (Ed.), *Bayesian Statistics 5*, Oxford, pp. 297–321. Oxford University Press.

Poulsen, R. (1999, March). Approximate maximum likelihood estimation of discretely observed diffusion processes. Working Paper Series 29, Cantre for Ananlytical Finance, University of Aarhus.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. London: Academic Press.

Protter, P. (1995). *Stochastic Integration and Differential Equations: A New Approach*. Springer.

Protter, P. and D. Talay (1997). The Euler scheme for Lévy stochastic differential equations. *Annals of Probability 25*, 393–423.

Raftery, A. E. (1995). Hypothesis testing and model selection. In W. R. Gilks, S. Richardson, and D. J. Spieghalter (Eds.), *In Markov Chain Monte Carlo in Practice*, pp. 163–187. London: Chapman & Hall.

Raftery, A. E. and S. M. Lewis (1992). One long run with diagnostics: implementation strategies for MCMC. *Statistical Science 7*, 493–497.

Raftery, A. E. and S. M. Lewis (1995a). How many iterations in the Gibbs sampler? In W. R. Gilks, S. Richardson, and S. D. J (Eds.), *In Markov chain Monte Carlo in practice*, pp. 765–776. London: Chapman & Hall.

Raftery, A. E. and S. M. Lewis (1995b). Implementing MCMC. In W. R. Gilks, S. Richardson, and S. D. J (Eds.), *In Markov chain Monte Carlo in practice*, pp. 115–140. London: Chapman & Hall.

Revuz, D. (1975). *Markov Chains*. Holland: North-Holland.

Revuz, D. and M. Yor (1991). *Continuous Martingales and Brownian motion*. Berlin Heidelberg New York: Springer.

Richtmyer, R. D. and K. W. Morton (1967). *Difference Methods for Initial-value Problems* (Second ed.), Volume 4 of *Interscience Tracts in Pure and Applied Mathematics, no. 4*. New York: Interscience Publishers John Wiley & Sons.

Ripley, B. D. (1987). *Stochastic Simulation* (First ed.). New York: John Wiley.

Roberts, G. O. (1992). Discussion on parameterization issues in Bayesian inference (by S. E. Hills and A. F. M. Smith). In J. M. Bernado, J. Berger, A. P. Dawid, and S. A. F. M (Eds.), *Bayesian Statistics*, Volume 4, pp. 241. Oxford: Oxford University Press.

Roberts, G. O. (1995). Markov chain concepts related to sampling algorithms. In W. R. Gilks, S. Richardson, and D. J. Spieghalter (Eds.), *In Markov Chain Monte Carlo in Practice*, pp. 45–57. London: Chapman & Hall.

Roberts, G. O. and N. G. Polson (1994). A note on the geometric convergence of the Gibbs sampler. *Journal of the Royal Statistical Society, Series B 56*, 377–384.

Roberts, G. O. and S. K. Sahu (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society B 59*, 291–317.

Roberts, G. O. and O. Stramer (1999, April). On inference for partially observed non-linear diffusion models using the Metropolis–Hastings algorithm. Submitted for publication.

Roberts, G. O. and R. L. Tweedie (1994). Geometric convergence of central limit theorems, for multidimensional Hastings and Metropolis–Hastings algorithms. Technical Report 94-9, University of Cambridge.

Rogers, L. and D. Williams (1994). *Diffusions, Markov Processes and Martingales, Vol. 1, Itô Calculus* (Second ed.). New York: John Wiley and Sons.

Rogers, L. and D. Williams (1996). *Diffusions, Markov Processes and Martingales, Vol. 2, Foundations* (Second ed.). New York: John Wiley and Sons.

Rogers, L. C. G. and D. Talay (Eds.) (1997). *Numerical Methods in Finance.* Publications of the Newton Institute. Cambridge: Cambridge University Press.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Annals of Mathematical Statistics 23*, 470–472.

Rossi, P. E. (Ed.) (1996). *Modelling Stock Market Volatility: Bridging the Gap to Continuous Time.* San Diego, California: Academic Press.

Ruanaidh, J. J. K. and W. J. Fitzgerald (1996). *Numerical Bayesian Methods Applied to Signal Processing.* Statistics and Computing. New York: Springer-Verlag.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method* (First ed.). New York: John Wiley.

Rydberg, T. H. (1997). Why financial data are interesting to statisticians. research report 384, Department of theoretical statistics, University of Aarhus.

Rydberg, T. H. (1999). Generalized hyperbolic diffusion processes with applications towards finance. *Mathematical Finance 9*, 183–201.

Saá-Requejo, J. (1993). The dynamics and the term structure of risk premia in the foreign exchange markets. Working paper, University of Chicago.

Sandmann, G. and S. J. Koopman (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *Journal of Econometrics 87*, 271–301.

Schervish, M. J. and B. P. Carlin (1992). On the convergence of successive substitution sampling. *Journal of Computation and Graphical Statistics 1*, 111–127.

Schwert, G. W. (1989). Why does stock market volatility change over time? *Journal of Finance 44*, 1115–1153.

Scott, L. O. (1987). Option pricing when the variance changes randomly: theory, estimation and an application. *Journal of Financial and Quantitative Analysis 26*, 459–438.

Scott, L. O. (1991). Random variance option pricing: empirical tests of the model and delta-sigma hedging. *Advances in Futures and Options Research 5*, 113–135.

Shanmugan, S. and A. Breipohl (1988). *Random Signals: Detection, Estimation and Data Ananlysis.* John Wiley and Sons.

Shephard, N. (1994). Partial non-Gaussian state space. *Biometrika 81*, 115–131.

Shephard, N. (1996). Statistical aspects of ARCH and stochastic volatility. In D. R. Cox, O. E. Barndorff-Nielsen, and D. V. Hinkely (Eds.), *Time Series Models in Econometrics, Finance and Other Fields*, pp. 1–67. London: Chapman and Hall.

Shephard, N. and S. Kim (1994). Comment on Bayesian analysis of stochastic volatility models by Jacquier, Polson and Rossi. *Journal of Business and Economic Statistics 11*, 406–410.

Shephard, N. and M. K. Pitt (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika 84*, 653–667.

Shiryayev, A. N. (1996). *Probability* (Second ed.), Volume 95 of *Graduate Texts in Mathematics*. New York: Springer.

Shoji, I. and T. Ozaki (1997). Comparitive study of estimation methods for continuous-time stochastic processes. *Journal of Time Series Analysis 18*(5), 485–506.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Smith, A. (1993). Estimating non-linear time series models using simulated vector autoregressions. *Journal of Applied Econometrics 8*, 63–84.

Smith, A. F. M. (1991). Bayesian computational methods. *Philisophical Transactions of the Royal Society of London, A 337*, 1079–1102.

Smith, A. F. M. and G. O. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, B 55*(1), 3–23.

Smith, A. F. M., A. M. Skene, J. E. H. Shaw, J. C. Naylor, and M. Dransfield (1985). The implementation of the Bayesian paradigm. *Communications in Statistics - Theory and Methods 14*, 1079–1102.

Smith, G. (1985a). *Numerical Solution of Partial Differential Equations: Finite-Difference Methods* (Third ed.). Oxford: Oxford University Press.

Smith, J. Q. (1985b). Diagnostic checks of non-standard time series models. *Journal of Forecasting 4*, 283–291.

Sørensen, M. (1997). Estimating functions for discretely observed diffusions: a review. In I. V. Basawa, V. P. Godambe, and R. L. Taylor (Eds.), *Selected Proceedings of the Symposium on Estimating Functions*, Monograph Series, pp. 305–325. IMS Lecture Notes.

Stein, E. M. and J. C. Stein (1991). Stock price distributions with stochastic volatility: an analytic approach. *Review of Financial Studies 4*, 727–752.

Stoyanov (1984). Problems of estimation in continous-discrete models. In *Proceedings of the 7th Conference on Probability Theory, Brazov*, Bucharest. Iosifescu Ed.

Strikwerda, J. (1990). *Finite Difference Schemes and Partial Differential Equations*. The Wadsworth & Brooks/Cole Mathematics Series. Pacific Grove, California: Wadsworth & Brooks/Cole.

Strook, W. S. and S. R. S. Varadhan (1979). *Multidimensional Diffusion Processes*. Berlin: Springer-Verlag. Corrected second printing, 1997.

Talay, D. (1995). Simulation and numerical analysis of stochastic differential systems: a review. In P. Krée and W. Wedig (Eds.), *Probabilistic Methods in Applied Physics*, Volume 451, Chapter 3, pp. 54–96. Springer-Verlag.

Talay, D. and L. Tubaro (1990). Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Analysis and Applications 8*, 94–120.

Tanner, M. A. (1996). *Tools for Statistical Inference: Methods for Exploration of Posterior Distributions and Likelihood Functions* (Third ed.). New York: Springer-Verlag.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of American Statistical Association 82*, 528–550.

Tauchen, G. E. (1997). New minimum chi-squared methods in empirical finance. In D. M. Kreps and K. F. Wallis (Eds.), *Advances in Economics and Econometrics: Theory and Applications, 7th World Congress*, Volume 3, Chapter 9, pp. 279–317. Cambridge: Cambridge University Press.

Tauchen, G. E. and M. Pitts (1983). The price variability-volume relationship on speculative markets. *Econometrica 51*, 485–505.

Taylor, S. (1986). *Modelling Financial Time Series*, Volume Chichester. John Wiley & Sons.

Taylor, S. J. (1982). Financial returns modelled by the product of two stochastic processes: a study of the daily sugar prices 1961-75. In O. D. Anderson (Ed.), *Time Series Ananlysis: Theory and Practice*, Volume I. Amsterdam: North-Holland.

Taylor, S. J. (1994). Modelling stochastic volatility: a review and comparative study. *Mathematical Finance 4*, 183–204.

Tierney, L. (1994). Markov chains for exploring the posterior distributions, (with discussion). *Annals of Statistics 22*, 1701–1762.

Tierney, L. (1995). Introduction to general state space Markov chain theory. In W. R. Gilks, S. Richardson, and D. J. Spieghalter (Eds.), *Markov Chain Monte Carlo in Practice*, pp. 59–74. London: Chapman & Hall.

van Dijk, H. K. and T. Kloek (1980). Further experience in Bayesian anaylsis using Monte Carlo integration. *Journal of Econometrics 14*, 307–328.

Vasicek, O. (1977). An equilibrium characterisation of the term structure. *Journal of Financial Economics 5*, 177–188.

Vines, S. K., W. R. Gilks, and P. Wild (1994). Fitting multiple random effects models. Technical report, Biostatistics unit-MRC, Cambridge, UK.

von Neumann, J. (1951). Various techniques in connection with random digits. *Applied Mathematics Series 12*, 36–38.

Wakefield, J. C. (1993). Discussion on the meeting on the Gibbs sampler and other Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, B 55*, 56–57.

Walker, A. M. (1969). On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society 31*, 80–88.

Wang, N. and J. M. Robins (1998). Large-sample theory for parametric multiple integration procedures. *Biometrika 85*, 935–948.

Watanabe, S. (1984). *On Stochastic Differential Equations and Malliavin Calculus*. Number 73 in Lectures on Mathematics and Physics, Eds. S. Raghavan and B. V. Sreekantan. Institute of Fundamental Research: TATA.

Watson, G. N. (1944). *A Treatise on the Theory of Bessel Functions*. Cambridge: Cambridge University Press.

Wiener, N. (1923). Differential space. *Journal of Mathematical Physics 2*, 131–174.

Wiggins, J. B. (1987). Option values under stochastic volatility. *Journal of Financial Economics 19*, 351–372.

Wilmott, P., J. Dewynne, and S. Howison (1993). *Option Pricing: Mathematical Models and Computation.* Oxford: Oxford University Press.

Wilmott, P., S. Howison, and J. Dewynne (1995). *The Mathematics of Financial Derivatives.* Cambridge: Cambridge University Press.

Wong, C. K. J. (1999). New methods for performing efficient inference for linear and log-normal volatility models. M.Phil. thesis, University of Oxford.

Yamada, T. and S. Watanabe (1971). On the uniqueness of solutions of stochastic differential equations. *Journal of Mathematics, Kyoto University 11*, 155–167.

Yoshida, N. (1992). Estimation for diffusion processes from discrete observations. *Journal of Multivariate Analysis 41*, 220–242.

Zellner, A. and P. E. Rossi (1984). Bayesian analysis of dichotomous quantal response models. *Journal of Econometrics 25*, 365–393.