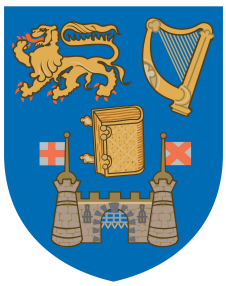
A New Method to Implement Bayesian Inference on Stochastic Differential Equation Models.

A Thesis presented for the degree of
Doctor of Philosophy



School of Computer Science and Statistics

Trinity College Dublin, Ireland
February 2011

Chaitanya Joshi

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where otherwise stated, the work described herein has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Chaitanya Joshi.

Chaitanya Joshi

Trinity College Dublin,

February 2011.

Summary

Stochastic differential equations (SDEs) are widely used to model numerous real-life phenomena. However, transition densities of most of the SDE models used in practice are not known, making both likelihood based and Bayesian inference difficult. Methods for Bayesian inference have mainly relied on MCMC based methods which are computationally expensive. There is a need to develop a computationally efficient method which will provide accurate inference.

This thesis introduces a new approach to approximate Bayesian inference for SDE models. This approach is ***not MCMC based*** and aims to provide a more efficient option for Bayesian inference on SDE models. This research problem was motivated by a civil engineering problem of modeling the force exerted by vehicles on the road surface as they traverse it.

Proposed here two new methods to implement this approach. These methods have been named as the Gaussian Modified Bridge Approximation (GaMBA) and its extension GaMBA- Importance sampling (GaMBA-I). This thesis provides an easy to use algorithm for both these methods, discusses their consistency properties, describes examples where these methods provide efficient inference and also illustrates situations where these methods would not yield efficient and accurate inference.

To illustrate how GaMBA-I could be used to model complex real life processes, this research attempts to model the dynamic force exerted by the vehicles on the road surface using SDE models. An SDE model based on one of the existing differential

equation models was used to fit a simulated force data using GaMBA-I. This was considered as a 'proof of concept' work to investigate if the SDE modeling of this problem is feasible.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Prof. Simon Wilson. He not only granted me the freedom to identify this research problem, but also provided invaluable guidance and support to persue it. He encouraged me to attend various international meetings, workshops and learning opportunities; the inputs gathered from which have immensely helped me in completing this work. He has always been helpful and patient and always had time to answer every query.

This research has been funded by Science Foundation Ireland, Research Frontiers Programme, contract no. $07.RFP.MATF.139$. This project was in collaboration with Prof. Eugene O'Brien of University College Dublin. I am grateful to him and his Ph.D. students Niall Harris and Rahim Taheri for their inputs and support.

A special ***Thank you*** to everyone in the (erstwhile) Department of Statistics for always being helpful and kind and for the interesting chats in the *coffee room*! Thanks to all the post-docs and post-grads, both past and present, for the numerous chats, discussions and the fun times.

Finally, this work wouldn't have been possible without the constant support and encouragement provided by my loving wife and dear parents. I can't thank them enough!

Chaitanya Joshi

Trinity College Dublin,

February 2011.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis is about Bayesian inference methods for stochastic differential equation (SDE) models. Specifically, it introduces a new approach to approximate Bayesian inference for SDE models. This research problem was motivated by a civil engineering problem of modeling the force exerted by vehicles on the road surface as they traverse.

Although motivated by an application, the methodologies developed in this thesis are for the general problem of Bayesian modeling for SDE models. The emphasis throughout has been in the context of the general statistical inference problem, and for the most part the application does not even need a mention.

1.1 Motivation

1.1.1 Motivation for the Applied Research

This research was in collaboration with with Prof. Eugene O'Brien of the School of Architecture Landscape and Civil Engineering at University College Dublin. The main

motivation of this project was to develop a better model to understand road degradation. The principal factor for road degradation is the 'road-vehicle interaction', which essentially refers to the forces exerted by the vehicles on the road surface as a result of the excitation caused by the surface.

In the engineering literature, differential equation models based on Newton's second law have been used to capture this interaction. However, it is extremely difficult to correctly capture all the dynamics in these models, and the models which can actually be used in practice only manage to capture a majority of these dynamics. These models assume a much simplified system and their solutions (often only the numerical solutions are possible) are derived using what is known as the 'finite element method', which essentially amounts to discretising the continuous space. In addition to these modeling constraints, the data available is sparce. The force is typically measured by fitting sensors within the road surface on a specially constructed patch of road. These sensors are typically placed every 1.5 meters, or so. Thus, the continuous process (of forces exerted by the vehicles) is only observed at a few discrete time points.

It was therefore thought that modeling this force using an SDE model might actually help capture the dynamics of the system more succinctly. Such a model could be built using one of the existing differential equation models.

1.1.2 Motivation for the Statistical Methodology Research

An SDE model can intuitively be understood as an extension of a differential equation model which incorporates randomness driven by the Weiner process. While the solution of a differential equation is a deterministic function, the solution of an SDE is a

continuous stochastic process known as the *diffusion process*. A diffusion process is a continuous time Markov process whose behaviour is governed by its transition density. This density is in turn governed by the values of the parameters in the SDE model. Statistical interest in the SDE models centres around the inference on these parameters.

SDE based models have become an increasingly popular choice for modeling real life processes because of their inherent incorporation of uncertainty. Though SDE models are an attractive modeling choice, closed form solution to many SDEs used in practice are not known. In fact, except for a few standard SDE models, the transition density for most others is not known. Statistical inference for SDE models is therefore not straightforward.

MCMC based methods have been the most widely used tool for Bayesian inference in the statistical community for about twenty years. These methods have very attractive mathematical properties and can be applied to many types of models. But MCMC methods can also be computationally very expensive for complex models and computational limitations of the time have often constrained the type of model being used for a particular problem.

Bayesian inference for SDE models has been centered around MCMC based methods. In fact, due to the mathematical properties of certain diffusion processes, implementation of MCMC based methods on SDE models is particularly tricky (let it be called the 'dependency problem' for now) and time consuming. This 'dependency problem', along with the ever increasing computational ability, has meant that Bayesian inference for SDE models has been a very active area of research for the last ten years.

Most of this research effort has concentrated around developing MCMC based methods which get around the 'dependency problem'. As a result, these methods are complicated to implement, computationally expensive and their use constraints the type of SDE model that can be used in practice. Thus, there is a need to develop computationally cheaper methods for Bayesian inference on SDE models.

1.2 Research Contributions

Following are the main research contributions of this thesis.

- This thesis explores a new approach to approximate Bayesian inference on SDE models. This approach is ***not MCMC based*** and is inspired from the work of Rue et al. (2009) on the Integrated Nested Laplace Approximation (INLA) for Gaussian Markov Random Field (GMRF) models. This thesis introduces two new methods to implement this approach. These methods have been named as the Gaussian Modified Bridge Approximation (GaMBA) and its extension GaMBA-Importance sampling (GaMBA-I). This thesis provides an easy to use algorithm for both these methods, discusses their consistency properties, describes examples where these methods provide efficient inference and also illustrates situations where these methods would not yield efficient and accurate inference.

- GaMBA provides a general framework which can be used for Bayesian inference on SDE models rather than using MCMC based methods. As the research progresses, the methodological advances can be incorporated into the GaMBA framework to make feasible even faster and even more accurate Bayesian inference for SDE models. Further, since GaMBA is computationally cheaper than the MCMC based methods, its use has the potential to make possible the inference

on several highly complex processes using SDE models.

- This research has attempted to model the dynamic force exerted by the vehicles on the road surface using SDE models. As far as the author and the collaborators are aware, this has not been done so far. An SDE model based on one of the existing differential equation models was used to fit a simulated force data using GaMBA. This was considered as a *'proof of concept'* work to investigate if the SDE modeling of this problem is feasible.

1.3 Overview of Chapters

The rest of this thesis is organised as follows.

Chapter 2 This chapter provides the basic overview of the preliminaries required to proceed to the topic of Bayesian inference on SDE models – the topic of this thesis. This includes material on SDEs, diffusion processes, numerical methods for SDEs as well as material on Monte Carlo methods for Bayesian Inference.

Chapter 3 provides a brief overview of the various statistical inference methods that have been used for SDE models. This includes both the *classical* likelihood based methods as well as methods for Bayesian inference. Special emphasis has been given on the methods specially relevant to this thesis, i.e. Simulated likelihood based methods using Importance sampling and Bayesian methods.

Chapter 4 is the main chapter of this thesis. It first describes the basic concept behind GaMBA and then discusses in detail the various issues encountered while implementing this concept in practice also providing an algorithm which can be used to

implement GaMBA on any one dimensional SDE model. It then introduces GaMBA-I and discusses its consistency properties. This chapter describes examples where these methods provide efficient inference and also illustrates situations where these methods wont be too efficient. Finally, this chapter discusses various practical aspects regarding the implementation of GaMBA and GaMBA-I including its limitations.

Chapter 5 is regarding the engineering application and illustrates how GaMBA can be effectively used in a real life modeling problem. It gives an overview of the existing differential equation models and then proceeds to describe how an SDE model can instead be built. This SDE model is then fitted using both GaMBA and the basic MCMC method on a simulated data for dynamic force. The performance of GaMBA is compared with the MCMC method, both in terms of accuracy and computational efficiency.

Chapter 6 concludes the thesis and provides a discussion on the possible areas of further research.

Chapter 2

Stochastic Differential Equations & Monte Carlo Methods for Bayesian Inference

This chapter provides the basic overview of the preliminaries required to proceed to the topic of Bayesian inference on stochastic differential equation (SDE) models – the topic of this thesis.

For probability theory, random processes, and stochastic calculus excellent texts in increasing order of mathematical rigour are: Grimmett and Stirzaker (2001), Koralov and Sinai (2007), and Dudley (2003). For a complete course on SDE's and stochastic calculus, refer to Oksendal (2007). Lacus (2008) provides very accessible introduction to all aspects of SDE modeling for practitioners; while Kloeden and Platen (1992) provide an exhaustive account of the numerical methods used for SDE's, and also provides a very accessible introduction to probability, random processes, and stochastic calculus.

For an excellent account of statistical inference and discussion on the strengths and the possible drawbacks of various approaches to statistical inference, refer to Cox (2006). Comprehensive texts on Bayesian inference include, among others, Gelman et al. (2003), Gilks et al. (1996) and Bernardo and Smith (2000). Robert and Casella (2004) provide a comprehensive account on the Monte Carlo methods in Statistics.

2.1 Stochastic Differential Equations (SDEs)

2.1.1 Why SDE models?

Differential equation models are used to model dynamical systems across a wide range of areas such as biology, ecology, engineering and economics. Often however, such differential equation models fail to completely capture the inherent uncertainties associated with the system to be modeled. Allowing randomness in some of the coefficients of a differential equation might result in a more realistic mathematical model. Consider the following example about a population growth model.

Let $N(t)$ be the size of the population at time t, and $a(t)$ be the relative rate of growth at time t, then a straightforward differential equation model might be as follows:

$$\frac{dN}{dt} = a(t) \cdot N(t), \quad N(0) = n_0 \quad \text{(constant)}.$$

It might happen that $a(t)$ is not completely known, but subject to some random environmental effects, so that

$$a(t) = r(t) + \text{"}noise\text{"},$$

where the exact behaviour of the *"noise"* term is not known, but its probability distribution is known. The function $r(t)$ is assumed to be non-random. The model now becomes

$$\frac{dN}{dt} = (r(t) + "noise") \cdot N(t).$$

More generally, one might be interested in solving equations of the form

$$\frac{dX_t}{dt} = f(X_t, t) + g(X_t, t) \cdot "noise" \tag{2.1}$$

where f and g are some given functions. The question is, whether it is mathematically possible to deal with such equations? If yes, then how?

The answer is of course in the affirmative and the resulting equation is called the stochastic differential equation (SDE). In fact, there are different ways to solve (if the solution exists) an SDE. The most commonly used approach in mathematical modeling and statistics is using *Itô's calculus*. The solution to a SDE using Itô's calculus is a random process; more precisely it is a diffusion process.

The general form of an SDE is given by

$$dX_t = f(X_t, t)dt + g(X_t, t) \cdot dW_t, \quad X_0 = x_0 \tag{2.2}$$

where W_t is the Weiner process defined in Section 2.2.1.

The question is how to interpret this equation? That is to say, how to solve this equation? This question is non-trivial because note that W_t is no-where differentiable. This is precisely the question that stochastic calculus aims to answer. The different approaches that can be undertaken for doing this are the *Itô's calculus*, its variational

relative the *Malliavin calculus* and the *Stratonovich integral*. The reader is referred to Oksendal (2007) for a comprehensive introduction to stochastic calculus.

For the purposes of this thesis, the stochastic differential equation (2.2) with some initial condition x_0 is to be conveniently interpreted to mean the integral equation

$$X_t = x_0 + \int_0^t f(X_s, s)\, ds + \int_0^t g(X_s, s)\, dW_s \tag{2.3}$$

where the final integral is defined as an Itô integral.

2.1.2 Existence of the Solution

This subsection states some basic results on the existence of a unique solution to an SDE. Unless specified otherwise, the following material is based on Lacus (2008), p. $33 - 35$ and Oksendal (2007), p. $68 - 72$.

Consider the stochastic differential Equation (2.2). The initial condition can be random or not. If random, say $X_0 = Z$, it should be independent of the σ-algebra $\mathcal{F}_\infty^m$ generated by W_t and satisfy the condition $E|Z|^2 < \infty$. The two deterministic functions $f(.,.)$ and $g(.,.)$ are called respectively the *drift* and *diffusion* coeffecients of the SDE, and it is henceforth assumed that, they are measurable.

As in the ODE case, an SDE may have no solution, or it may have one or more. In fact, fairly mild conditions on f and g are sufficient to ensure that (2.2) has a unique solution.

Theorem 1 **: Existence and Uniqueness Theorem :-** Consider the stochastic differential Equation (2.2). If the functions f and g satisfy the following conditions:

Global Lipschitz : For all $x, y \in \Re$ and $t \in [0, T]$, there exists a constant $K < \infty$ such that

$$|f(x,t) - f(y,t)| + |g(x,t) - g(y,t)| < K|x - y|,$$

Linear Growth : For all $x, y \in \Re$ and $t \in [0, T]$, there exists a constant $C < \infty$ such that

$$|f(x,t)| + |g(x,t)| < C(1 + |x|).$$

then (2.2) has a unique and continuous *strong solution* X_t adapted to the σ-algebra $\mathcal{F}_t^z$ generated by Z and W_t and such that

$$E\left[\int_0^T |X_t|^2 dt\right] < \infty.$$

The Lipschitz condition ensures that the solution has continuous paths, and the linear growth condition controls the behaviour of the solution so that X_t does not explode to infinity in a finite time.

The result above states that the solution X_t is of *strong* type. This implies that the solution is *pathwise* unique. X_t is *strong* because the version W_t of the Wiener process is given in advance, and the solution X_t constructed from it is $\mathcal{F}_t^z$ adapted.

Instead, if the Wiener process version is not assumed to be known then the solution X_t is called a *weak* solution. If there are two weak solutions X^1 and X^2, then they may not necessarily be pathwise identical, however their distributions would be. Thus, weak solutions are often enough from a statistical inference point of view. Of course strong solutions are also weak solutions, but the contrary is not necessarily true.

Sometimes, the global Lipschitz condition of Theorem 6 can be too restrictive (Kutoyants (2004), p.25) and can be weakened using a *Local* Lipschitz condition

Local Lipschitz : For any $M > 0$ and all $x, y \in \Re$ such that $|x| < M$, $|y| < M$, and $t \in [0, T]$, there exists a constant $K_M < \infty$ such that

$$|f(x,t) - f(y,t)| + |g(x,t) - g(y,t)| < K_M|x - y|.$$

Then there exists a unique solution $\{X_t\}$ of Equation (2.2) under the local Lipschitz condition (Friedman (1975), p.104).

If the SDEs of interest are (time) homogeneous SDE's of the form

$$dX_t = f(X_t)dt + g(X_t)dW_t. \tag{2.4}$$

then, the *weak* solutions exist under fairly mild conditions as stated below.

Theorem 2 **: Existence of weak solution :-** (see for e.g. Durett (1996), p. 210 or Kutoyants (2004), p. 25) For time homogeneous SDEs such as Equation (2.4). Let f be locally bounded, and g be continuous and positive. For some A, if the functions f and g satisfy the following condition:

$$xf(x) + g^2(x) \leq A(1 + x^2)$$

for any $x \in \Re$, then (2.4) has a unique *weak solution*.

Finally, it can be shown that the solution of an SDE (if it exists) is a continuous Markov process (Friedman (1975), p.109).

2.1.3 Itô's formula & Lamperti Transform

Unlike ordinary calculus, it is not possible in stochastic calculus to switch at will between the two approaches of differential equations and integration. In stochastic

calculus, the useful range of techniques is practically restricted to those that deal with integral equations. Of these, an important technique is what is known as Itô's formula, which can be seen as a stochastic chain rule and is given by the following theorem (Grimmett and Stirzaker (2001), pg. 545).

Theorem 3 **:Itô's formula :-** If X is a diffusion process satisfying the SDE of Equation (2.7) and $Y_t = h(X_t, t)$, where h is twice continuously differentiable on $[0, \infty) \times \Re$ then Y is also a diffusion process given by

$$dY_t = [h_x(X_t, t)f(X_t, t) + h_t(X_t, t) + \frac{1}{2}h_{xx}(X_t, t)g^2(X_t, t)]dt + h_x(X_t, t)g(X_t, t)dW_t \tag{2.5}$$

where $h_x(X_t, t)$ and $h_t(X_t, t)$ denote the derivatives of h w.r.t. its first and second arguments respectively and evaluated at (X_t, t), whereas $h_{xx}(X_t, t)$ denotes the second derivative of h w.r.t. X.

There is one particular application of Itô's formula that is of interest in statistical estimation problems and is often used (see for e.g. Roberts and Stramer (2001),Ait-Sahalia (2002),Lacus (2008)). Suppose we have the stochastic differential equation

$$dX_t = f(X_t, t)\ dt + g(X_t, t)\ \ dW_t,$$

with a non-constant diffusion coefficient. Such an SDE can be transformed into one with unitary diffusion coefficient by applying the Lamperti transform,

$$Y_t = h(X_t) = \int_z^{X_t} \frac{1}{g(u, t)} du. \tag{2.6}$$

Here z is any arbitrary value in the state space of X. Indeed the process solves the SDE

$$dY_t = b(Y_t, t)dt + dW_t,$$

where

$$b(y,t) = \frac{f(h^{-1}(y),t)}{g(h^{-1}(y),t)} - \frac{1}{2}g_x(h^{-1}(y),t),$$

where $g_x = dg(.)/dx$.

To obtain this result, one should use the Itô's formula with

$$h(x,t) = \int_z^{X_t} \frac{1}{g(u,t)}du, \;\; h_t(x,t) = 0, \;\; h_x(x,t) = \frac{1}{g(x,t)}, \;\; h_{xx}(x,t) = -\frac{g_x(x,t)}{g^2(x,t)}.$$

2.2 Diffusion Processes

A particle is said to be *diffusing* about a space $\Re^n$ whenever it experiences erratic and disordered motion through the space; for example, one may speak of radioactive particles diffusing through the atmosphere, or even a rumour diffusing through a population. Random processes which try to model such phenomena are called ***diffusion processes***. They are continuous both in state space ($\Omega = \Re^n$) as well as the index set ($T = \Re^d$). A diffusion process can be defined as follows (Stirzaker (2005), pg. 224):

Definition : A random process $X = \{X_t : t \geq 0\}$ is called a diffusion process, if it is a continuous (a.s) Markov process satisfying

$$P(|X_{t+h} - X_t| > \epsilon | X_t = x) = o(h) \quad \text{for all } \epsilon > 0, \tag{2.7}$$

$$E(X_{t+h} - X_t | X_t = x) = a(X_t,t)h + o(h), \tag{2.8}$$

$$E([X_{t+h} - X_t]^2 | X_t = x) = b^2(X_t,t)h + o(h). \tag{2.9}$$

The functions a and b are called the 'instantaneous mean' (or 'drift') and 'instantaneous variance' (or 'diffusion') of X respectively.

If dX_t is used as a convenient shorthand to denote the small increment in X_t over a small interval dt, then using the properties (2.7) to (2.9) above dX_t can be expressed as

$$dX_t = a(X_t, t)\ dt + b(X_t, t)\ \ dW_t \tag{2.10}$$

where W_t is the Weiner process defined in the next subsection. Equation (2.10) is in fact the general form of an SDE. Thus, another way to look at diffusion processes, and the one that will be adhered to in this thesis is as a solution to a stochastic differential equation using Itô's calculus. Ornstein-Uhlenbeck process, geometric Brownian motion process, etc are some such examples of diffusion processes. Such processes are used to model phenomena in a wide range of areas such as economics, biology and engineering.

2.2.1 Weiner Process

The Wiener process is the archetypal diffusion process and is in fact the process incorporated in all SDE's. Its development was motivated by the need to model the erratic random motion of tiny particles observed by the Scottish botonish Robert Brown in 1827; therefore it is also commonly referred to as the ***Brownian Motion***. It is defined as follows (Grimmett and Stirzaker (2001), pg. 516):

Definition : A Wiener process $W = \{W_t : t \geq 0\}$, starting from $W_0 = w$, say, is a real valued Gaussian process such that:

(a) W has independent increments;

(b) $W_{s+t} - W_s$ is distributed as $N(0, \sigma^2 t)$ for all $s, t \geq 0$ where σ^2 is a positive constant;

(c) the sample paths of W are (Hölder) continuous.

The process W is called a *standard Wiener process* if $\sigma^2 = 1$ and $W_0 = 0$. If W is non-standard then $\hat{W}_t = (W_t - W_0)/\sigma$ is standard.

Clearly (a) and (b) specify the *fdds* of the Wiener process which is Gaussian. An immediate implication of (b) is that the Wiener process has stationary increments, since the distribution of $W_{s+t} - W_s$ depends on t alone.

The autocovariance function of W is given by :

$$\begin{aligned} c(s,t) &= E([W_s - W_0][W_t - W_0]) \\ &= E([W_s - W_0]^2 + [W_s - W_0][W_t - W_s]) \\ &= \sigma^2 s + 0 \quad \text{if } 0 \leq s \leq t, \end{aligned}$$

which is to say that,

$$c(s,t) = \sigma^2 min(s,t) \quad \text{for all } s,t \geq 0. \tag{2.11}$$

There are two types of statements to be made about random processes. The first deals with sample path properties, and the second with distributional properties. While (a) and (b) in the definition specify the *fdds* of the Wiener process; some of the path properties are immediately clear as well. For example, Equation (2.11) implies that

$$E([W_{s+t} - W_s]^2) \to 0 \quad \text{as } t \to 0;$$

i.e. W is *continuous in mean squared*.

Following is the list of important properties of the Wiener process:

- it is a Markov process;
- it is a Gaussian process;

- it has stationary and independent increments;
- its sample paths are (Hölder) continuous almost everywhere;
- but they are differentiable (a.s) nowhere.

2.2.2 Ornstein-Uhlenbeck (OU) Process

This process is popular generalisation of the Wiener process model and is given by

$$dX_t = (\theta_1 - \theta_2 X_t)\, dt + \theta_3\, dW_t, \quad X_0 = x_0, \tag{2.12}$$

with $\theta_3 \in \Re_+$, and $\theta_1, \theta_2 \in \Re$. With $\theta_1 = 0$, the OU process first originated in Physics (Uhlenbeck and Ornstein (1930)), where it was founded on the assumption that the velocity of the particle (rather than its position) undergoes a random walk and that the motion of the particle is damped by the frictional resistance of the fluid. Vasicek (1977) later used the OU process to model evolution of interest rates.

The transition density of the OU process $p_t(X_t|X_0 = x_0)$; i.e. the density of the distribution of X_t given $X_0 = x_0$, is Gaussian with mean and variance respectively (Lacus (2008), pg. 45)

$$\mu_t(x) = \frac{\theta_1}{\theta_2} + \left(x_0 - \frac{\theta_1}{\theta_2}\right) e^{-\theta_2 t}$$

and,

$$\sigma_t^2(x) = \frac{\theta_3^2(1 - e^{-2\theta_2 t})}{2\,\theta_2}.$$

The OU process has several interesting properties. Contrary to the Wiener process, it is a process with finite variance for all $t \geq 0$. Also, OU process is ergodic and its invariant law is Gaussian with mean θ_1/θ_2 and variance $\theta_3^2/2\theta_2$. For $\theta_2 > 0$, the process is *mean reverting*, meaning that the process tends to oscillate around some equilibrium

state. In fact θ_2 governs the rate at which the process reverts back to it's mean.

2.2.3 Diffusion Bridge

Although diffusion processes are continuous time processes, while modeling a real life phenomemenon, they might often be observed only at discrete time points. Thus, it is often of considerable interest to determine the distribution of the path taken by a diffusion process between the two observed time points. Such a diffusion process conditioned on the values taken by the process at two distinct time points is called a 'tied down diffusion process' or a 'diffusion bridge'.

Definition : Let $X = \{X_t\}$ be a diffusion process such that $X_{t_1} = a$ and $X_{t_2} = b$ for $t_1 < t_2$. Then the ***diffusion bridge*** $Y = \{Y_t\}$ is a random process which has the same distribution as $\{X_t | X_{t_1} = a, X_{t_2} = b; t_1 \leq t \leq t_2\}$. In a short-hand notation, such a bridge is often referred to as a (t_1, a, t_2, b) bridge (Bladt and Sorensen (2010)).

The diffusion bridge corresponding to the Wiener process is called a ***Brownian bridge*** and its *fdds* are Gaussian as given by the following theorem (Oksendal (2007)).

Theorem 4 **:** Let $B = \{B_t : t_1 \leq t \leq t_2\}$ be a process with continuous sample paths and the same *fdds* as the Wiener process $W = \{W_t : t_1 \leq t \leq t_2\}$ conditioned on $W_{t_1} = a$ and $W_{t_2} = b$. The process B is a diffusion process which solves the SDE:

$$dB_t = \frac{(b - B_t)}{(t_2 - t)}\, dt + \, dW_t, \;\; B_0 = a.$$

It turns out that B_t has a Gaussian transition density and with instantaneous mean

μ_t and variance σ_t^2 given by:

$$\mu_t = a + \frac{(t - t_1)}{(t_2 - t_1)}(b - a) \quad \text{for } t_1 \leq t \leq t_2,$$

$$\sigma_t^2 = \frac{(t - t_1)(t_2 - t)}{(t_2 - t_1)} \quad \text{for } t_1 \leq t \leq t_2.$$

Note that μ_t is just the linear interpolation between a and b, and the variance $\sigma_t^2 \to 0$ as $t \downarrow t_1$ or $t \uparrow t_2$.

It is also possible to derive a closed form expression for an OU-bridge (diffusion bridge corresponding to the OU process) when $\theta_1 = 0$. It can be shown (Bladt and Sorensen (2010)) that the $(0, a, 1, b)$ OU-bridge X_t is a solution to the SDE:

$$dX_t = \frac{\theta_2 \left(X_t - 2(X_t - b\, e^{-\theta_2(1-t)})\right)}{(1 - e^{-2\theta_2(1-t)})}\, dt + \theta_3\, dW_t, \quad X_0 = a.$$

In general, there are some results available (e.g. Bladt and Sorensen (2010), Lyons and Zheng (1990)) to obtain SDEs for diffusion bridges corresponding to general diffusion processes. Consider a diffusion corresponding to a general SDE in Equation (2.2). Then as stated in Delyon and Ying (2006), the distribution of a discretised (t_1, a, t_2, b) diffusion bridge corresponding to this SDE is same as that of another diffusion Y_t satisfying

$$dY_t = \tilde{f}(Y_t, t)\, dt + g(Y_t, t)\, dW_t, \quad Y_0 = a, \quad t_1 \leq\, t \leq\, t_2,$$

where

$$\tilde{f}(X_t, t) = f(X_t, t) + [gg'](X_t, t)\Delta_x(\log\, p(t, X_t, t_2, b)),$$

and $p(t, X_t, t_2, b)$ is the transition density of X_t. In practice, however, theses results are often of limited use since they require the transition density of the original diffusion to be known in closed form. Therefore apart from a few exceptions such as the

Wiener process, or the OU process, it is not possible in general to derive closed form expressions for diffusion bridges.

For this reason, simulating diffusion bridges (exact or approximate) is an important problem of the statistical inference on SDE models.

2.3 Numerical Methods for SDEs

An exact closed form solution is often not known for many SDEs used in practice. For such SDEs, approximate solutions can be simulated using numerical methods which are usually based on the time discrete approximations of the continuous solution. Thus, to simulate a solution over the time interval $[T_0, T]$, the time interval is divided into N parts such that $T_0 = t_0 < t_1 < \cdots < t_{N-1} < t_N = T$. For the sake of simplicity, it is assumed here that the time points $t_0, t_1, \cdots, t_N$ are equally spaced with an interval $\delta = t_i - t_{i-1}$, for $i = 1, \cdots, N$ between any two consecutive time points. Figure 2.1 illustrates this time discretisation. However,it is important to note that this condition is not necessary, and these methods could be defined for a more general time discretisation (see Kloeden and Platen (1992)).

Approximation provided by a method could be assessed at two levels: *strong convergence* and *weak convergence*. The *strong convergence* criterion is useful when the purpose of the simulation is to approximate the true path as closely as possible. On the other hand, when the objective is to approximate the distributional properties of the true diffusion, the *weak convergence* criterion is more appropriate. Thus, for the purposes of this thesis, the *weak convergence* criterion is of primary interest. The rate

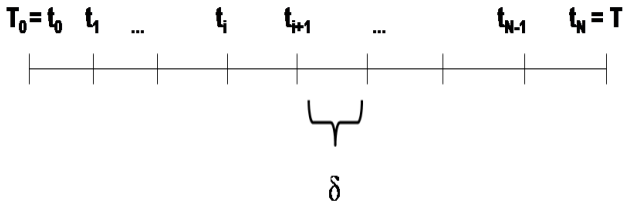


Figure 2.1: Time discritisation for numerical methods.

at which a method converges to the true process is determined by its *order of convergence*. Methods of approximation of some order that strongly converge, usually have a higher order of weak convergence. The reader is referred to Kloeden and Platen (1992) for a detailed exposition on this topic.

Definition : A time discrete approximation X_δ of a continuous process X, with the time increment δ of the discretisation, is said to be of general *strong order of convergence* γ to X, if for any fixed time horizon T it holds true that

$$E|X_\delta(T) - X(T)| \leq C\delta^\gamma, \;\; \forall \delta < \delta_0, \tag{2.13}$$

with $\delta_0 > 0$ and C a constant not depending on δ.

Definition : A time discrete approximation X_δ is said to converge *weakly of order* β to X, if for any fixed time horizon T and any $2(\beta + 1)$ continuous differentiable function g of polynomial growth, it holds true that

$$|E(g(X(T)) - E(g(X_\delta(T)))| \leq C\delta^\beta, \;\; \forall \delta < \delta_0, \tag{2.14}$$

with $\delta_0 > 0$ and C a constant not depending on the time increment of the discretisation δ.

Two widely used numerical methods will be defined in this section. Consider the SDE (2.2) with initial condition $X_{T_0} = X_0$, and a discretisation of the time interval $[T_0, T]$ as shown in Figure 2.1.

2.3.1 The Euler approximation

One of the simples time discrete approximations of a diffusion process is the *Euler approximation* or the *Euler-Maruyama approximation* as it is sometimes called. It can also be interpreted as a strong Taylor approximation of order 0.5 (Kloeden and Platen (1992)).

The Euler approximation of Equation (2.2) is a continuous stochastic process $Y = \{Y_t, T_0 \leq t < T\}$ with $Y_0 = X_0$, and satisfying the following iterative scheme

$$Y_{i+1} = Y_i + f(Y_i, t_i)\delta + g(Y_i, t_i)(W_{i+1} - W_i), \tag{2.15}$$

for $i = 1, \cdots, N-1$, where W_t denotes the Weiner process and where the notation $Y_i = Y_{t_i}$, $W_i = W_{t_i}$ has been used.

Under the assumption that the criteria for existence of a solution as described in Theorem 1, Section 2.1.2 are satisfied, Euler approximation is strongly convergent with order $\gamma = 1/2$ (see Theorem 10.2.2 in Kloeden and Platen (1992)). Under an additional assumption that the coefficients f and g are four times continuously differentiable, Euler approximation is weakly convergent with order $\beta = 1$. Further, with milder assumptions on f and g a weak convergence of order $\beta < 1$ can be achieved (see Chapter 14 of Kloeden and Platen (1992)).

Euler method is preferred by many statistical inference methods because its transition density is Gaussian which is easier to analyse and evaluate. This transition density is given by

$$P(Y_{i+1}|Y_i, \Theta) = N(\mu_{eu}, \sigma^2_{eu}), \quad j = 0, 1, ..., (m-2) \tag{2.16}$$

where

$$\mu_{eu} = Y_i + f(Y_i, t_i) \cdot \delta$$

$$\sigma_{eu} = g(Y_i, t_i) \cdot \delta^{1/2}.$$

2.3.2 The Milstein approximation

Usually the Euler scheme gives good numerical results when the drift and diffusion coefficients are nearly constant (over the concerned time interval). In general, however, it may not be particularly satisfactory, and thus the use of higher order approximations is recommended (see Kloeden and Platen (1992), pg. 342). The Milstein approximation is one such method. It can also be interpreted as a strong Taylor approximation of order 1.

This method uses Itô's lemma to increase the accuracy of the approximation by adding the second-order term. Let g_x denote the partial derivative of $g(X_t, t)$ with respect to x. Then the Milstein approximation of Equation (2.2) is a continuous stochastic process $Y = \{Y_t, T_0 \leq t < T\}$ with $Y_0 = X_0$, and satisfying the following iterative scheme

$$Y_{i+1} = Y_i + f(Y_i, t_i)\delta + g(Y_i, t_i)(W_{i+1} - W_i) + \frac{1}{2}g(Y_i, t_i)g_x(Y_i, t_i)\{(W_{i+1} - W_i)^2 - \delta\}, \tag{2.17}$$

for $i = 1, \cdots, N-1$, where W_t denotes the Weiner process. Simplified notation of $Y_{t_i} = Y_i$ and $W_{t_i} = W_i$ has been used.

Under the assumption that the criteria for existence of a solution as described in Theorem 1, Section 2.1.2 are satisfied, and additional continuity assumptions on coefficients f and g, Milstein scheme is strongly convergent with order $\gamma = 1$ (see Theorem 10.6.3 in Kloeden and Platen (1992)). Further, under the added continuity conditions on f and g, it is also weakly convergent with order $\beta = 1$ (see Chapter 14 of Kloeden and Platen (1992)).

2.3.3 Effect of step size δ

By definition, the accuracy of both the pathwise approximation as well as the *weak* approximation depends firstly on the rate parameter and secondly on the step size δ. For a given method, the rate parameter is usually known and constant (for e.g. Euler's method has $\gamma = 0.5$ and $\beta = 1$). However the practitioner can control the accuracy of the approximation by changing the step size δ.

For Euler's method, Equations (2.13) and (2.14) imply that while the pathwise approximation $E|X_\delta(T) - X(T)|$ has an upper bound of $\mathbf{o}(\delta^{0.5})$, the *weak* approximation $|E(g(X(T)) - E(g(X_\delta(T)))|$ has an upper bound of $\mathbf{o}(\delta)$. Thus, both these approximations get better as δ becomes smaller. Refer to Kloeden and Platen (1992) for detailed illustrations on the effect of δ over the quality of the approximations.

This property has an important consequence for statistical inference. As will be discussed in Chapters 3 and 4, an approach in statistical inference involves imputing

latent variables simulated using Euler's approximation. For this inference to be accurate it is imperative that the step size is small enough. Though this increases the computational demand of the inference method, it also imparts the desirable asymptotic consistency properties to the method.

2.4 Monte Carlo Methods for Approximating Posterior Distributions

Bayesian inference involves computing probabilities and expectations. For continuous probability distributions, this implies evaluating integrals. Although, exact analytical evaluation of such integrals would be preferred, often they are non-standard or analytically intractable. When analytical evaluation is impossible, numerical integration is an option. However, Bayesian model specifications can often produce high-dimensional integrals for which numerical methods become computationally involved. This is because the number of function evaluations required to achieve a certain degree of approximation increases exponentially in the dimension of the problem. Therefore, it is important to consider other methods for evaluating integrals which do not suffer so directly from the increase in the dimensionality of the problem. *Monte Carlo methods* are a class of such methods.

The term *Monte Carlo methods* refers to a general class of computational methods that rely on repeated random sampling to compute their results. In statistics, these methods are used to approximate analytically intractable, high dimensional integrals which need to be evaluated to obtain probabilities, moments, etc. Their implementation typically involves two steps. First step involves *sampling* a large number of draws

from the desired probability distributions. The second step involves *approximating the integral* required to obtain the necessary inference.

This section provides a brief overview of the Monte Carlo methods relevant to this thesis. They are - Monte Carlo integration, Importance sampling and *Markov Chain Monte Carlo* (MCMC) methods.

2.4.1 Monte Carlo Integration

Monte Carlo integration is a Monte Carlo method for obtaining integration based summaries. Suppose X is a random variable with probability density $p(X)$. Then Monte Carlo integration can be used to estimate the the following integral

$$J(X) = \int h(X)p(X)\,dX = \mathbf{E}_p[h(X)] \tag{2.18}$$

as

$$\hat{J}(X) = \frac{1}{N}\sum_{i=1}^{N} h(X_i) \tag{2.19}$$

where $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} p(X)$. Then by the Strong Law of Large Numbers

$$\hat{J}(X) \overset{a.s}{\to} J(X) \quad \text{as } N \to \infty$$

Further, when h^2 has a finite expectation under p, the speed of convergence of $\hat{J}$ can be assessed (see Robert and Casella (2004), Section 3.2) since the variance

$$\mathbf{V}[\hat{J}(X)] = \frac{1}{N}\int (h(X) - \mathbf{E}_p[h(X)])^2 p(X)\,dX$$

can also be estimated using $X_1, X_2, \cdots, X_n$ through

$$\hat{\mathbf{V}}[\hat{J}(x)] = \frac{1}{N^2}\sum_{i=1}^{N}[h(x_i) - \hat{J}(x)]^2. \tag{2.20}$$

This is a very useful property in practice, since it provides the rate at which the variance of the Monte Carlo estimate decreases as the sample size N increases. Further, unlike numerical approximation methods this rate does not depend on the dimensionality of X.

Because the Monte Carlo integration is a direct consequence of the Strong Law of Large Numbers, it requires the samples to be independent. However, in practice it might still provide a good approximation when the samples are not strictly independent. Further, the *strong law of large numbers* is available for stationary Markov chains (see Roberts and Rosenthal (2004)). Thus, the Monte Carlo integration method works for both the direct simulation methods (which generate independent samples) as well as the Markov chain Monte Carlo methods (which generate correlated samples).

Monte Carlo integration has many uses in Bayesian inference since many integrals of interest can be approximated by expectations e.g. it appears when evaluating predictive density for a new observation, marginalizing out parameters, obtaining posterior moments, etc.

2.4.2 Importance Sampling

When implementing Monte Carlo integration, samples are drawn directly from the distribution of interest $p(\cdot)$. However, in certain situations (see Robert and Casella (2004), Section 3.3), it may be more appropriate to sample from a different distribution – often called the *importance sampling density* (see Geweke (1989)) – instead and then to modify the representation of the integral as an expectation with respect to this density.

The method of *importance sampling* involves representing Equation 2.18 as

$$J(X) = \int h(X)\frac{p(X)}{q(X)}\, q(X)\, dX = \mathbf{E}_p[h(X)], \tag{2.21}$$

where $q(X)$ is the importance sampling density and then estimating $J(X)$ as

$$\hat{J}(X) = \frac{1}{N}\sum_{i=1}^{N} h(X_i)\frac{p(X_i)}{q(X_i)} \tag{2.22}$$

where $X_1, X_2, \cdots, X_n \overset{i.i.d}{\sim} q(X)$. Geweke (1989) has shown that as long as the support of $q(\cdot)$ contains the support of $p(\cdot)$, then $\hat{J}(X) \overset{a.s}{\to} J(X)$ as $N \to \infty$ irrespective of the choice of $q(\cdot)$. Importance sampling is therefore a very appealing method (Robert and Casella (2004), Section 3.3) as it puts very little restriction on the importance sampling density and it can be chosen as the density which is easier to sample. In practice, however, a well chosen importance density can make the estimation more efficient and Geweke (1989) provides come guidelines in this regard.

Importance sampling gives more weights to regions where $p(X) > q(X)$ and down-weighs regions where $p(X) < q(X)$. Thus, if $q(\cdot)$ is chosen so as to have the same mode as $p(\cdot)$ but with a larger variance, then more samples will be drawn from the *high density* region of $p(\cdot)$. This can be a desirable property.

If the probability measure corresponding to the importance density is absolutely continuous with respect to the probability measure of interest, then importance sampling can be implemented using the change of measure formula (*Radon Nikodym derivative*).

Let $X = X(\omega)$ be *measurable* on probability space $(\Omega, \mathcal{F}, P)$. Let $Q(\omega)$ be a probability measure absolutely continuous with respect to $P(\omega)$ and let $\tilde{P} = dP/dQ$

be the Radon-Nikodym derivative of $P(\cdot)$ with respect to $Q(\cdot)$. Then, Equation (2.21) can be expressed as a Lebesgue integral

$$\begin{aligned} J(X) = \int h(X(\omega))\, dP(\omega) &= \int h(X(\omega)) \frac{dP}{dQ}\, dQ(\omega) \\ &= \mathbf{E_Q}\left[h(X(\omega))\tilde{P}(\omega)\right], \end{aligned} \tag{2.23}$$

which can be approximated using importance sampling as

$$\hat{J}(X) = \frac{1}{N}\sum_{i=1}^{N} h(X(\omega_i))\tilde{P}(\omega_i) \tag{2.24}$$

where, $\omega_1, \cdots, \omega_N \sim Q(\cdot)$.

2.4.2.1 Girsanov's Theorem

Girsanov's theorem is a change of measure theorem for stochastic processes (Lacus (2008)). Because it provides an analytical expression for the change of measure, it is useful in statistical inference on stochastic processes and is often used in inference on SDE models (see Prakasa Rao (1999), Kutoyants (2004)). The theorem stated below is from Lacus (2008), pg. 41.

Consider three SDEs:

$$\begin{aligned} dX_t &= b_1(X_t)\, dt + \sigma(X_t)\, dW_t, \quad X_0^1, \quad 0 \le t \le T, \\ dX_t &= b_2(X_t)\, dt + \sigma(X_t)\, dW_t, \quad X_0^2, \quad 0 \le t \le T, \\ dX_t &= \sigma(X_t)\, dW_t, \quad X_0, \quad 0 \le t \le T, \end{aligned}$$

and let P_1, P_2 and P denote the three probability measures induced by the solutions of these three SDEs respectively.

Girsanov's theorem: Assume that the coefficients of each of the above SDEs satisfy conditions for existence of a weak solution mentioned in Section 2.1.2. Assume further that the initial values are either random variables with densities $f_1(\cdot)$, $f_2(\cdot)$ and $f(\cdot)$ with the same common support or non-random and equal to the same constant. Then the three measures P_1, P_2 and P are all equivalent and the corresponding Radon-Nikodym derivatives are

$$\frac{dP_1}{dP}(X) = \frac{f_1(X_0)}{f(X_0)} \exp\left\{ \int_0^T \frac{b_1(X_s)}{\sigma^2(X_s)}\, dX_s - \frac{1}{2}\int_0^T \frac{b_1^2(X_s)}{\sigma^2(X_s)}\, ds \right\}, \tag{2.25}$$

and

$$\frac{dP_2}{dP_1}(X) = \frac{f_2(X_0)}{f_1(X_0)} \exp\left\{ \int_0^T \frac{b_2(X_s) - b_1(X_s)}{\sigma^2(X_s)}\, dX_s - \frac{1}{2}\int_0^T \frac{b_2^2(X_s) - b_1^2(X_s)}{\sigma^2(X_s)}\, ds \right\}. \tag{2.26}$$

When importance sampling is to be used for inference on SDE models, Girsanov's theorem can be used to evaluate the Radon-Nikodym derivative if the probability measure corresponding to the importance density is absolutely continuous with respect to the probability measure of the diffusion process.

2.4.3 Markov Chain Monte Carlo (MCMC) Methods

The Monte Carlo integration and *importance sampling* methods discussed above illustrate how Monte Carlo methods could be used to approximate the integrals, but these methods rely on the samples having been already drawn from the desired distribution. Even though direct sampling methods such as rejection sampling and the inverse transform method etc. draw samples exactly from the desired distribution, these methods break down as the dimensionality of the distribution increases. In such cases approximate sampling methods such as MCMC can be used.

MCMC methods are an elegant way of sampling from high-dimensional distributions and thus approximating high-dimensional integrals. They are based on the premise (see Cappé et al. (2005)) that simulating an i.i.d sequence $X_1, X_2, \cdots, X_n$ with common probability distribution π is not the only way of approximating π in the sense of being able to approximate the expectation of a π integrable function. In particular, one may consider an ergodic Markovian sequence $\{X_i\}$ with π as its stationary distribution instead. An MCMC method for the simulation of a distribution π can be defined as (see Robert and Casella (2004)) any method producing an ergodic Markov chain $\{X_i\}$ whose stationary distribution is π.

Implementing these methods typically involves running several Markov chains for a large number of transitions until each of them can be considered to have reached stationarity. It is important to note that the samples drawn using these methods typically represent realisations of a set of identically distributed, correlated random variables.

Based on the mechanism used to generate the Markov chains, these methods can be classified into two main types: the *Gibbs sampling* and the *Metropolis-Hastings (MH) algorithm*.

2.4.3.1 Gibbs Sampling

It is the simplest MCMC method to implement and originated in statistical physics where it was known as the *heat bath algorithm*. It relies on the assumption that each of the full-conditional distributions $f(\theta_i|\theta_{-i})$ is a distribution from which samples can be easily drawn. Its steps can be summarised as follows:

Step I : Initialise $\theta_2^0, \cdots, \theta_n^0$.

Step II : for $r = 1$ to R

- sample θ_1^r from $f(\theta_1|\theta_2^{r-1}, \cdots, \theta_n^{r-1})$,
- sample θ_2^r from $f(\theta_2|\theta_1^r, \theta_3^{r-1}, \cdots, \theta_n^{r-1})$,

  $\vdots$
- sample θ_n^r from $f(\theta_n|\theta_1^r, \cdots, \theta_{n-1}^r)$.

Step III : Repeat Step II for a large (R) number of times, until the stationary distribution is reached for every θ_i.

Step IV : Continue sampling using Step II, the samples can now be considered to be draws from their respective stationary distributions, as desired.

Gibbs sampling is often the first choice when the full conditionals are available in closed form. It is not straightforward to use Gibbs sampling if this is not the case. However, in certain cases it is possible to get around this problem, for e.g. when the full conditional distributions are univariate log concave or nearly log concave, methods such as ARS and ARMS (see Gilks et al. (1994)) can be used to implement Gibbs sampling. A drawback of Gibbs sampling is that when θ_i's are highly correlated, the convergence can be excruciatingly slow. This is because high-correlation implies that new samples drawn from the full conditionals render values very close to the previous samples and thus the chains move very slowly. In such cases, the model needs to be suitably reparameterised before Gibbs sampling can be used. See Robert and Casella (2004) for some suitable reparameterisations.

2.4.3.2 Metropolis-Hastings (MH) Algorithm

MH algorithm is a generalization by Hastings (1970) to the Markov chain method first proposed by Metropolis et al. (1953). The idea is to sample from the joint distribution

$f(\theta_1, \theta_2, \cdots, \theta_n)$ instead of the full conditionals. Let $\Theta = (\theta_1, \theta_2, \cdots, \theta_n)$, and the target distribution $\pi(\Theta) = f(\theta_1, \theta_2, \cdots, \theta_n)$. This method can be implemented whenever it is possible to evaluate $\pi(\Theta^*)/\pi(\Theta)$, normalising constants need not be known. The basic idea is as follows.

Given that the Markov chain is in state Θ^r, the algorithm draws a proposal state Θ^* from a proposal distribution $q(\cdot|\Theta^r)$. The proposal state is then accepted with probability:

$$\alpha(\Theta^r, \Theta^*) = \min\left(1, \frac{\pi(\Theta^*)\, q(\Theta^r|\Theta^*)}{\pi(\Theta^r)\, q(\Theta^*|\Theta^r)}\right).$$

If the proposal state is not accepted then the chain remains in the same state. The transition probabilities of this Markov chain can be written as $P_{\Theta^r,\Theta^*} = \alpha(\Theta^r, \Theta^*) \cdot q(\Theta^*|\Theta^r)$.

The MH algorithm steps can be summarised as follows:

Step I : Initialise Θ^0.

Step II : for $r = 1$ to R

- Given the chain is in state Θ^r, generate a proposed value Θ^* by sampling from $q(\Theta^*|\Theta^r)$,
- Compute the acceptance probability $\alpha(\Theta^r, \Theta^*)$,
- Accept Θ^* with probability $\alpha(\Theta^r, \Theta^*)$, in which case $\Theta^{r+1} = \Theta^*$, else $\Theta^{r+1} = \Theta^r$.

Step III : Repeat Step II for a large (R) number of times, untill a stationary distribution is reached for every θ_i.

Step IV : Continue sampling using Step II, the samples can now be considered to be draws from their respective stationary distributions, as desired.

The originally proposed algorithm by Metropolis et al. (1953), known as the *Metropolis algorithm* is a special case where the proposal distribution is symmetric i.e. $q(\Theta^r|\Theta^*) = q(\Theta^*|\Theta^r)$, so the acceptance probability reduces to

$$\alpha(\Theta^r, \Theta^*) = \min\left(1, \frac{\pi(\Theta^*)}{\pi(\Theta^r)}\right).$$

The Gibbs sampler can also be considered as a special case of the MH algorithm where the proposal distribution is such that the proposed value is always accepted with probability 1, i.e. $\alpha(\Theta^r, \Theta^*) = 1$, always.

A widely used example of the proposal density is the *random walk* in which case $q(\Theta^r|\Theta^*) = q(|\Theta^r - \Theta^*|)$ and in this case, the MH algorithm is often referred to as the *Random walk Metropolis-Hastings (RWMH).*

The choice of the proposal generating distribution and its parameters is important because its relationship to the target density dictates the rate of convergence. Relatively small proposal moves can result in high acceptance rates (the percentage of moves accepted), long time to convergence and poor mixing i.e. the full support of the target distribution will not be properly explored and low probability areas will be under-sampled. Conversely, when the proposal moves are large in relation to the spread of the target density, the proposed states will have low acceptance rate and the chain will take long to converge. Many families of proposal distributions have been defined. More discussion of this issue can be found in Gilks et al. (1996), Gelman et al. (2003), and Chib and Greenberg (1995).

2.4.3.3 Metropolis within Gibbs Algorithm

Though MH algorithm is widely applicable and easy to understand, sampling from n dimensional distribution can often be computationally very expensive. It might be

the case though, that for some of these n parameters the full conditional distributions are available in closed form. Such parameters can be sampled using a Gibbs sampling step while for the rest of the parameters can be jointly sampled using the MH algorithm.

Even when full conditionals are not available for any of the parameters, each of the parameters can be individually sampled as in Gibbs sampling, by replacing Step II of Gibbs sampling algorithm with Step II of the MH algorithm. This modification is referred to as *Metropolis within Gibbs* algorithm.

A common situation where Metropolis within Gibbs algorithm is used is in the models where data has been augmented. *Data augmentation* involves introducing latent variables into the model and is often used to facilitate approximation of the likelihood (in situations where it is not known in closed form) but also in order to facilitate sampling and improve mixing. Suppose that the target distribution

$$P(\Theta|\mathbf{X}) \propto P(\Theta)P(\mathbf{X}|\Theta)$$

is computationally difficult, or intractable, but the data $\mathbf{X}$ can be augmented with a latent variable $\mathbf{Z}$ so that:

$$P(\Theta, \mathbf{Z}|\mathbf{X}) \propto P(\Theta)P(\mathbf{X}, \mathbf{Z}|\Theta)$$

where the new likelihood $P(\mathbf{X}, \mathbf{Z}|\Theta)$ is now tractable, as also are $P(\Theta|\mathbf{X}, \mathbf{Z})$, $P(\mathbf{Z}|\Theta, \mathbf{X})$.

The advantage of data augmentation is that even when $P(\Theta, \mathbf{Z}|\mathbf{X})$ is difficult, Gibbs sampling, or Metropolis withing Gibbs sampling can be used to alternatively sample from $P(\Theta|\mathbf{X}, \mathbf{Z})$ and $P(\mathbf{Z}|\Theta, \mathbf{X})$. It will be illustrated in Chapter 4, that data augmentation is very commonly used in the MCMC methods for Bayesian inference on the SDE models.

2.4.3.4 Theoretical insight into MCMC Convergence

The MCMC methods briefly described above are of course very attractive and have been extensively used to obtain Bayesian inference. Several questions emerge though, such as : why do MCMC methods work? Is it possible to mathematically prove that the Markov chains generated by these methods indeed converge to the intended stationary distribution? If they do, then is it possible to quantify the rate of convergence? Finding answers to such questions has been an active area of research and an excellent review of the work done can be found in Roberts and Rosenthal (2004).

Markov chains generated by the MCMC methods are discrete time, but typically with continuous state space Ξ. MCMC methods are designed so that the stationary distribution exists. However, as Roberts and Rosenthal (2004) point out, a Markov chain with a stationary distribution may not converge to it, if the chain is *reducible*. Further, a Markov chain also needs to be aperiodic and Harris recurrent for the ergodic theorem (essential for the convergence of Monte Carlo integration for Markov chains) to hold (Robert and Casella (2004)). As Tierney (1996) and Roberts and Rosenthal (2004) point out both the MH algorithm and the Gibbs sampler satisfy the conditions of *ϕ-irreducibility*, *aperiodicity*, and *Harris-recurrence* and so their convergence to a stationary distribution is thus mathematically established. In fact because of Harris-recurrence, the convergence is guaranteed from *any* starting point in the state space. Convergence of MCMC methods in general should not be taken for granted, as Roberts and Rosenthal (2006) point out the conditions under which the Metropolis within Gibbs algorithm is *not* Harris recurrent and therefore the convergence is not guaranteed when the chain is started from a *ϕ-null* set in Ξ.

Though, the conditions for mathematical convergence have been established for

many MCMC algorithms, the rate at which they do so can not yet be quantified in all the cases. Roberts and Rosenthal (2004) review some of the results that have been obtained in this regard, but also point out that more results are needed and also that the results obtained so far can not often be useful in practice.

From the point of view of implementing MCMC methods in practice some results regarding the optimal acceptance rate for MH algorithms are available. An optimal acceptance rate is the one which ensures fastest convergence to the stationary distribution by the MCMC algorithm (though, in practice, one can never be completely sure that the stationarity has been achieved). This is nonetheless of practical significance, because the acceptance rate can be easily monitored using a computer and can be controlled by the the practitioner, for example, by changing the variance of the proposal distribution. Roberts et al. (1997) and Roberts and Rosenthal (2001) proved that under certain assumptions the optimal acceptance rate for RWMH algorithm was 0.234. Further Roberts and Rosenthal (2004) argue that even when these assumptions are not strictly met, this optimal rate can still provide a good approximation to the true optimal rate. More recently, Neal and Roberts (2006) have proved that this optimal rate of 0.234 also holds for Metropolis within Gibbs algorithms.

2.4.3.5 Convergence Diagnostics for MCMC

The results mentioned in the previous section are important as mathematical proofs of the validity of the MCMC methods, however, they provide little guidance to a practitioner on *when to stop* running an MCMC algorithm and produce results.

For MCMC methods to produce accurate inference it is imperative that the algo-

rithm has explored the state space well (Robert and Casella (2004)). In practice, for a MH based method, this can be monitored by controlling the acceptance rate. An acceptance rate closer to 1 might indicate that the chain is exploring the space in very small steps which means that the chain will take much longer to explore the entire space. On the other hand an acceptance rate closer to 0 could indicate that the chain is taking very large steps and therefore getting stuck in a particular state for too long. The optimal rate mentioned above, can act as a guidance to achieving the acceptance rate at which the chain will efficiently explore the state space. However, achieving a near optimal acceptance rate does not mean that the chain has converged, and various diagnostic tools have been proposed to access the convergence.

The trace plots of the Markov chains are often examined, they are a visual tool for evaluating how well the chain has mixed. Further, as noted in the previous section, the starting point of the chain plays an important role in determinining its convergence, specially when the chain is not Harris recurrent. For such chains, the convergence can be tricky to determine as the chain that appears to have converged might be stuck in a region of the parameter space determined by the starting point of the chain. This can also happen when models are over-parameterized or have multi-modal posterior densities. It is therefore advisable to run the chain from different starting points as suggested by Gelman and Rubin (1992).

The concept of mixing is a qualitative one and it is therefore natural to assess it with visual tools, although quantitative diagnostics have been developed, for example the *Gelman-Rubin statistic* Gelman and Rubin (1992). Non-parametric tests, such as for example the *Kolomogrov-Smrionov test* can also be used to assess convergence. It is important to note, that many such tests have been developed to be used on *i.i.d*

samples, and Robert and Casella (2004) suggest ways in which these tests can be used to assess convergence of Markov chains.

Comparative reviews of MCMC diagnostics can be found in Cowles and Carlin (1996), and Robert and Casella (2004) for example, among numerous available texts. However, as Roberts and Rosenthal (2004) point out none of the diagnostic techniques provide any rigorous guarantees, and can also introduce bias in resulting estimates by, at times, prematurely claiming convergence. Therefore these techniques should not be over-relied upon.

2.4.3.6 Parallel Processing of MCMC algorithms

Due to their Markovian nature, MCMC algorithms are not straightforward to *parallelise* and therefore still remain computationally expensive to implement. Nevertheless, the easy availability of *multi-core* processors has generated much interest in developing ways to parallel-process MCMC algorithms to save computational effort. Some methods have already been proposed - for example see Brockwell (2006), and Jacob et al. (2010). At present however, applicability of such methods is still limited. The author is not aware of the use of such methods on SDE models.

Chapter 3

Statistical Inference on Stochastic Differential Equations

This chapter aims to provide a brief overview of the various statistical inference methods that have been used for stochastic differential equation (SDE) models.

Consider the SDE

$$dY_t = f(Y_t, t, \Theta)\, dt + g(Y_t, t, \Theta)\, dW_t, \;\; Y_0 = y_0, \;\; t \geq 0, \tag{3.1}$$

where W is an r-dimensional Wiener process, $\Theta \in \Xi \subset \Re^d$ is an unknown parameter, $f(\cdot, \cdot, \Theta) : \Re^p \times [0, \infty) \longmapsto \Re^p$ and $g(\cdot, \cdot, \Theta) : \Re^p \times [0, \infty) \longmapsto \mathbf{M}^{p \times r}$ ($\mathbf{M}^{p \times r}$ being the set of $p \times r$ matrices).

It is henceforth assumed that a *weak* solution to Equation (3.1) exists. In particular, it is assumed that the $f(\cdot)$ and $g(\cdot)$ satisfy the following:

A 1 : For any $R > 0$ and all $x, y \in \Re^p$ such that $|x| < R$, $|y| < R$, and $t \in [0, T]$,

there exists a constant $K_R < \infty$ such that

$$|f(x,t) - f(y,t)| + |g(x,t) - g(y,t)| < K_R|x - y|,$$

A 2 : For all $x, y \in \Re^p$ and $t \in [0,T]$, there exists a constant $C < \infty$ such that

$$|f(x,t)| + |g(x,t)| < C(1 + |x|).$$

Additional assumptions, wherever required would be specified in the following discussion.

3.1 Inference on Continuously Observed Diffusions

Diffusion processes are continuous time processes. In principle, a diffusion process could be observed as continuous paths or only at discrete time points. Statistical inference procedures based on these two types of data are quite different. The focus of this thesis is only on the discretely observed case – which is also the one most often encountered in practice. The rest of this chapter will mainly review statistical inference methods concerned with discretely observed diffusion processes. However, for the sake of completion, some basic results from the continuous case are briefly mentioned below.

First note that if Y was observed continuously from time 0 until time T, then in principle, at least, the quadratic variation of the process completely determines (rather than estimates) the diffusion coefficient, since it is well known (for e.g Sorensen (2004) ,Oksendal (2007)) that for any $t \in [0,T]$,

$$\sum_{i=1}^{n} [Y_{iT2^{-n}} - Y_{(i-1)T2^{-n}}]^2 \xrightarrow{a.s} \int_0^T g^2(Y_t, t, \Theta)\, dt \quad \text{as } n \to \infty. \tag{3.2}$$

If it is now assumed that Equation 3.1 also satisfies the following assumption:

A 3 **:** $g(\cdot)$ is independent of Θ and is a constant,

then, under the assumption **A** 3, g can be determined using the sample path as (Prakasa Rao (1999),Polson and Roberts (1994)):

$$\frac{1}{T}\sum_{i=1}^{n}[y_{iT2^{-n}} - y_{(i-1)T2^{-n}}]^2 \xrightarrow{a.s} g^2 \quad \text{as } n \to \infty.$$

Once the diffusion coefficient $g(Y_t, t)$ is completely known (or is determined), the likelihood function for Θ based on the continuous observations Y in the time interval $[0, T]$ is given by (for e.g. Sorensen (2004), Pedersen (1995b), Lacus (2008))

$$l_T = \int_0^T f(Y_s, s, \Theta)'(g(Y_s, s)g(Y_s, s)')^{-1}\, dY_s - \frac{1}{2}\int_0^T f(Y_s, s, \Theta)'(g(Y_s, s)g(Y_s, s)')^{-1} f(Y_s, s, \Theta)\, ds \tag{3.3}$$

Note that, Equation 3.3 is just the Radon-Nikodym derivative (Equation 2.25) obtained using Grisanov's theorem.

Consider the following two assumptions:

A 4 **:** SDE is time homogeneous, i.e. $f(Y_t, t, \Theta) = f(Y_t, \Theta)$ and $g(Y_t, t, \Theta) = g(Y_t, \Theta)$ for $\forall\, t \geq 0$,

A 5 **:** the functions f and g satisfy the following condition:

$$yf(y) + g^2(y) \leq A(1 + y^2)$$

for any $y \in \Re$ and some $A > 0$.

If the SDE in equation 3.1 satisfies these assumptions in addition to $A\,1$ and $A\,2$ and also the sufficient conditions for recurrence (see Kutoyants (2004), p. 40) then it has ergodic properties with the invariant density given by

$$\pi_{\Theta}(y) = \frac{1}{M(\Theta)g^2(y,\Theta)}\, exp\{2\int_{y_0}^{y} \frac{f(x,\Theta)}{g^2(x,\Theta)}\, dx,\} \tag{3.4}$$

where $M(\Theta)$ is the normalising constant.

The reader is referred to Prakasa Rao (1999) for a detailed review of the methods for statistical inference on continuously observed diffusion processes.

As mentioned earlier, the main interest of this thesis is on the methods for discretely observed diffusions. These methods are now classified and reviewed.

3.2 Likelihood Based Methods

In most applications, the data are observed only at discrete time points $0 = t_0 < t_1 < \cdots < t_n$. Let $\mathbf{Y} = \{y_0, y_1, \cdots, y_n\}$ be the corresponding observations. For the sake of notational convenience, it is also assumed that these time points are equally spaced; Δ being the time difference between each of them, although most of the methods work equally well when this is not the case.

If the transition densities $p(s, x, t, y, \Theta)$ of Y_t are known, then the log-likelihood function

$$l_n(\Theta) = \sum_{i=1}^{n} log(p(t_{i-1}, y_{i-1}, t_i, y_i, \Theta)) \tag{3.5}$$

can be used for estimation of Θ. Sufficient conditions for the maximum likelihood estimator (MLE) thus obtained to be consistent and asymptotically normal have been given (see for e.g. Prakasa Rao (1999), Kutoyants (2004)). For the few SDE models, such as for e.g. the O-U, GBM, and CIR models, for which an exact likelihood is

known, numerical methods are often used for maximising the log-likelihod (see Lacus (2008) for details).

However, for many SDEs used in practice, their transition densities are not known in the closed form and therefore Equation 3.5 can not be used. Several different approaches have been used to get around this problem. Various statistical inference methods thus developed can be classified on the basis of the approach they follow.

Approximated Likelihood Methods try to provide an approximation to the true transition density. The *Estimating Functions* approach consists of constructing a function of $\mathbf{Y}$ and Θ which imitates the score function and then to approximate the MLE using this function. One of the prominent non-likelihood based approach is the *Method of Moments* approach. Finally, a wide variety of methods have also been proposed to carry out *Bayesian Inference*.

The remainder of this chapter provides an overview of these various approaches. The method developed as part of this thesis is a method to carry out Bayesian inference, but it also has strong links to the importance sampling methods proposed to carry out the approximated likelihood methods. Therefore Bayesian methods and the importance sampling method are reviewed in more detail than the rest.

3.3 Approximated Likelihood Methods

These methods try to approximate the true transition density using a known density. Two such methods have been reviewed in this section.

3.3.1 Simulated Likelihood Methods

The simulated likelihood method was independently proposed by Pedersen (1995b) and Santa-Clara (1995) – see also Brandt and Santa-Clara (2002). The basic idea is to approximate the true (unknown) transition density $p(s, x_s, t, x_t, \Theta)$ by a sequence of transition densities $p^{(M)}(s, x_s, t, x_t, \Theta)$ of the Euler's approximation that converge to $p(s, x_s, t, x_t, \Theta)$ as $M \to \infty$ and then to define the approximate log-likelihood functions

$$l_n^{(M)}(\Theta) = \sum_{i=1}^{n} log(p^{(M)}(t_{i-1}, y_{i-1}, t_i, y_i, \Theta)). \tag{3.6}$$

For $M = 1$, the density of the Euler approximation $p^{(1)}(s, x_s, t, x_t, \Theta)$ is not an accurate approximation of $p(s, x_s, t, x_t, \Theta)$, unless $\Delta = (t - s)$ is sufficiently small. When Δ is too large, the idea is to consider a smaller $\delta << \Delta$, for example, $\delta = \Delta / M$ for M large enough. This is done by imputing $M - 1$ latent variables $X_1, \cdots, X_{M-1}$ and simulating $x_1, \cdots, x_{M-1}$ using Euler's scheme with $x_0 = x$, and $x_M = y$ corresponding to time points $\tau_1, \cdots, \tau_{M-1}$ with $\tau_0 = s$ and $\tau_M = t$. Figure 1 illustrates the data impuation.

$p^{(M)}(s, x_s, t, x_t, \Theta)$ is then given by

$$p^{(M)}(s, x_s, t, x_t, \Theta) = \int_{\Re^{p(M-1)}} \prod_{j=1}^{M} p^{(1)}(\tau_{j-1}, x_{j-1}, \tau_j, x_j, \Theta)\, dx_1 \cdots dx_{M-1} \tag{3.7}$$

where the one-step transition density of the Euler approximation $p^{(1)}(\tau_{j-1}, x_{j-1}, \tau_j, x_j, \Theta)$ is given by

$$p^{(1)}(\tau_{j-1}, x_{j-1}, \tau_j, x_j, \Theta) = N\left(x_{j-1} + f(x_{j-1}, \tau_{j-1}, \Theta) \cdot \delta\,,\, g^2(x_{j-1}, \tau_{j-1}, \Theta) \cdot \delta\right). \tag{3.8}$$

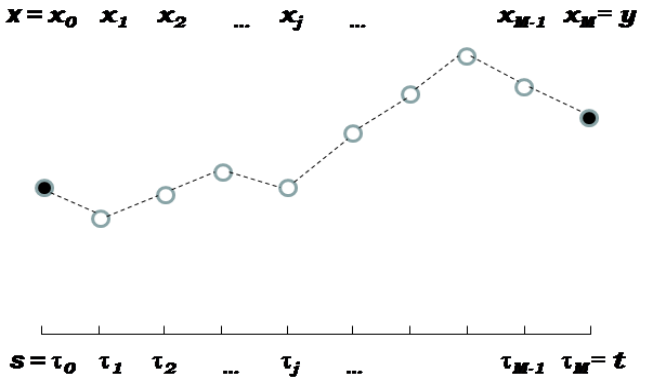


Figure 3.1: Imputing latent variables.

Using the Markovian property of the Euler approximation and the Chapman-Kolmogorov equations, Equation (3.7) can be interpreted as (see Pedersen (1995b), Theorem 1)

$$p^{(M)}(s, x_s, t, x_t, \Theta) = \mathbf{E}_{x_{M-1}}[p^{(1)}(\tau_{M-1}, x_{M-1}, t, x_t, \Theta)] \tag{3.9}$$

In practice, Monte-Carlo integration is used to approximate the integral. Thus, starting from $Y_s = x$, M trajectories of the process Y are simulated using the Euler's method and then $p^{(M)}(s, x_s, t, x_t, \Theta)$ is approximated using

$$p^{(M)}(s, x_s, t, x_t, \Theta) \approx= \frac{1}{K} \sum_{k=1}^{K} p^{(1)}(\tau_{M-1}, x_{M-1}^{(k)}, t, x_t, \Theta). \tag{3.10}$$

Strong Law of Large number implies that this Monte-Carlo approximation can be made arbitrarily accurate by using a large enough M.

Let $a(t, x, \Theta) = g(t, x, \Theta) \cdot g(t, x, \Theta)'$ denote the covariance structure for the multivariate diffusion. Then, Pedersen (1995b) proves the following results which provide the sufficient conditions under which the approximation $l_n^{(M)}(\Theta)$ becomes asymptotically exact.

Theorem 1 **:** In addition to $A\,1$ and $A\,2$ assume that for all $\Theta \in \Xi$ that:

A 6 : (**i**) : $(t, x) \longmapsto f(t, x, \Theta)$ is continuous.

(**ii**) : $a(t, x, \Theta) = a(\Theta)$ is positive definite.

Then $p(s, x_s, t, x_t, \Theta)$ exists, and for all $0 \leq s < t$, $x \in \Re^p$ and $\Theta \in \Xi$

$$p^{(M)}(s, x_s, t, \cdot, \Theta) \to p(s, x_s, t, \cdot, \Theta)$$

in $\mathbf{L}^1$ as $M \to \infty$.

The above theorem assumes $a(t, x, \Theta)$ to be independent of t and x. Pedersen (1995b) shows that it is possible to prove the above convergence by letting $a(\cdot)$ depend on x, if the SDE is assumed to be time-homogeneous as stated in the following result.

Theorem 2 **:** Assume that for all $\Theta \in \Xi$ that:

A 7 : (**i**) : f and g are time homogeneous, i.e. $f(t, x, \Theta) = f(x, \Theta)$, and $g(t, x, \Theta) = g(x, \Theta)$.

(**ii**) : $f(x, \Theta)$ and $g(x, \Theta)$ are bounded with bounded derivatives of any order.

(**iii**) : $a(x, \Theta)$ is strongly positive definite, that is there exists an $\epsilon(\Theta) > 0$ such that $a(x, \Theta) - \epsilon(\Theta)\, I_p$ is a non-negative definite for all $x \in \Re^p$.

Then $p(s, x_s, t, x_t, \Theta)$ exists, and for all $t \geq 0$, $x_s, x_t \in \Re^p$ and $\Theta \in \Xi$

$$p^{(M)}(s, x_s, t, \cdot, \Theta) \to p(s, x_s, t, \cdot, \Theta)$$

in $\mathbf{L}^1$ as $M \to \infty$. Furthermore,

$$p(s, x_s, t, x_t, \Theta) = \lim_M \inf p^{(M)}(s, x_s, t, x_t, \Theta)$$

for almost all $y \in \Re^d$, and so if $p^{(M)}(s, x_s, t, \cdot, \Theta)$ converges pointwise, then it converges to $p(s, x_s, t, \cdot, \Theta)$.

Pedersen (1995b) further states that though the assumptions in Theorem 2 are very restrictive, it should be possible to replace the boundedness condition on f and g by conditions such as $A\,1$ and $A\,2$, and also that only few derivatives of $f(\cdot,\Theta)$ and $g(\cdot,\Theta)$ are really needed for the proof.

Once the asymptotic convergence of $p^{(M)}$ to p is thus established, the main justification for using $l_n^{(M)}(\Theta)$ as a substitute for $l_n(\Theta)$ for large values of M is given by the following result:

Theorem 3 : If $p^{(M)}(s,x_s,t,\cdot,\Theta) \rightarrow p(s,x_s,t,\cdot,\Theta)$ in $\mathbf{L}^1$ as $M \rightarrow \infty$ for all $0 \leq s < t$, $x_s \in \Re^p$ and $\Theta \in \Xi$, then $l_n^{(M)}(\Theta) \xrightarrow{P} l_n(\Theta)$ as $M \rightarrow \infty$, for all $\Theta \in \Xi$ and $n \in \mathbf{M}$.

Thus, for a large class of multi-variate SDEs which satisfy the regularity conditions of either Theorem 1 or Theorem 2, the true likelihood function can be approximated arbitrarily closely for a large enough M and K using the simulated likelihood method. Further, Pedersen (1995a) proves that the maximum likelihood estimator thus obtained is consistent and asymptotically normal without requiring any assumption on the distance between the discrete observation time-points.

To implement this method in practice, $l_n^{(M)}(\Theta)$ needs to be calculated at a finite number of points $\Theta \in \Xi$. This can be done using some numerical optimization method, for example, Newton's method. Pedersen (1995b) suggests finding the initial values of Θ for a numerical maximisation by maximising $l_1^{(M)}(\Theta)$. It is then possible to compute both the Hessian as well as the gradient explicitly and thus employ a numerical optimization method. While doing so, it is advisable to use the same random numbers

ϵ_t in the Euler discretisation to simulate sample paths for each value of Θ explored. Brandt and Santa-Clara (2002) provide details for this implementation.

Though this approach has very appealing theoretical properties, as Lacus (2008) points out, in practice, it can be computationally very expensive because for each pair of observations, K trajectories of length M are needed. Usually, a large number of simulations K are needed, and M may vary from 5 to 10 for reasonable estimates. If this approximation has to be used as a function of the parameters Θ in order to get the maximum likelihood estimates of the parameters, this task may require a very large amount of time.

3.3.1.1 Using Importance Sampling

Durham and Gallant (2002) illustrated that most of the paths simulated using the Euler method method conditional on $Y_s = x_s$ miss the next observed point $Y_t = x_t$, rendering this method computationally inefficient. They proposed two different approaches to improve the efficiency for one dimensional SDE models. The first approach is to use bias reduction techniques which basically involve choosing a better approximation to the transition density than the one offered by Euler's approximation. The second approach aims to reduce the variance by using the Importance Sampling method to obtain more efficient Monte-Carlo estimates of Equation (3.9). The importance sampling approach is described in detail below.

Consider the approximation in Equation (3.9). Often, a very large number K of sample paths need to be simulated in order to get an accurate approximation rendering it numerically inefficient. One way to improve the numerical efficiency is by using the

Importance sampling method. Let $q(x_1, \cdots, x_{M-1})$ denote the proposal density to be used for the importance sampling. Let $\{\mathbf{x}_k = (x_{k,1}, \cdots, x_{k,M-1}),\ k = 1, \cdots, K\}$ be independent draws from q. Then Equation (7) can be approximated as

$$p^{(M,K)}(s, x, t, y, \Theta) = \frac{1}{K} \sum_{k=1}^{K} \frac{\prod_{j=1}^{M} p^{(1)}(\tau_{j-1}, x_{k,j-1}, \tau_j, x_{k,j}, \Theta)}{q(x_{k,1}, \cdots, x_{k,M-1})}, \tag{3.11}$$

where $x_{k,0} = x$ and $x_{k,M} = y$ for all k. Let $(X_1, \cdots, X_{M-1})$ be a random vector with density q. Consider the following assumption:

A 8 :

$$\mathbf{E}\left[\frac{\prod_{j=1}^{M} p^{(1)}(\tau_{j-1}, X_{j-1}, \tau_j, X_j, \Theta)}{q(X_1, \cdots, X_{M-1})}\right] < \infty.$$

Then under assumption **A** 8, $P^{(M,K)} \xrightarrow{a.s} P^{(M)}$ as $K \to \infty$, using the Strong Law of Large Numbers, as long as the support of $q(\cdot)$ contains the support of $p(\cdot)$.

Durham and Gallant (2002) propose three different proposal densities that could be used in the importance sampling set-up for efficient approximation of the likelihood. The first sampler is based on the Brownian bridge. The second sampler is a modification of the first sampler and is called the *Modified Brownian bridge (MBB)* sampler. The final sampler is the one proposed by Elerian et al. (2001).

Brownian Bridge Sampler : The Brownian bridge sampler for the SDE of Equation (3.1) conditioned on $Y_s = x_s$ and $Y_t = x_t$ for any $0 \leq s < t$ is given by the Euler discretisation of the following SDE:

$$d\tilde{Y}_\tau = \tilde{f}(\tilde{Y}_\tau, \tau, \Theta)\, d\tau + g(\tilde{Y}_\tau, \tau, \Theta)\, dW_\tau, \quad \tilde{Y}_s = x_s, \tilde{Y}_s = x_s \ \ 0 \leq s \leq \tau \leq t, \tag{3.12}$$

where, the drift is given by

$$\tilde{f}(x_\tau, \tau, \Theta) = \frac{x_t - x_\tau}{t - \tau}.$$

Again, using the Markov property of the Euler's approximation, the proposal density is given by

$$q(x_1, \cdots, x_{M-1}) = \prod_{j=1}^{M} q^{(1)}(\tau_{j-1}, x_{j-1}, \tau_j, x_j, \Theta), \tag{3.13}$$

with $x_0 = x_s$ and $x_M = x_t$ and where $q^{(1)}(\tau_{j-1}, x_{j-1}, \tau_j, x_j, \Theta)$ is the density of the one-step Euler's discretisation of Equation (3.12) given by

$$q^{(1)}(\tau_{j-1}, x_{j-1}, \tau_j, x_j, \Theta) = N\left(x_{j-1} + \tilde{f}(x_{j-1}, \tau_{j-1}, \Theta) \cdot \delta\,,\, g^2(x_{j-1}, \tau_{j-1}, \Theta) \cdot \delta\right). \tag{3.14}$$

Thus, $p^{(M,K)}$ of Equation (3.11) can now be evaluated where $q(x_1, \cdots, x_{M-1})$ is now given by Equation (3.13).

Note that, Equation (3.12) is a true Brownian bridge (recall Theorem 4 in Chapter 2) only if $g(\cdot)$ is unity, and that one way to achieve this is by transforming the SDE. Further, Durham and Gallant (2002) point out that because the Brownian bridge sampler has the same diffusion coefficient as the original SDE, they are locally equivalent. Therefore, the Radon-Nikodym derivative of the two measures can be obtained in closed form using Girsanov's theorem. Thus, $p^{(M,K)}$ can also be computed in this case, by using the Euler discretisation of this derivative process and they provide the details of this implementation.

Modified Brownian Bridge (MBB) Sampler : The MBB sampler is also a discretised Gaussian density which is derived as follows. Consider the distribution of X_j conditional on $X_{j-1} = x_{j-1}$ and $X_M = x_t$. This can be expressed as

$$P(X_j|x_{j-1}, x_t, \Theta) = \frac{P(x_t|X_j, \Theta) \cdot P(X_j|x_{j-1}, \Theta)}{P(x_t|x_{j-1}, \Theta)}. \tag{3.15}$$

$P(X_j|x_{j-1}, \Theta)$ can be approximated using the Euler's density as

$$P(X_j|x_{j-1}, \Theta) \approx N\left(x_{j-1} + f(x_{j-1}, \tau_{j-1}, \Theta) \cdot \delta\,,\, g^2(x_{j-1}, \tau_{j-1}, \Theta) \cdot \delta\right). \tag{3.16}$$

Similarly, considering the time lag of $(M-j+1)\cdot\delta$, $P(x_t|x_{j-1}, \Theta)$ can be approximated using Euler's density as

$$P(x_t|x_{j-1}, \Theta) \approx N\left(x_{j-1} + f(x_{j-1}, \tau_{j-1}, \Theta) \cdot (M-j+1) \cdot \delta\,,\, g^2(x_{j-1}, \tau_{j-1}, \Theta) \cdot (M-j+1) \cdot \delta\right). \tag{3.17}$$

Finally, $P(x_t|X_j, \Theta)$ is approximated using an 'Euler like' (but not Euler) approximation

$$P(x_t|X_j, \Theta) \approx N\left(X_j + f(x_{j-1}, \tau_{j-1}, \Theta) \cdot (M-j) \cdot \delta\,,\, g^2(x_{j-1}, \tau_{j-1}, \Theta) \cdot (M-j) \cdot \delta\right). \tag{3.18}$$

Then, it is possible to construct the approximate joint density of X_j and x_t (conditional upon x_{j-1}) using the multivariate Normal conditioning results (Golightly and Wilkinson (2007)) which yield

$$P(X_j, x_t|X_{j-1}) \approx N\left(\mu, \Sigma\right), \tag{3.19}$$

where

$$\mu = \begin{pmatrix} x_{j-1} + f(x_{j-1}, \tau_{j-1}, \Theta) \cdot \delta \\ x_{j-1} + f(x_{j-1}, \tau_{j-1}, \Theta) \cdot (M-j+1) \cdot \delta \end{pmatrix}$$

and

$$\Sigma = \begin{pmatrix} g^2(x_{j-1}, \tau_{j-1}, \Theta) \cdot \delta & g(x_{j-1}, \tau_{j-1}, \Theta) \cdot \sqrt{\delta} \\ g(x_{j-1}, \tau_{j-1}, \Theta) \cdot \sqrt{\delta} & g^2(x_{j-1}, \tau_{j-1}, \Theta) \cdot (M-j+1) \cdot \delta \end{pmatrix}$$

Now, by conditioning Equation (3.19) on x_t, $P(X_j|x_{j-1}, x_t, \Theta)$ can approximated as

$$P(X_j|x_{j-1}, x_t, \Theta) \approx N\left(x_{j-1} + \tilde{f}(x_{j-1}, \tau_{j-1}, \Theta)\cdot\delta\,,\, g^2(x_{j-1}, \tau_{j-1}, \Theta)\cdot\frac{(M-j)}{(M-j+1)}\delta\right), \tag{3.20}$$

where $\tilde{f}$ is the same drift as mentioned in Equation (3.12). Again, using the Markov property of the Euler's approximation, the proposal density is given by

$$q(x_1, \cdots, x_{M-1}) = \prod_{j=1}^{M} q^{(1)}(\tau_{j-1}, x_{j-1}, \tau_j, x_j, \Theta), \tag{3.21}$$

with $x_0 = x_s$ and $x_M = x_t$ and where

$$q^{(1)}(\tau_{j-1}, x_{j-1}, \tau_j, x_j, \Theta) = N\left(x_{j-1} + \tilde{f}(x_{j-1}, \tau_{j-1}, \Theta)\cdot\delta\,,\, g^2(x_{j-1}, \tau_{j-1}, \Theta)\cdot\frac{(M-j)}{(M-j+1)}\delta\right)$$

of Equation (3.20).

Thus, $p^{(M,K)}$ of Equation (3.11) can now be evaluated where $q(x_1, \cdots, x_{M-1})$ is now given by Equation (3.21).

Note that, the MBB sampler is identical to the Brownian bridge sampler except for the term $\frac{(M-j)}{(M-j+1)}$ in the diffusion coefficient. However, Durham and Gallant (2002) argues that this difference makes MBB a more efficient sampler than the Brownian bridge sampler. They point out that because of this extra term, for $j = M$, the variance of the MBB sampler is zero, making sure that the samples drawn using MBB will never miss the observed data point x_t. It is for this reason that, by using the MBB as the proposal density the simulated likelihood method can be implemented with greater computation efficiency using importance sampling.

The third importance sampler proposed by Durham and Gallant (2002) is the one proposed by Elerian et al. (2001). It is also a discretised Gaussian density. The idea is

to approximate the target density by a multivariate Gaussian density with mean and variance based on a second-order Taylor expansion of the log target density around its mode. The key feature of this sampler is that it draws paths in one shot, rather than recursively. Please refer to Elerian et al. (2001) for details.

Finally, it is important to note that for the importance sampling methods to significantly improve efficiency, it is required that the tails of the proposal distribution are not too thin. Also, as mentioned in Section 2.4.2, it is also required that the support of the importance density contains the support of the target density. See Geweke (1989) and Robert and Casella (2004) for details. Also, note that, it might be possible to improve the accuracy of the simulated likelihood method by using the transition density corresponding to higher order numerical methods such as , for example the Milstein scheme; see for Durham and Gallant (2002) for details.

3.3.2 Hermite Polynomials expansion of the likelihood

Although, the simulated likelihood method described above is widely applicable and easy to implement, it can be computationally expensive and does not provide a closed form analytical approximation to the true transition density. Ait-Sahalia (2002) proposed a method to approximate the transition density of a one dimensional diffusion process using closed form analytical approximations obtained by constructing a convergent series of Hermite polynomials.

Consider an one dimensional SDE of the same form as Equation (3.1). In addition to assumption **A** 7, following assumption is needed.

A 9 : $f(x, \Theta)$ and $g(x, \Theta)$ are three times continuously differentiable in Θ for all $x \in \Re$ and for all $\Theta \in \Xi$.

The first step in implementing this method is to transform the diffusion process X_t into a process Y_t with a constant diffusion term using the Lamperti transform

$$Y_t = h(X_t, \Theta) = \int_z^{Y_t} \frac{1}{g(u, \Theta)} du. \tag{3.22}$$

Using, Itô's formula, it can be seen that the process Y_t solves the SDE

$$dY_t = b(Y_t, \Theta)dt + dW_t, \quad Y_0 = y_0$$

where

$$b(y, \Theta) = \frac{f(h^{-1}(y), \Theta)}{g(h^{-1}(y), \Theta)} - \frac{1}{2} g_x(h^{-1}(y), \Theta),$$

where $g_x = dg(.)/dx$.

This transformed process, needs to satisfy the following assumption:

A 10 : For all $\Theta \in \Xi$, $f(y, \Theta)$ and its derivatives with respect to y and Θ have at most polynomial growth near the boundaries and satisfies the specific boundary conditions mentioned in assumption 2 of Ait-Sahalia (2002).

The next step is then to transform the process Y_t into a process Z_t, which for a fixed Δ is given as follows:

$$Z = \Delta^{-1/2}(Y - y_0).$$

Ait-Sahalia (2002) shows that for a fixed Δ, Z defined above happens to be close enough to a $N(0, 1)$ variable to make it possible to create a convergent series of expansions for its density p_Z around a $N(0, 1)$.

Let $P_Y(\Delta, y|y_0, \Theta)$ denote the conditional density of $Y_{t+\Delta}|Y_t$, and define the density function of Z as

$$p_Z(\Delta, z|y_0, \Theta) = \Delta^{1/2} p_Y(\Delta, \Delta^{1/2} z + y_0|y_0, \Theta). \tag{3.23}$$

The density p_Z can now be approximated using a Hermite series expansion as follows. The classical Hermite polynomials can be expressed as a series

$$H_j(z) = e^{z^2/2} \frac{d^j}{dz^j}[e^{-z^2/2}], \quad j \geq 0. \tag{3.24}$$

Let $\phi(z) = e^{-z^2/2}/\sqrt{2\pi}$ denote the $N(0,1)$ density function. Then define

$$p_Z^J(\Delta, z|y_0, \Theta) = \phi(z) \sum_{j=0}^{J} \eta_z^j(\Delta, y_0, \Theta)\, H_j(z) \tag{3.25}$$

as the Hermite expansion of the density function p_Z for fixed Δ, y_0, and Θ, and where the coefficients η_z^j are given by

$$\eta_z^j(\Delta, y_0, \Theta)\, H_j(z) = \frac{1}{j!} \int_{-\infty}^{\infty} H_j(z) p_Z(\Delta, z|y_0, \Theta)\, dz. \tag{3.26}$$

Therefore, the sequence of approximations to P_Y is given by

$$p_Y^J(\Delta, y|y_0, \Theta) = \Delta^{-1/2}\, p_Z^J(\Delta, \Delta^{-1/2}(y - y_0).|y_0, \Theta) \tag{3.27}$$

Finally, p_Y can be approximated using

$$p_Y^J(\Delta, x|x_0, \Theta) = g(x, \Theta)^{-1} p_Y^J(\Delta, h(x, \Theta)|h(x_0, \Theta), \Theta) \tag{3.28}$$

Then, Ait-Sahalia (2002) proves that the following theorem holds.

Theorem 4 **:** Under assumptions **A** 7, **A** 9 and **A** 10, for every $\Theta \in \Xi$ and $\Delta \in (0, \bar{\Delta})$, (see Ait-Sahalia (2002) for definition of $\bar{\Delta}$ and other details):

$$p_Y^J(\Delta, x|x_0, \Theta) \to p_Y(\Delta, x|x_0, \Theta)$$

as $j \to \infty$. In addition, the convergence is uniform in Θ over Ξ.

Ait-Sahalia (2002) also study the properties of the sequence of the MLE's obtained using p_Z^J and provide conditions under which it converges to the true MLE. Further, he also provides details on how to implement this method in practice and show empirically that, often in practice, accurate approximations can be obtained by using very small number ($J = 2$ or 3) of polynomials.

Though this method can provide very accurate inferences, however, implementing this method for multivariate SDE's may not often be possible because of the transformation of (3.21). Also, as Lacus (2008) points out, implementing this method can also become computationally expensive for many models.

3.4 Estimating Functions

When the true transition density (and hence the likelihood) of the diffusion process are explicitly known in closed form, the score function

$$S_n(\mathbf{Y}, \Theta) = \frac{\partial}{\partial \theta} l_n(\Theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log(p(t_{i-1}, y_{i-1}, t_i, y_i \Theta))$$

can be used (by solving for $S_n(\mathbf{Y}, \Theta) = 0$ for Θ) to find the MLE. For most SDE models, however, the likelihood is not available in closed form and so also is the score function. The idea behind the estimating functions approach is to build a function $f_n(\mathbf{Y}, \Theta)$ which, as Sorensen (2004) puts it, 'tries to mimic' the score function S_n, so

that the MLE can be approximated by solving $f_n(\mathbf{Y}, \Theta) = 0$ for Θ.

In order for the corresponding estimator to have the desirable asymptotic properties, it is crucial that the estimating function is unbiased and is able to distinguish the true parameter value from other values of Θ:

$$\mathbf{E}_{\Theta_0}\left[f_n(\mathbf{Y}, \Theta)\right] = 0 \quad \text{if and only if } \Theta = \Theta_0. \tag{3.29}$$

For a detailed review on estimating functions, please refer to Heyde (1997). Several different estimating functions have been proposed, two important classes of estimating functions are briefly reviewed below.

3.4.1 Martingale Estimating Functions

Martingale estimating functions are a popular class of estimating functions and their general form can be expressed as (Lacus (2008)):

$$f_n(\mathbf{Y}, \Theta) = \sum_{i=1}^{n} f(Y_{i-1}, Y_i, \Theta),$$

such that the function f_n satisfies the Martingale property:

$$\mathbf{E}_{\Theta}[f_n(\mathbf{Y}, \Theta)|\mathcal{G}_{n-1}] = f_{n-1}(\mathbf{Y}, \Theta),$$

with respect to the discrete-time filtration $\mathcal{G}_n$.

As Sorensen (2004) points out, there are at least two good reasons for looking at estimating functions that are martingales. Firstly, the score function is a martingale, and secondly, theory for Martingales can be used to prove the asymptotic behaviour of the estimators thus generated. Sorensen (1999) proves the asymptotic properties of

the such estimators.

Nevertheless, evaluating Martingale estimating functions in practice usually involves some integral of the transition density. These integrals are usually unknown and some numerical procedure must be used in practice. Martingale estimating functions therefore have the drawback of being time consuming (Kessler (2000)).

3.4.2 Simple Estimating Functions

Kessler (2000) proposed a different class of estimating functions, which do not involve numerical evaluation of intergrals in practice and are therefore faster and easier to implement than the Martingale estimating functions. These functions are known as *simple estimating functions*.

A *Simple estimating function* can be defined as (see Sorensen (2004),Kessler (2000) for e.g.) as an estimating function of the form

$$f_n(\mathbf{Y}, \Theta) = \sum_{i=1}^{n} f(Y_{i-1}, \Theta),$$

where the function f takes only one state variable as argument, and satisfies the assumption $\int f(x, \Theta)\, d\pi(\Theta) = 0$, where $\pi(\Theta)$ is the invariant distribution of the ergodic diffusion process Y.

For such estimating functions, the crucial condition (3.29) simplifies to

$$\mathbf{E}_{\Theta_0}\left[f(x_0, \Theta)\right] = 0 \quad \text{if and only if } \Theta = \Theta_0.$$

As Kessler (2000) points out, in addition to the computational advantage, simple estimating functions also have an advantage that they can be generated simply by

choosing $f = \mathbf{L}_\Theta\, h$, where $\mathbf{L}_\Theta$ is the infinitesimal generator of the ergodic diffusion process given by

$$\mathbf{L}_\Theta\, h(x,\Theta) = f(x,\Theta)\, h_x(x,\Theta) + \frac{1}{2}\, g^2(x,\Theta)\, h_{xx}(x,\Theta)$$

where $h(x,\Theta)$ is twice continuously differentiable function (w.r.t Y) and h_x and h_{xx} denote, respectively, its first and second derivatives. See Lacus (2008) for details regarding the implementation of this method.

However, Sorensen (2004) points out that simple estimating functions can not be used on diffusion processes which are not ergodic, or for which the data has not yet reached stationarity.

Many different kinds of estimating functions have been proposed, for example, estimating functions based on polynomials, or based on eigenfunctions, etc. Not all of them are either Martingales or *simple*. For a detailed review on estimating functions refer to Prakasa Rao (1999) and Lacus (2008).

3.5 Other Approaches

Several different approaches have been proposed to obtain inference on diffusion process parameters using discretely observed data that have not been reviewed here. Principally among them are: approximating the true transition density by numerically solving the Kolmogorov forward equations of the diffusion process, methods of *indirect inference* which try to approximate the inference by using an auxiliary model, and the *generalised method of moments* which is a generalisation of the method of moments that is based on matching of theoretical moments and sample moments. Finally, note

that there are also several methods proposed for the *non-parametric* inference.

The reader is referred to Prakasa Rao (1999), Lacus (2008) and Sorensen (2004) for a detailed review of these methods.

3.6 Bayesian Inference

The common theme underlying many of the methods developed thus far to obtain Bayesian inference on discretely observed diffusions is to interpret the inference problem as that of a hidden Markov model where the unobserved Markov process forms the set of latent variables. Most of these methods are Monte-Carlo based methods which rely on drawing large number of samples of both the latent variables as well as the diffusion process parameters. Many of these methods sample only the discretised paths of the unobserved diffusion, whereas some others can sample the continuous paths. A lot of research has focused on two main aspects: (i) developing Monte-Carlo algorithms which get around the problem of dependency between the latent variables and the diffusion parameters, and (ii) developing more efficient and more accurate proposals for the latent variables which in turn results in more accurate and efficient inference on the diffusion process parameters. This section provides a quick overview of some of the important methods developed thus far.

The inference set-up used by Bayesian methods based on time discretisation is very similar to the set-up used by the simulated likelihood methods described in section 3.3.1. The diffusion process is observed at time points $t_0, t_1, \cdots, t_n$ and $\mathbf{Y} = \{y_0, y_1, \cdots, y_n\}$ denote these observations. $M-1$ latent variables are imputed between every pair of consecutive observations. These latent variables are simulated using some

simulation mechanism (for e.g. the Euler's method). This is typically achieved by partitioning each interval $[t_i, t_{i+1}]$ into M equal parts as $t_i = \tau_{0_i} < \tau_{1_i} < \cdots < \tau_{M_i} = t_{i+1}$, such that $\tau_{(j+1)_i} - \tau_{j_i} = \delta_\tau$ is small enough for the simulation mechanism to be reasonably accurate. This is illustrated in Figure 3.2. Equal spacing is not necessary but has been used only to make the notation easier. Let $\mathbf{X^{(i)}} = X_{1_i}, X_{2_i}, \ldots, X_{(M-1)_i}$ denote the latent variables corresponding to times $\tau_{1_i}, \tau_{2_i} < \cdots < \tau_{(M-1)_i}$, between the observations y_i and y_{i+1} for $i = 0, 1, \cdots, n-1$.

3.6.1 MCMC based methods

One of the earliest MCMC based methods was proposed by Elerian et al. (2001) and consists of the following the following *Metropolis within Gibbs* updating scheme.

Metropolis within Gibbs scheme of Elerian et al. (2001) :

Step 1**:** Initialise $\mathbf{X} = \{\mathbf{X^{(0)}}, \cdots, \mathbf{X^{(n-1)}}\}$ and Θ.

Step 2**:** Update $\mathbf{X^{(i)}}$ from $P(\mathbf{X^{(i)}}|y_i, y_{i+1}, \Theta)$ for $i = 0, \cdots, n-1$.

Step 3**:** Update Θ from $P(\Theta|\mathbf{X}, \mathbf{Y})$.

Step 4**:** Repeat steps 2 and 3 until the chains satisfy a convergence criteria.

Elerian et al. (2001) proposed the use of Euler's method for simulating the latent variables $\mathbf{X}$ in Step 2 and argued that since the paths simulated using the Euler's method converge weakly to the true diffusion, by using a large M, the inference obtained using this MCMC method would become more and more accurate. The proposal density for $\mathbf{X}$ was obtained by approximating the target density at the mode using

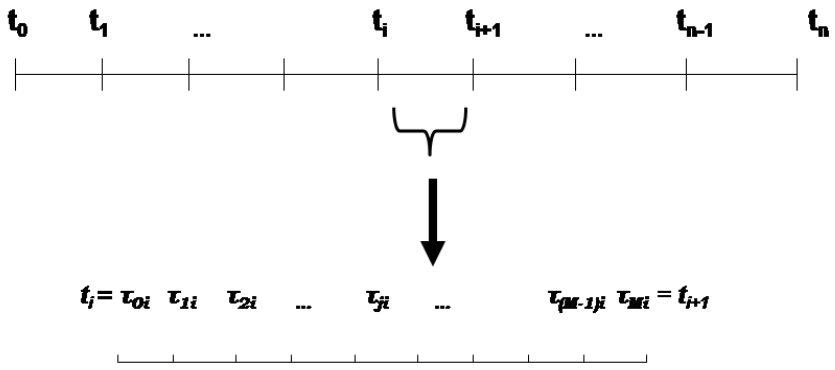


Figure 3.2: Latent data points between each pair of observed data.

a multivariate Gaussian or a multivariate t-distribution using the Newton-Raphson method. The reader is referred to their paper for details. They also suggested updating $\mathbf{X}$ in blocks to improve efficiency. This method appears to be very intuitive and appealing. Chib and Shephard (2002) proposed implementing this algorithm by using the density of the modified Brownian bridge (MBB) construct proposed by Durham and Gallant (2002) as a proposal distribution for the latent variables $\mathbf{X}$.

However, Roberts and Stramer (2001) pointed out that the quadratic variation of the process completely determines the diffusion coefficient (recall Equation (3.2)). Therefore, as $M \to \infty$ the data augmentation scheme in Gibbs sampling based methods (such as that of Elerian et al. (2001)) is reducible; the imputed data merely confirm the current value of the diffusion coefficient which then remains unaltered since it is in turn determined by the quadratic variation of the sample path of $\mathbf{X}$.

Roberts and Stramer (2001) proposed that this dependency problem can be avoided by first transforming the SDE using the Lamperti transform to obtain the SDE with a unit diffusion coefficient and then implementing the MCMC scheme on this trans-

formed process. They also proposed using an independence sampler based either on a Brownian bridge density or on the OU bridge density as a proposal density for $\mathbf{X}$. Though, this approach gets around the problem of dependency between $\mathbf{X}$ and the diffusion coefficients, the required transformation may not be possible for many non-linear multivariate SDEs (Golightly and Wilkinson (2007)).

Chib et al. (2006) proposed an innovative solution to eliminate the issue of dependency. They suggested sampling the Wiener process components W_t instead of the latent variables $\mathbf{X}$ and then constructing the $\mathbf{X}$'s deterministically using the Euler approximation. Since the Wiener process is independent of the diffusion parameters σ, there is no dependency and the MCMC scheme can now be used without the need of any transformation.

Golightly and Wilkinson (2007) instead used the modified Brownian bridge (MBB) construct of Section 3.3.1.1 to deterministically obtain $\mathbf{X}$'s instead of using the Euler's method. The advantage of doing this, they argue, is that unlike the Euler's method, the paths generated by MBB do not miss the observations and hence this becomes a more efficient method of reparameterizing the latent variables.

Re-parameterisation scheme of Golightly and Wilkinson (2007) :

Step 1: Initialise $\mathbf{X} = \{\mathbf{X^{(0)}}, \cdots, \mathbf{X^{(n-1)}}\}$ and Θ.

Step 2: Update $\mathbf{X^{(i)}}$ from $P(\mathbf{X^{(i)}}|y_i, y_{i+1}, \Theta)$ for $i = 0, \cdots, n-1$.

Step 3: Using the MBB of Equation (3.20) determine $W^{(i)}$ deterministically as

$$W_{j_i} - W_{j-1_i} = \left(g(x_{j-1}, \tau_{j-1}, \Theta) \cdot \sqrt{\frac{(M-j)}{(M-j+1)}\delta} \right)^{-1} \cdot \Big((X_{j_i} - X_{j-1_i}) - \tilde{f}(x_{j-1}, \tau_{j-1}, \Theta) \cdot \delta \Big).$$

Step 4**:** Update Θ from $P(\Theta|\mathbf{W}, \mathbf{Y})$.

Step 5**:** Repeat steps 2, 3 and 4 until the chains satisfy a convergence criteria.

Both these methods claim to in fact improve with m rather than worsen. Though these methods do get around the problem of dependency, it is at the cost of added complexity to the existing MCMC algorithms and are as a result computationally expensive.

3.6.2 Other Bayesian Methods

Filtering based methods have also been proposed; see for e.g. Del Moral et al. (2002), Golightly and Wilkinson (2006) and Sarkka and Sottinen (2008). These methods have an advantage that the inference framework does not need to be restarted from scratch as the new data becomes available. However, computation efficiency of such methods is an ongoing research problem (Golightly and Wilkinson (2006)). Also, methods which use a diffusion specific importance sampler (e.g Sarkka and Sottinen (2008)) are less widely applicable than some of the methods described above.

Finally, it is important to note that methods which can simulate from the exact density of the diffusion have been proposed (Beskos et al. (2006), Beskos et al. (2008), and Fernhead et al. (2010)). These methods can be applied to a wide range (though, not all) of diffusion processes. The advantage that these methods have over the time-discretisation based methods described above is that these methods are free of the discretisation error introduced by discretising the underlying continuous time process.

3.6.3 Illustration of an MCMC method

This section illustrates the use of an MCMC method for SDE models using a toy problem. Results obtained using this MCMC scheme will serve as a benchmark against which the results obtained using GaMBA can be compared, both in terms of the accuracy as well as the computational efficiency.

A standard MCMC based method which can be used for SDE models has now been described. It is mainly based on the method described in Chib and Shephard (2002) which is an extension of the method proposed by Elerian et al. (2001). It is a straightforward method which does not incorporate modifications suggested in Chib et al. (2006) or Golightly and Wilkinson (2007) and this makes it simpler and computationally cheaper to implement than the modified methods. The issue of dependency between the latent variables and the diffusion coefficient can be dealt with by transforming the process as suggested by Roberts and Stramer (2001).

Consider the SDE of Equation (3.1). Assume that the diffusion coefficient is not constant. Then in order to avoid the problem of dependence between the latent variables and the diffusion coefficients , the process X_t is transformed into a process $Y_t = h(X_t)$ with unitary diffusion coefficient using the Itô's formula. This transformation is described in Section 2.1.3. The new process is given by

$$dY_t = \left[\frac{f(h^{-1}(Y_t),\mu)}{g(h^{-1}(Y_t),\sigma)} - \frac{1}{2}g'(h^{-1}(Y_t),\sigma)\right]dt + dW_t. \tag{3.30}$$

After transforming the SDE an MCMC scheme based on Chib and Shephard (2002) is implemented for a fixed M such that all the $M-1$ latent variables between every pair

of consecutive observations are updated in one block. It is described below. Between every pair of consecutive observations $\{Y_i, Y_{i+1}\}$, $M-1$ equally spaced latent variables denoted by $\mathbf{X}^{(i)} = \{X_{1_i}, \cdots, X_{(M-1)_i}\}$ are imputed. The entire set of latent variables for $n+1$ observations $\{Y_0, Y_1, \cdots, Y_n\}$ is denoted by $\mathbf{X} = \{\mathbf{X}^{(i)}\}_{i=1,\cdots,n-1}$.

MCMC algorithm :

Step I: Initialize Ξ, and generate intial $\mathbf{X}$'s using Modified Brownian Bridge density $P_{MBB}(\mathbf{X}^{(i)}|Y_i, Y_{i+1}, \Xi)$

Step II: For $i = 0, 1, ..., (n-1)$,

(a) Propose $\mathbf{X}^{(i)*}$ using Modified Brownian Bridge density $P_{MBB}(\mathbf{X}^{(i)*}|Y_i, Y_{i+1}, \Xi)$

(b) Accept $\mathbf{X}^{(i)*}$ with probability

$$\alpha = \min\left(\left[\frac{P_{MBB}(\mathbf{X}^{(i)}|Y_i, Y_{i+1}, \Xi) \cdot P_E(Y_{i+1}|\mathbf{X}^{(i)*}, Y_i)}{P_{MBB}(\mathbf{X}^{(i)*}|Y_i, Y_{i+1}, \Xi) \cdot P_E(Y_{i+1}|\mathbf{X}^{(i)}, Y_i)}\right], 1\right)$$

Step III: Propose Ξ^* using a proposal density $Q(\Xi \to \Xi^*)$ and accept Ξ^* using

$$\beta = \min\left(\left[\frac{P_E(\mathbf{X}, \mathbf{Y}|\Xi^*) \cdot P(\Xi^*) \cdot Q(\Xi^* \to \Xi)}{P_E(\mathbf{X}, \mathbf{Y}|\Xi) \cdot P(\Xi) \cdot Q(\Xi \to \Xi^*)}\right], 1\right)$$

Step IV: Repeat steps **II** and **III** until the chains satisfy a convergence criteria.

P_E refers to the density corresponding to the Euler approximation. The implementation of this MCMC algorithm is now illustrated by using the following toy example.

Example: Euro-Dollar interest rate data

The Cox-Ingersoll-Ross (CIR) Process is the solution to the stochastic differential equation

$$dX_t = (\theta_1 - \theta_2 X_t)\, dt + \theta_3 \sqrt{X_t}\, dW_t \tag{3.31}$$

with $\theta_1, \theta_2, \theta_3 \in \Re^+$. The transition density for this process is known and is non-central Chi-squared. This process is used in financial applications to model interest rates.

This process is now used to model the simulated Euro-Dollar interest rate data. This data set of size 100 has been simulated using parameter values $\theta_1 = 0.00036$, $\theta_2 = 0.0047$, $\theta_3 = 0.012$, $m = 10$, and $Y_0 = 8$. These chosen parameter values are the posterior means obtained by Roberts and Stramer (2001) after analysing a real-life Euro-Dollar interest rate data. Figure 3.3 shows the simulated data.

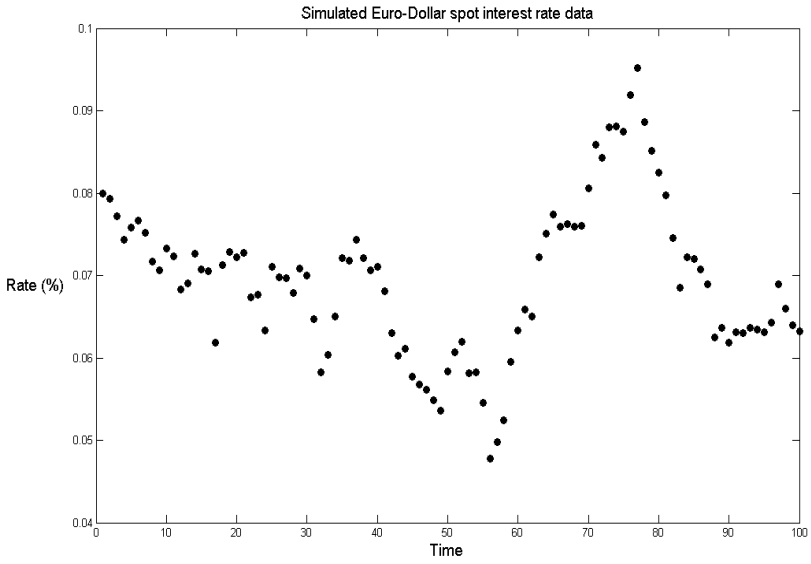


Figure 3.3: Simulated Euro-Dollar interest rate data.

The CIR process can be transformed to a process $Y_t = 2\sqrt{X_t}/\theta_3$. Using Itô's formula the new SDE has a constant diffusion coefficient and is given by

$$dY_t = \left[\frac{(\theta_1 - \theta_2 X_t)\sqrt{X_t}}{\theta_3} - \frac{\theta_3 X_t^{-1/2}}{4} \right] + dW_t. \tag{3.32}$$

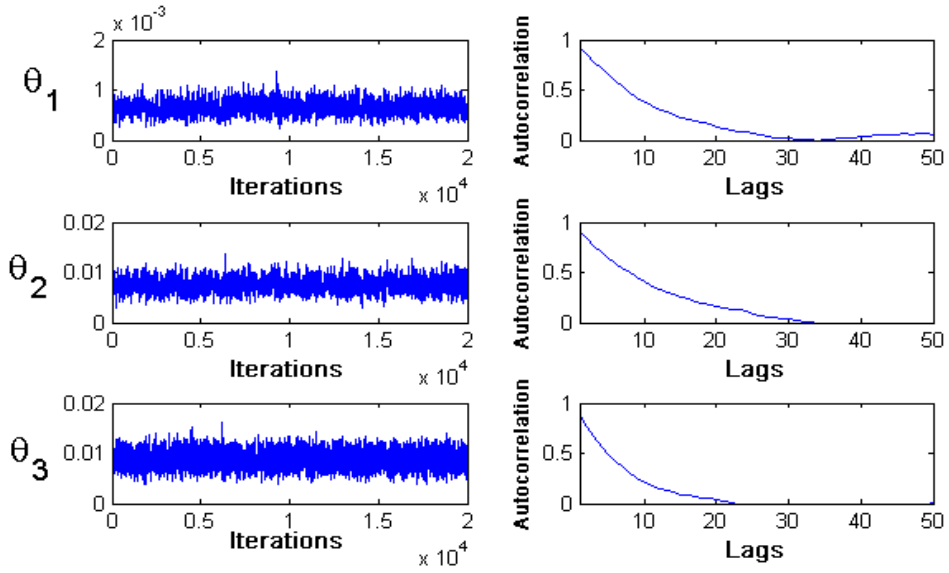


Figure 3.4: MCMC chains for the Euro-Dollar interest rate data along with their Correlograms.

The MCMC algorithm can now be implemented. Visual check of the MCMC trace plot along with the correlograms were used to assess stationarity. After discarding the first $10,000$ samples as a 'burn-in' period, the next $10,000$ samples were considered to be correlated draws from the stationary distribution. Figure 3.4 shows the MCMC chains along with their correlograms, and Figure 3.5 the marginal posteriors. The vertical lines in Figure 3.5 indicate the true values used to simulated this data. Thus, it can be seen that, after transforming the SDE, this basic MCMC based method yields accurate results. However, this implementation (using MATLAB 7.5.0) takes about 14 minutes on a standard personal computer.

3.6.4 Need for a New Method

Thus, it can be seen that although the MCMC based methods can be used for accurate Bayesian inference on SDE models, in general they are computationally expensive. In

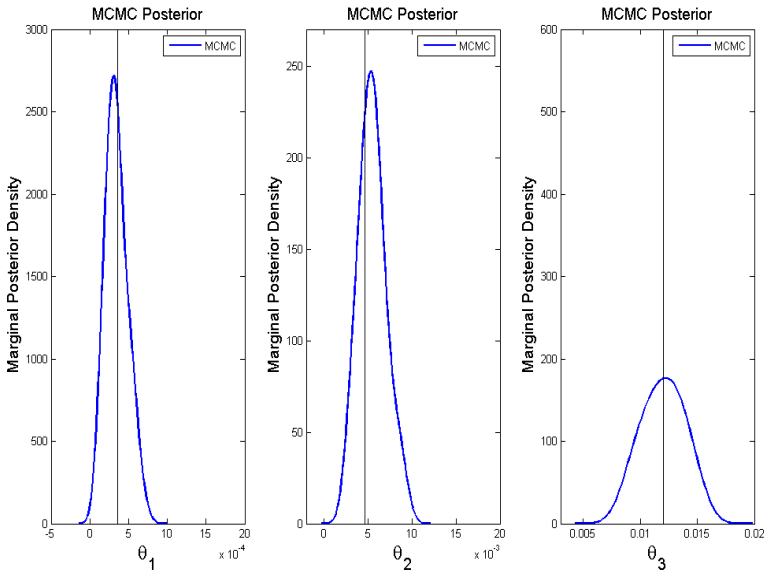


Figure 3.5: MCMC posteriors for the Euro-Dollar interest rate data - vertical lines indicating the true value of the parameter.

fact, a simple application of an MCMC method on an SDE model can be very inefficient and is not advisable. One option is to first transform the SDE before using a Gibbs sampler based method. But such a transformation is usually not possible for many multi-dimensional SDEs. The other option is to use some other form of reparameterisation that gets around the *dependency problem*. However methods which use this option are computationally very expensive and complicated to implement in practice.

Thus, there is a need to evaluate if some non-MCMC based approach could be applicable to a wide range of SDE models and used to obtain computationally efficient but accurate inference on SDE models. This was the motivation behind this PhD research. The methodology thus developed has been described in Chapter 4.

Chapter 4

Efficient Bayesian Inference for Stochastic Differential Equation Models

4.1 Introduction

Stochastic differential equations (SDEs) and related theory have been reviewed in Chapter 2. A broad overview of the statistical inference methods for SDE models has been provided in Chapter 3. This is the main chapter of this thesis. It describes the new method for statistical inference developed as part of this research work.

For the purposes of this chapter, the following has been assumed.

- It is assumed that a *weak* solution to Equation (3.1) exists. In particular, it is assumed that the $f(\cdot)$ and $g(\cdot)$ satisfy the following:

  A 1 **:** For any $R > 0$ and all $x, y \in \Re^p$ such that $|x| < R$, $|y| < R$, and $t \in [0, T]$,

there exists a constant $K_R < \infty$ such that

$$|f(x,t) - f(y,t)| + |g(x,t) - g(y,t)| < K_R|x - y|,$$

A 2 : For all $x, y \in \Re^p$ and $t \in [0,T]$, there exists a constant $C < \infty$ such that

$$|f(x,t)| + |g(x,t)| < C(1 + |x|).$$

- It is also assumed that the following holds:

  A 7 : (**i**) : f and g are time homogeneous, i.e. $f(x,t,\Theta) = f(x,\Theta)$, and $g(x,t,\Theta) = g(x,\Theta)$.

  (**ii**) : $f(x,\Theta)$ and $g(x,\Theta)$ are bounded with bounded derivatives of any order.

  (**iii**) : $a(x,\Theta) = g(x,\Theta) \cdot g(x,\Theta)'$ is strongly positive definite, that is there exists an $\epsilon(\Theta) > 0$ such that $a(x,\Theta) - \epsilon(\Theta)\, I_p$ is a non-negative definite for all $x \in \Re^p$.

The statistical interest in SDE modeling centers around the inference on the parameters which govern the drift and diffusion coefficients. To make this dependence explicit, the general form of a one-dimensional SDE can be expressed as:

$$dX_t = f(X_t, \mu)\, dt + g(X_t, \sigma)\, dW_t, \quad X(t_0) = x_0 \tag{4.1}$$

where $f : \Re \times \Re \to \Re$ is the *drift* of the SDE, $g : \Re \times \Re \to \Re$ is the *diffusion* of the SDE, and W_t is a one-dimensional process having independent scalar Wiener Process components. The stochastic process $\{X_t(w)\}$ defined on the probability space $(\Omega, \mathcal{F}_t, P)$, which satisfies Equation (4.1) is called a (Itô) *diffusion process.* If such a process exists, then it is a continuous time Markov process with continuous sample paths *a.s.* Its transition density is governed by the parameters $\Theta = \{\mu, \sigma\}$ of the SDE. Note that μ

and σ could be vector valued.

As attempts are being made to make models more realistic by introducing natural inherent uncertainties, SDE models are being increasingly used to model real life phenomena. However, not only are such models often non-trivial to solve, the statistical inference on their parameters Θ is also not straightforward in most cases.

This chapter focuses on the Bayesian inference for diffusion process parameters. This research area has already seen a lot of activity in the last few years which has been reviewed in Section 3.6. But despite this research activity, there remains a significant need for accurate but computationally cheaper methods. Proposed here is a new method of approximate Bayesian inference for diffusion process parameters, named as GaMBA (for '*Gaussian Modified Bridge Approximation*') and its extension called GaMBA-I (for *GaMBA-Importance sampling*). The objective behind developing these methods is to develop a method which is easy to implement and computationally cheaper compared to the existing alternatives.

This chapter is organised as follows. Section 4.2 introduces the basic idea behind GaMBA. Sections 4.3 and 4.4 describe in detail how this could be actually implemented in practice. Section 4.5 describes the extension GaMBA-I, and describes the conditions under which the posterior obtained using GaMBA-I would have the desirable convergence properties. While Section 4.6 provides some examples to illustrate the use of GaMBA and GaMBA-I for Bayesian inference on standard SDE models, Section 4.7 illustrates situations where GaMBA and GaMBA-I would not yield efficient inference. Finally Section 4.8 discusses various practical aspects concerning GaMBA and GaMBA-I including their limitations.

4.2 GaMBA : The Basic Idea

For the purposes of this thesis, it is assumed that the diffusion process is observed only at discrete points in time and that it is observed without error.

As described in Section 3.6 , the common theme underlying many of the methods developed thus far to obtain Bayesian inference on discretely observed diffusions is to interpret the inference problem as that of a hidden Markov model where the unobserved Markov process forms the set of latent variables. GaMBA uses this very set-up. As earlier described in section 3.6, under this setting, the observations are denoted by $\mathbf{Y} = \{y_0, y_1, \cdots, y_n\}$. $M-1$ latent variables are imputed between every pair of consecutive observations by partitioning each interval $[t_i, t_{i+1}]$ into M equal parts as $t_i = \tau_{0_i} < \tau_{1_i} < \cdots < \tau_{M_i} = t_{i+1}$, such that $\tau_{(j+1)_i} - \tau_{j_i} = \delta_\tau$ is as small as possible. This is illustrated in Figure 3.2. Equal spacing is not necessary but has been used only to make the notation easier. Let $\mathbf{X^{(i)}} = \{X_{1_i}, X_{2_i}, \ldots, X_{(M-1)_i}\}$ denote the latent variables corresponding to times $\tau_{1_i}, \tau_{2_i} < \cdots < \tau_{(M-1)_i}$, between the observations y_i and y_{i+1} for $i = 0, 1, \cdots, n-1$. The entire set of latent variables is denoted by $\mathbf{X} = \{\mathbf{X^{(0)}}, \mathbf{X^{(1)}}, \cdots, \mathbf{X^{(n-1)}}\}$. As described in Section 4.1, the diffusion parameters are denoted by Θ.

For such a set-up, the joint posterior can be expressed as:

$$P(\mathbf{X}, \Theta \,|\mathbf{Y}) \propto P(\mathbf{Y}, \mathbf{X} \,|\Theta) \cdot P(\Theta), \tag{4.2}$$

where $P(\Theta)$ is an appropriate prior distribution. The joint posterior can be factorised

as:

$$P(\mathbf{X}, \Theta \,|\mathbf{Y}) = P(\mathbf{X} \,|\Theta, \mathbf{Y}) \cdot P(\Theta \,|\mathbf{Y}). \tag{4.3}$$

Thus, using Equations 4.2 and 4.3, the posterior $P(\Theta|\mathbf{Y})$ can be approximated, up to the proportionality constant by using the following identity:

$$P(\Theta|\mathbf{Y}) \;\; \propto \;\; \frac{P(\mathbf{Y}, \mathbf{X}|\Theta) \cdot P(\Theta)}{P(\mathbf{X}|\Theta, \mathbf{Y})}, \tag{4.4}$$

valid for any $\mathbf{X} \sim P(\mathbf{X}|\Theta, \mathbf{Y})$. Equation 4.4 can be used to obtain Bayesian inference on parameters Θ. However, evaluating Equation (4.4) analytically is usually too complicated in practice.

Let Θ^* be the (unknown) modal value of the posterior density $P(\Theta|\mathbf{Y})$, and Ξ^* be the subspace of the parameter space Ξ containing Θ^* such that $\int_{\Xi \setminus \Xi^*} P(\Theta|Y) d\Theta \approx 0$. If for a given value of Θ, say Θ^o, it is possible to compute the right hand side of Equation (4.4), then the posterior probability of Θ^o given the observed data $\mathbf{Y}$ can thus be obtained up to a proportionality constant. Repeating this procedure for a large sample of Θ values carefully sampled from the subspace Ξ^* would thus give an approximation to the unknown posterior density $P(\Theta|\mathbf{Y})$. One way to efficiently sample from Ξ^* is by using a grid-sampling method; i.e. by approximating the continuous Ξ^* space using a grid $\mathcal{G}_{\Xi^*}$ consisting of finite number of points. This is the basic idea behind GaMBA. This basic idea can be summarised in the following procedure.

GaMBA : Inference Procedure

1. Determine Ξ^*.

2. Define a discrete grid $\mathcal{G}_{\Xi^*}$ on Ξ^*.

3. For each point on the grid $\Theta_j \in \mathcal{G}_{\Xi^*}$,

   - Sample $\mathbf{X}$ from $P(\mathbf{X}|\Theta_j, \mathbf{Y})$,
   - For this $\mathbf{X}$ evaluate

$$P(\Theta_j|\mathbf{Y}) \propto \left. \frac{P(\mathbf{Y}, \mathbf{X}|\Theta_j) \cdot P(\Theta_j)}{P(\mathbf{X}|\Theta_j, \mathbf{Y})} \right|_{\mathbf{X}}$$

   .

4. Normalise to obtain $P(\Theta|\mathbf{Y})$ over $\mathcal{G}_{\Xi^*}$.

While trying to put this simple idea into practice though, following practical considerations need to be made.

4.2.1 Identifying Ξ^*

In the implementation of GaMBA, identifying Ξ^* is a very important non-trivial step. This is because, if the support is not identified accurately, then the inference thus obtained can not be correct. If there is credible prior information, then Ξ^* can be determined using this information. In the absence such information, one approach to determine Ξ^* can possibly be to first find the mode of Equation (4.4) using some numerical method such as for example, Newton's method, and then approximate $P(\Theta/\mathbf{Y})$ using Laplace's approximation around this mode by computing the Hessian. It is then possible to determine Ξ^* such that $\int_{\Xi \setminus \Xi^*} P(\Theta|Y)\, d\Theta \leq \epsilon$ for some pre-determined small value $\epsilon \geq 0$. Such a procedure has been used obtain Ξ^* for latent GMRF models by Rue et al. (2009). Note that, this is only one way to determine Ξ^*; there possibly can not be just one 'right' Ξ^*. Also note that, this method may not be suitable when $P(\Theta/\mathbf{Y})$ is likely to be multi-modal.

Identification of Ξ^* is a very important research problem in itself, but this is not the focus of this research work. The main aim of this research work is to explore how step 3 of the inference procedure mentioned above can be implemented for SDE models. The key question that this thesis aims to answer is, how to evaluate $P(\mathbf{Y}, \mathbf{X}|\Theta)$ and $P(\mathbf{X}|\Theta, \mathbf{Y})$ either exactly or approximately? Further, if an approximation is possible, then how its accuracy could be determined? Some possible answers to these questions are suggested in the next section.

For the purposes of this thesis we assume that credible prior information is available and that it is possible to identify Ξ^* based on this information.

4.2.2 Evaluating $P(\mathbf{Y}, \mathbf{X}|\Theta)$

The complete likelihood $P(\mathbf{Y}, \mathbf{X}|\Theta)$ can be factorised as

$$P(\mathbf{Y}, \mathbf{X}|\Theta) = P(y_0|\Theta) \cdot \prod_{i=1}^{n} P(y_i|X_{M-1_{i-1}}, \Theta) \cdot \prod_{i=1}^{n} P(X_{1_{i-1}}|y_{i-1}, \Theta) \cdot \prod_{i=1}^{n} \prod_{j=2}^{M-1} P(X_{j_{i-1}}|X_{j-1_{i-1}}, \Theta). \tag{4.5}$$

Since, y_0 is considered observed, $P(y_0|\Theta)$ can be considered as constant and thus

$$P(\mathbf{Y}, \mathbf{X}|\Theta) \propto \prod_{i=1}^{n} P(y_i|X_{M-1_{i-1}}, \Theta) \cdot \prod_{i=1}^{n} P(X_{1_{i-1}}|y_{i-1}, \Theta) \cdot \prod_{i=1}^{n} \prod_{j=2}^{M-1} P(X_{j_{i-1}}|X_{j-1_{i-1}}, \Theta). \tag{4.6}$$

Equation (4.6) can be evaluated exactly if the exact transition density is known. However in many cases, the exact transition density is not available, and in such cases, Equation (4.6) can be approximated using Euler's density up to the unknown constant

$P(y_0|\Theta)$ as

$$P(\mathbf{Y},\mathbf{X}|\Theta) \approx \prod_{i=1}^{n} P_E(y_i|X_{M-1_{i-1}},\Theta)\cdot\prod_{i=1}^{n} P_E(X_{1_{i-1}}|y_{i-1},\Theta)\cdot\prod_{i=1}^{n}\prod_{j=2}^{M-1} P_E(X_{j_{i-1}}|X_{j-1_{i-1}},\Theta), \tag{4.7}$$

where P_E is the Euler density of Equation (2.16). Section 2.3 describes the conditions under which such an approximation is weakly convergent and also the rate of this convergence.

4.2.3 Sampling $\mathbf{X} \sim P(\mathbf{X}|\Theta,\mathbf{Y})$

Equation 4.4 is valid for any $\mathbf{X}$ sampled from its full conditional distribution $P(\mathbf{X}|\Theta,\mathbf{Y})$. However, for most of the SDE models used in practice, $P(\mathbf{X}|\Theta,\mathbf{Y})$ and $P(\mathbf{Y},\mathbf{X}|\Theta)$ are not known in closed form and need to be approximated. $P(\mathbf{Y},\mathbf{X}|\Theta)$ can be approximated using the Euler approximation as described above. Section 4.3.1 will illustrate why an approximation to $P(\mathbf{X}|\Theta,\mathbf{Y})$ can not be obtained by conditioning the Euler density and a completely different approximation (say P_B) is required instead. Sampling an $\mathbf{X}$ randomly from $P_B(\mathbf{X}|\Theta,\mathbf{Y})$ can make the inference unreliable in the sense that repeating GaMBA on the same set of data and same Ξ^* can yield different posteriors. This is specially true when P_B and P_E do not have the same mode and also if the tails of P_B are not thicker than those of P_E. Figure 4.1 provides a simplistic illustration of this.

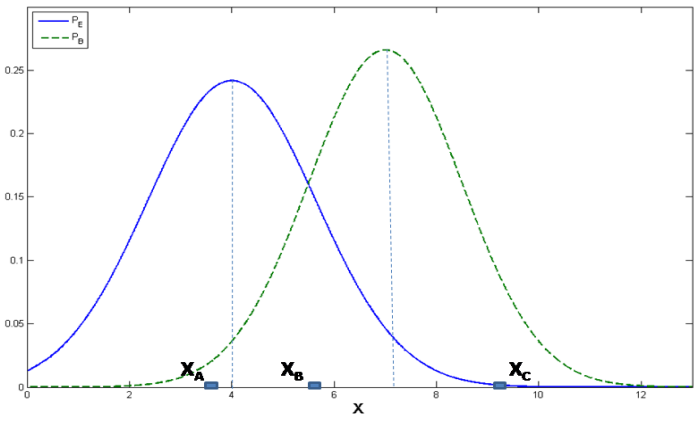


Figure 4.1: Simplistic illustration of P_E(line) and P_B (dashed) densities with different modes.

The ratio $P_E(\mathbf{Y}, \mathbf{X}|\Theta)/P_B(\mathbf{X}|\Theta, \mathbf{Y})$ which determines the likelihood component of the equation (4.4) will yield completely different values if the sampled $\mathbf{X}$ was (for example) X_A or X_B or X_C. Specifically, in this case the following would hold:

$$\frac{P_E(\mathbf{Y}, \mathbf{X}_A|\Theta)}{P_B(\mathbf{X}_A|\Theta, \mathbf{Y})} > \frac{P_E(\mathbf{Y}, \mathbf{X}_B|\Theta)}{P_B(\mathbf{X}_B|\Theta, \mathbf{Y})} \approx 1 > \frac{P_E(\mathbf{Y}, \mathbf{X}_C|\Theta)}{P_B(\mathbf{X}_C|\Theta, \mathbf{Y})}.$$

GaMBA posterior obtained using (say) $\mathbf{X}_A$ would be very different from the one obtained using (say) $\mathbf{X}_B$ and so on. Therefore, as a practical consideration, it is desirable to identify a mechanism of sampling $\mathbf{X}$ which results in reliable inference using GaMBA.

This thesis explores two possible ways in which this could be achieved. The first approach of doing this as described in Section 4.4 is to choose the $\mathbf{X}$ which maximises $P_B(\mathbf{X}|\Theta, \mathbf{Y})$ over $\mathbf{X}$, i.e. the modal value X^* of the density $P_B(\mathbf{X}|\Theta, \mathbf{Y})$. Not only is this approach computationally cheaper than the second approach, but it also has an additional advantage that when choosing not to 'integrate out' the $\mathbf{X}$'s, it is intuitively appealing to instead evaluate the expressions at the modal value of $P_B(\mathbf{X}|\mathbf{Y}, \Theta)$ rather than at any random X. The reader is referred to Rue et al. (2009) where this approach has been successfully used.

The second approach considered in this thesis is to sample several (say K) $\mathbf{X}$'s instead of just one and then approximate the likelihood using the average

$$\frac{1}{K}\sum_{k=1}^{K}\frac{P_E(\mathbf{Y},\mathbf{X}_k|\Theta)}{P_B(\mathbf{X}_k|\Theta,\mathbf{Y})}.$$

This approach is referred to as GaMBA-I and is described in Section 4.5 where its link with importance sampling is also established. Though this approach is computationally less efficient compared to the first one, it is possible to prove the consistency properties of GaMBA-I if certain conditions are met. Note that GaMBA-I is a stochastic approach unlike GaMBA which is deterministic.

4.3 Evaluating $P(\mathbf{X}|\Theta,\mathbf{Y})$

Because of the Markovian nature of a diffusion process, each observation is conditionally independent to other observations given its previous observation. Thus when concerned with the distribution of $P(\mathbf{X}|\mathbf{Y},\Theta)$, one is in fact dealing with n independent discretised diffusion bridges, each conditioned only on the corresponding pair of successive observations $\{y_i, y_{i+1}\}$. Using these discretised bridges, $P(\mathbf{X}|\Theta,\mathbf{Y})$ can be factorised as

$$P(\mathbf{X}|\mathbf{Y},\Theta)=\prod_{i=0}^{n-1}P(\mathbf{X}^{(i)}|y_i,y_{i+1},\Theta). \tag{4.8}$$

Since the SDE is assumed be time-homogeneous, each of the $(\mathbf{X}^{(i)}|y_i,y_{i+1},\Theta)$ are also identically distributed. Therefore, without loss of generality and for notational ease, only the variables corresponding to the second diffusion bridge will be considered here for the deliberation on how to approximate $P(\mathbf{X}|\Theta,\mathbf{Y})$. While $\{y_1,y_2\}$ are the observations corresponding to this bridge, $\{X_1,\cdots,X_{M-1}\}$ denote the corresponding

missing variables of the discretised diffusion bridge with $X_0 = y_1$ and $X_M = y_2$ (see Figure 4.2). Thus, the question 'how to evaluate $P(\mathbf{X}|\Theta, \mathbf{Y})$ in Equation 4.4?' will be answered by elaborating how possibly $P(X_1, \cdots, X_{M-1}|y_1, y_2, \Theta)$ can be evaluated.

Note that using the Markovian property of the diffusion process $P(X_1, \cdots, X_{M-1}|y_1, y_2, \Theta)$ can be written as

$$P(X_1, \cdots, X_{M-1}|y_1, y_2, \Theta) = \quad P(X_1|y_1, y_2, \Theta) \cdot P(X_2|X_1, y_2, \Theta) \cdots P(X_{M-1}|X_{M-2}, y_2, \Theta). \tag{4.9}$$

It is important to note that the above simplification is possible because the data have been assumed to be observed without error. Such simplification may not be possible if the data were observed with errors. In that case the dependence structure between the errors and their distributional assumptions would also have to be taken in to account. However, this case has not been considered in this thesis and it is assumed that the data have been observed without error.

Note that Equation (4.9) can be written as

$$P(X_1, \cdots, X_{M-1}|y_1, y_2, \Theta) = \prod_{j=1}^{M-1} P(X_j|X_{j-1}, X_M, \Theta) \tag{4.10}$$

where $X_0 = y_1$ and $X_M = y_2$. In order to implement GaMBA, $P(X_1, \cdots, X_{M-1}|y_1, y_2, \Theta)$ needs to be evaluated either exactly or approximately. However, note that this also involves sampling (simulating) X_j from $P(X_j|X_{j-1}, X_M, \Theta)$ for every j.

There might be different ways in which this could be achieved. It would be ideal to have a diffusion bridge construct, which will be easy to implement, be applicable to a wider class of diffusion processes, computationally not too expensive and be reasonably accurate as well. With these objectives in mind, the currently available options

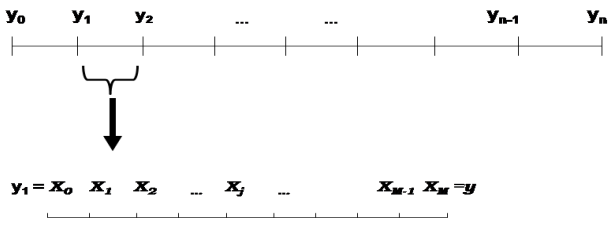


Figure 4.2: Second discretised diffusion bridge.

of simulating a diffusion bridge are reviewed below.

4.3.1 Constructing a bridge based on Euler's approximation

Since paths simulated using Euler's method converge in distribution to the true diffusion, as $M \to \infty$ ($\delta_\tau \to 0$), a natural option to approximate $P(X_1, \cdots, X_{M-1}|Y_1, Y_2, \Theta)$ is by using a density obtained by conditioning Euler's paths.

For every j, $P(X_j|X_{j-1}, X_M, \Theta)$ can be expressed as

$$\begin{aligned} P(X_j|X_{j-1}, X_M, \Theta) &= \frac{P(X_M|X_j, X_{j-1}, \Theta) \cdot P(X_j|X_{j-1}, \Theta)}{P(X_M|X_{j-1}, \Theta)} \\ &= \frac{P(X_M|X_j, \Theta) \cdot P(X_j|X_{j-1}, \Theta)}{P(X_M|X_{j-1}, \Theta)}. \end{aligned} \tag{4.11}$$

Now, X_M is observed, and while sampling X_j, X_{j-1} has already been sampled, and hence known, so $P(X_M|X_{j-1}, \Theta)$ is a constant, and therefore

$$P(X_j|X_{j-1}, X_M, \Theta) \propto P(X_M|X_j, \Theta) \cdot P(X_j|X_{j-1}, \Theta) \tag{4.12}$$

Using an Euler approximation for time lag δ_τ, $P(X_j|X_{j-1},\Theta)$ can be approximated as

$$P(X_j|X_{j-1},\Theta) \approx N_{X_j}(X_{j-1} + f(X_{j-1},\mu)\cdot\delta_\tau, \{g(X_{j-1},\sigma)\cdot\sqrt{\delta_\tau}\}^2), \tag{4.13}$$

where N_{X_j} denotes that random variable X_j has a Gaussian distribution with the mean and variance specified between the brackets.

Similarly, considering a time lag of $(M-j)\delta_\tau$, $P(X_M|X_j,\Theta)$ can be approximated as

$$P(X_M|X_j,\Theta) \approx N_{X_M}(X_j + f(X_j,\mu)\cdot(M-j)\delta_\tau, \{g(X_j,\sigma)\cdot\sqrt{(M-j)\delta_\tau}\}^2). \tag{4.14}$$

and therefore, $P(X_j|X_{j-1},X_M,\Theta)$ can be approximated using $P_{EB}(X_j|X_{j-1},X_M,\Theta)$, where

$$\begin{aligned} P_{EB}(X_j|X_{j-1},X_M,\Theta) \propto N_{X_M}(X_j + f(X_j,\mu)\cdot(M-j)\delta_\tau, \{g(X_j,\sigma)\cdot\sqrt{(M-j)\delta_\tau}\}^2) \\ \cdot N_{X_j}(X_{j-1} + f(X_{j-1},\mu)\cdot\delta_\tau, \{g(X_{j-1},\sigma)\cdot\sqrt{\delta_\tau}\}^2) \end{aligned} \tag{4.15}$$

where the subscript EB has been used to denote the bridge based on Euler's approximation. For Equation (4.15) to be used as a diffusion bridge construct, it needs to be simplified into a known closed form distribution from which X_j's can be sampled.

Limitations of EB : This approximation has two major drawbacks. Firstly, note that in the approximation of Equation (4.14), even as $M \to \infty$ (i.e $\delta_\tau \to 0$), $(M-j)\delta_\tau$- the time-gap between X_M and X_j remains a constant and therefore this approximation would not assymptotically converge to the true conditional distribution $P(X_M|X_j,\Theta)$. As a result the approximation $P_{EB}(X_j|X_{j-1},X_M,\Theta)$ would not converge to the desired conditional distribution $P(X_j|X_{j-1},X_M,\Theta)$ even as $M \to \infty$. Secondly, except for a linear SDE with a constant diffusion term (such as a 2 parameter OU process), it would

not be possible to simplify Equation (4.15) into a known closed form distribution from which X_j's could be easily sampled.

Therefore this approach does not seem promising and is not further pursued.

4.3.2 Modified Brownian Bridge (MBB)

The next diffusion bridge considered here is called the Modified Brownian Bridge, a construct proposed by Durham and Gallant (2002). MBB has been discussed and derived in Section 3.3.1.

Using the above notation, $P(X_j|X_{j-1}, X_M, \Theta)$ can be approximated using the MBB as

$$P_{MBB}(X_j|X_{j-1}, X_M, \Theta) \approx N_{X_j}\left(X_{j-1} + \left(\frac{X_M - X_{j-1}}{\tau_M - \tau_{j-1}}\right)\delta_\tau, \left\{g(X_{j-1}, \sigma)\sqrt{\left(\frac{M-j}{M-j+1}\right)\delta_\tau}\right\}^2\right) \tag{4.16}$$

where the subscript MBB has been used to denote the density corresponding to the MBB.

Just like the EB construct discussed in the previous section, the MBB construct does not have the desirable asymptotic properties; i.e the approximation $P_{MBB}(X_j|X_{j-1}, X_M, \Theta)$ would not converge to the desired conditional distribution $P(X_j|X_{j-1}, X_M, \Theta)$ even as $M \to \infty$. However, unlike the EB, it is widely applicable and has been successfully used as a proposal distribution for sampling diffusion bridges (Durham and Gallant (2002),Chib et al. (2006),and Golightly and Wilkinson (2007)). MBB can be applied to all one dimensional SDE's and extensions to multivariate SDEs is possible.

For the purposes of implementing GaMBA, MBB density can be used to approximate Equation (4.10) as

$$P(X_1, \cdots, X_{M-1}|y_1, y_2, \Theta) \approx \prod_{j=1}^{M-1} P_{MBB}(X_j|X_{j-1}, X_M, \Theta) \tag{4.17}$$

where P_{MBB} is the density of Equation (4.16). To ensure reliable inference (as described in Section 4.2.3), X_j can also be chosen to be X_j^* – the modal value of $P_{MBB}(X_j|X_{j-1}, X_M, \Theta)$.

It is important to note that because the mean of the MBB construct is just the linear interpolation between the two points, the posterior obtained using GaMBA may not be very accurate for non-linear processes. This is a limitation of GaMBA and will be illustrated in Section 4.7 using appropriate examples.

4.3.3 Other Approaches

Beskos et al. (2008) proposed an MCMC approach for generating paths of nonlinear diffusion bridges. This method is applicable to a wide class, but not all, of diffusions and compliments the current research work in the area of MCMC methods for high dimensional diffusions. Although this method produces exact diffusion bridges without even the discretisation error, being an MCMC based method, it is not suitable to be used within the GaMBA framework to sample diffusion bridges, as the aim is to develop a method which is much faster than the MCMC based methods.

More recently, Bladt and Sorensen (2010) proposed a method which relies on the time reversibility property of the ergodic diffusion processes and essentially consists

of simulating two paths of a diffusion process, one forward in time and another one moving backward in time. If the two trajectories intersect, then the combined path is a realisation of the bridge. The bridge generated using this method will have the distribution which will be close to the distribution of the true diffusion bridge. This is essentially a rejection sampling algorithm, which is very easy to implement and is available for all one dimensional ergodic diffusions. As the authors note however, the quality of approximation depends on the probability (denoted by π in the paper) that the bridge under consideration is hit by an independent diffusion process (the SDE under consideration).

Using this method within the GaMBA framework to simulate diffusion bridges does not appear to be attractive for two reasons. Firstly, determining π is not straightforward and its value depends on the unknown parameters which we wish to infer. Secondly, as described in the paper, even when π is close to one, the rejection rate could still be high making it a rather inefficient method to be used within the GaMBA framework.

4.4 Implementing GaMBA

This section summarises the discussion on diffusion bridge approximations provided in the last section and then provides and algorithm that can be used in practice to implement Bayesian inference on SDE models using GaMBA.

Even though, constructing a diffusion bridge approximation by conditioning Euler's path (EB) may seem attractive; as seen in Section 4.3.1, such an approximation would not have the desirable asymptotic properties. Also for most of the SDE models, it would not be possible to obtain a closed form solution to such an approximation,

severely limiting its usability. Therefore the EB construct is not a very attractive alternative.

As noted in Section 4.3.2 using MBB instead to simulate discretised diffusion bridges seems a good option. However, it is important to note that the MBB construct does not have the desirable asymptotic properties as well. Further, if the strategy of choosing the modal value of $P_{MBB}(X_j|X_{j-1}, X_M, \Theta)$ is adopted instead of sampling from it, then the resulting bridge turns out to be the linear interpolation between two observed points. Therefore this strategy has to be used with caution since it may not be yield accurate inference for non-linear SDE models.

There is a clear advantage of using the GaMBA framework, which is that it gets around the problem of dependency between the latent variables and the diffusion coefficients. This is because it only involves sampling from $P(\mathbf{X}|\Theta, \mathbf{Y})$ — that too using a fixed set of Θ values sampled using the grid $\mathcal{G}_{\Xi^*}$ on parameter space Ξ^*. Unlike MCMC based methods, GaMBA does not involve sampling from $P(\Theta|\mathbf{X}, \mathbf{Y})$, thus avoiding the reducibility problem due to the dependency between $\mathbf{X}$ and Θ.

There are other methods available to simulate diffusion bridges, some of which have been briefly reviewed in Section 4.3.3. However, these methods are computationally expensive compared to both the EB and the MBB approaches. Therefore, there does not seem much computational advantage in using these methods within the GaMBA framework.

GaMBA can thus be summarised into the following procedure.

GaMBA algorithm :

1. Identify Ξ^* (as described in Section 4.2.1.)
2. Define a discrete grid $\mathcal{G}_{\Xi^*}$ on Ξ^*.
3. For each point on the grid $\Theta_j \in \mathcal{G}_{\Xi^*}$,
   (a) Sample $\mathbf{X} \sim P_{MBB}(\mathbf{X}|\Theta_j, \mathbf{Y})$ or choose $\mathbf{X}$ to be $\mathbf{X}^* = \arg\max_X(P_{MBB}(\mathbf{X}|\Theta_j, \mathbf{Y}))$ as described in Section 4.3.2
   (b) For this $\mathbf{X}$ evaluate
   $$P_{GaMBA}(\Theta_j|\mathbf{Y}) \propto \frac{P_E(\mathbf{Y}, \mathbf{X}^*|\Theta_j) \cdot P(\Theta_j)}{P_{MBB}(\mathbf{X}^*|\Theta_j, \mathbf{Y})}$$
   .
4. Normalise to obtain $P(\Theta|\mathbf{Y})$ over $\mathcal{G}_{\Xi^*}$.

where:

- $P_E(\mathbf{Y}, \mathbf{X}|\Theta_j)$ is the Euler's density as in Equation (4.7);
- $P(\Theta_j)$ a suitable prior density;
- $P_{MBB}(\mathbf{X}|\mathbf{Y}, \Theta_j)$ is the MBB density as in Equation (4.17).

Note that, the posterior thus obtained will be a 'discretised' approximation to the true continuous posterior. If desired, a *continuous looking* posterior can be obtained by using standard kernal *smoothing* techniques which are easily available in any statistical package.

4.5 Link to Importance Sampling

This thesis is concerned with inference on parameters Θ of an SDE model given time-discrete observations $\mathbf{Y} = \{y_0, y_1, \cdots, y_n\}$. If the transition densities were known, one would find the posterior simply as

$$P(\Theta|\mathbf{Y}) = P(\mathbf{Y}|\Theta) \cdot P(\Theta)$$

where $\log(P(\mathbf{Y}|\Theta))$ is the log likelihood $l_n(\Theta)$ of Equation 3.5. But transition densities are unavailable for all but a few standard SDE models. As described earlier, many inference methods get around this problem by introducing latent variables $\mathbf{X}$ so that the transition densities could be approximated by densities corresponding some numerical approximation(such as Euler's). Once these latent variables have been introduced several different approaches could lead to the likelihood based inference on Θ.

The approach taken by simulated likelihood methods, as described in Section 3.3.1 is to obtain the likelihood by integrating out the latent variables from the complete likelihood, i.e

$$P(\mathbf{Y}|\Theta) = \int P(\mathbf{Y}, \mathbf{X}\,|\Theta)\; d\mathbf{X}, \tag{4.18}$$

where Equation 4.18 is usually solved using Monte-Carlo integration.

The approach taken by MCMC based methods, as described in section 3.6 is to interpret the problem as that of a hidden Markov model and employ Metropolis within Gibbs type of methods which alternately sample $P(\mathbf{X}|\mathbf{Y}, \Theta)$ and $P(\Theta|\mathbf{Y}, \mathbf{X})$ respectively. This usually involves some form of reparameterisation to avoid the dependency problem.

On the other hand, the approach taken by GaMBA is to approximate the likelihood as

$$P(\mathbf{Y}|\Theta) \approx \frac{P_E(\mathbf{Y}, \mathbf{X}|\Theta)}{P_{MBB}(\mathbf{X}|\mathbf{Y}, \Theta)} \tag{4.19}$$

and then to evaluate Equation 4.19 at $\mathbf{X}^* = \arg\max_X P_{MBB}(\mathbf{X}|\mathbf{Y}, \Theta)$ for a given value of Θ. This procedure is then repeated for large number of Θ's deterministically chosen from the support Ξ^*.

However, it is possible to show that there is a link between the approach taken by GaMBA and the one taken by simulated likelihood method. In fact, it can be shown that GaMBA can be easily extended so as to be interpreted as a grid based implementation of simulated likelihood method using importance sampling. This can be shown as follows.

4.5.1 GaMBA-I (GaMBA with Importance Sampling)

Consider the GaMBA algorithm described in Section 4.4. In step 3 of the algorithm an approximation to $P(\mathbf{Y}|\Theta)$ is evaluated using Equation 4.19 for every predetermined value of Θ, by sampling (or choosing) $\mathbf{X}$ ***once*** from $P_{MBB}(\mathbf{X}|\mathbf{Y}, \Theta)$. Instead, a better approximation, could possibly be obtained by sampling multiple (say K) values of $\mathbf{X}$ for every predetermined value of Θ, and then evaluating

$$P(\mathbf{Y}|\Theta) \approx \frac{1}{K} \sum_{k=1}^{K} \frac{P_E(\mathbf{Y}, \mathbf{X}_k|\Theta)}{P_{MBB}(\mathbf{X}_k|\mathbf{Y}, \Theta)}. \tag{4.20}$$

GaMBA algorithm can be extended using this modification. This extended algorithm is as follows:

GaMBA-I algorithm :

1. Identify Ξ^* (as described in Section 4.2.1.)

2. Define a discrete grid $\mathcal{G}_{\Xi^*}$ on Ξ^*.

3. For each point on the grid $\Theta_j \in \mathcal{G}_{\Xi^*}$

   (a) for $k = 1, \cdots, K$,

   Sample $\mathbf{X}_k \sim P_{MBB}(\mathbf{X}_k|\Theta_j, \mathbf{Y})$ as described in Section 4.3.2

   (b) Evaluate

$$P_{GaMBA-I}(\Theta_j|\mathbf{Y}) \propto \frac{1}{K}\sum_{k=1}^{K}\frac{P_E(\mathbf{Y}, \mathbf{X}_k|\Theta_j)\cdot P(\Theta_j)}{P_{MBB}(\mathbf{X}_k|\Theta_j, \mathbf{Y})}$$

   .

4. Normalise to obtain $P(\Theta|\mathbf{Y})$ over $\mathcal{G}_{\Xi^*}$.

where:

- $P_E(\mathbf{Y}, \mathbf{X}|\Theta_j)$ is the Euler's density as in Equation (4.7);
- $P(\Theta_j)$ a suitable prior density;
- $P_{MBB}(\mathbf{X}|\mathbf{Y}, \Theta_j)$ is the MBB density as in Equation (4.17).

Again, note that the posterior thus obtained will be a 'discretised' approximation to the true continuous posterior. If desired, a *continuous looking* posterior can be obtained by using standard kernal *smoothing* techniques.

Note that the approximation in Equation (4.20) is same as the one used in Equation (3.11) where MBB density of Equation (3.20) has been used as the proposal density $q(.)$. Thus, GaMBA-I (*GaMBA - Importance sampling*) can be seen as the novel implementation of the simulated likelihood method proposed by Durham and Gallant (2002) used in the Bayesian context. The novelty being that the parameter space is explored

using a predetermined set of points chosen using a grid over Ξ^*. On the other hand the approach taken so far (Pedersen (1995b),Santa-Clara (1995),Brandt and Santa-Clara (2002) and Durham and Gallant (2002)) has been to explore the parameter space using numerical optimisation methods such as a Newton-Raphson method, for example.

The advantage of interpreting GaMBA-I as a novel implementation of the simulated likelihood method is that statements regarding the asymptotic behaviour of GaMBA-I posterior can now be made using the results already known.

4.5.2 Convergence of GaMBA-I

By establishing link with the Importance sampling, GaMBA-I creates a possibility for exploring its consistency properties.

Some consistency properties are already known. Recall the notations $p^M(\Theta)$, $p^{M,K}(\Theta)$, $l_n^{(M)}(\Theta)$ and $l_n(\Theta)$ from Section 3.3.1. Note that $P_{GaMBA-I}(\Theta_j|\mathbf{Y})$ can be expressed as

$$P_{GaMBA-I}(\Theta_j|\mathbf{Y}) = \prod_{i=0}^{n-1} p^{M,K}(t_{i-1}, y_{i-1}, t_i, y_i, \Theta_j) \cdot P(\Theta_j) \tag{4.21}$$

Stramer and Yan (2007) have shown that for SDE's with unit diffusion, under the assumptions **A** 1, **A** 2, **A** 7 and **A** 8 that

1. $p^{M,K}$ converges weakly to a non-centered Gaussian random variable implying that the variability associated with the Importance sampling estimate is uniformly bounded in M, and
2. the total error in this estimation is of the order $O(1/M) + O(1/\sqrt{K})$.

Based on the results above, Stramer and Yan (2007) suggest choosing $K = M^2$ as an asymptotically optimal choice. Since most one dimensional SDEs with non-constant

diffusion coefficients could be transformed using the Lamperti transform (Section 3.6.2) into SDEs with unit diffusion coefficient, the result above is applicable to such SDEs as well. However, they point out that these asymptotic results depend on how closely the data are observed. They also illustrate that in practice, these results may take effect only for very high values of M.

For SDEs which can not be transformed into ones with unit diffusion coefficients, Stramer and Yan (2007) argue that it would be difficult to prove a general result such as the one above and that such results would need to be proved on a case to case basis. Though, this thesis does not aim to prove this convergence for any particular model, it is shown empirically in the next section that the marginal posteriors obtained using GaMBA-I do appear to become more accurate for the parameters of the un-transformed CIR process. These results are consistent with the results obtained by Stramer and Yan (2007) on the un-transformed CIR process.

Further, Section 4.7 empirically shows the examples where the 95% probability interval obtained using P_E and P_{MBB} do not overlap - indicating that GaMBA and GaMBA-I would not yield efficient and accurate inference for certain SDEs when the data are too sparsely observed. These examples further illustrate the point made by Stramer and Yan (2007).

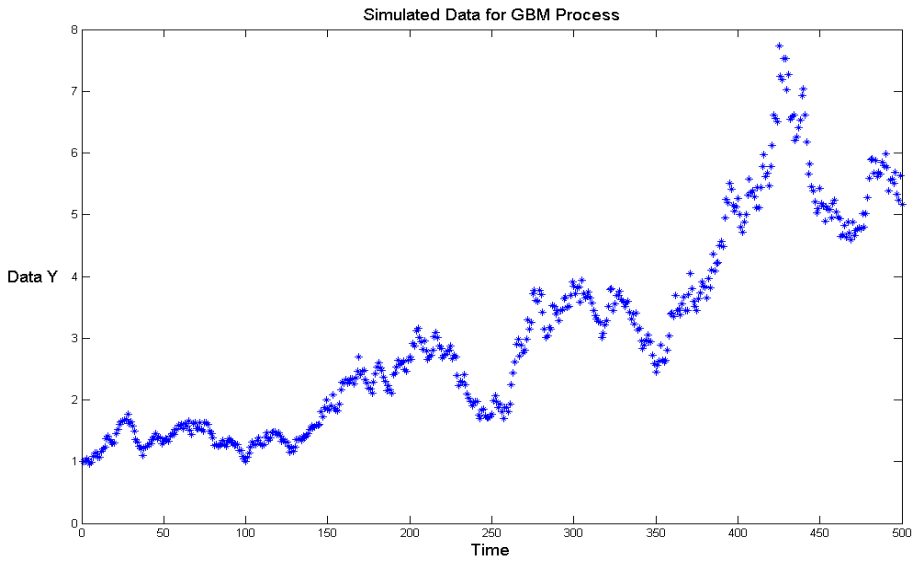


Figure 4.3: Simulated data for GBM process.

4.6 Examples

4.6.1 Geometric Brownian Motion (GBM) Process

Geometric Brownian Motion process is the solution to the stochastic differential equation,

$$dX_t = \theta_1 X_t \, dt + \theta_2 X_t \, dW_t \tag{4.22}$$

with $\theta_1 \in \Re$, and $\theta_2 \in \Re^+$. This process is also known as the Black-Scholes-Merton model after its introduction in the financial context to model asset prices where θ_1 is interpreted as the constant interest rate while θ_2 as the volatility of risky activity. For this process, the MBB density can be used to approximate $P(\mathbf{X}|\Theta, \mathbf{Y})$ in step 3 of GaMBA.

A dataset of 500 observations was simulated using parameter values $\theta_1 = 0.005$, $\theta_2 = 0.05$ and starting value of $Y_0 = 1$. This dataset was simulated using Euler's method with $M = 10$ and $\delta_{tau} = 0.1$ (i.e $\Delta_t = 1$). Figure 4.3 shows the simulated

data. The parameter space $\Xi^* = \Xi_1^* \times \Xi_2^*$ was chosen based on prior knowledge as $\Xi^* = [-0.05, 0.08] \times [0.01, 0.1]$, and a grid $\mathcal{G}_{\Xi^*}$ was considered with $\Delta\Xi_1 = 0.002$ and $\Delta\Xi_2 = 0.002$. GaMBA was implemented on the 3036 points thus sampled from Ξ^* and marginal posteriors distributions were obtained.

For this process, the true transition density $P(Y_{t+\Delta_t}|Y_t, \Theta)$ is log-normal with log-mean and log-variance given by,

$$\begin{aligned} \mu_{GBM} &= log(Y_t) + \left(\theta_1 - \frac{1}{2}\theta_2^2\right) \cdot \Delta_t \\ \sigma^2_{GBM} &= \theta_2^2 \cdot \Delta_t \end{aligned}$$

Therefore, it is possible to evaluate the true posterior inference on $\mathcal{G}_{\Xi^*}$. This true posterior and the GaMBA posterior computed for $M = 5$ were plotted together along with the true parameter values used in simulating the data. Figure 4.4 shows the results where the vertical lines denote the true parameter values. Posteriors obtained using GaMBA have smaller variance than the true posteriors, however their modes closely match with the true parameter value. In this example, Uniform priors were used for both GaMBA and the exact method.

Thus, it can be said that in this example GaMBA has correctly captured the mean but not the variance of the marginal posteriors.

4.6.2 Ornstein-Uhlenbeck (O-U) Process

This process was introduced in Section 2.2.2. When the first parameter $\theta_1 = 0$, the process is *mean reverting* around 0, the rate of mean-reversion being controlled by the

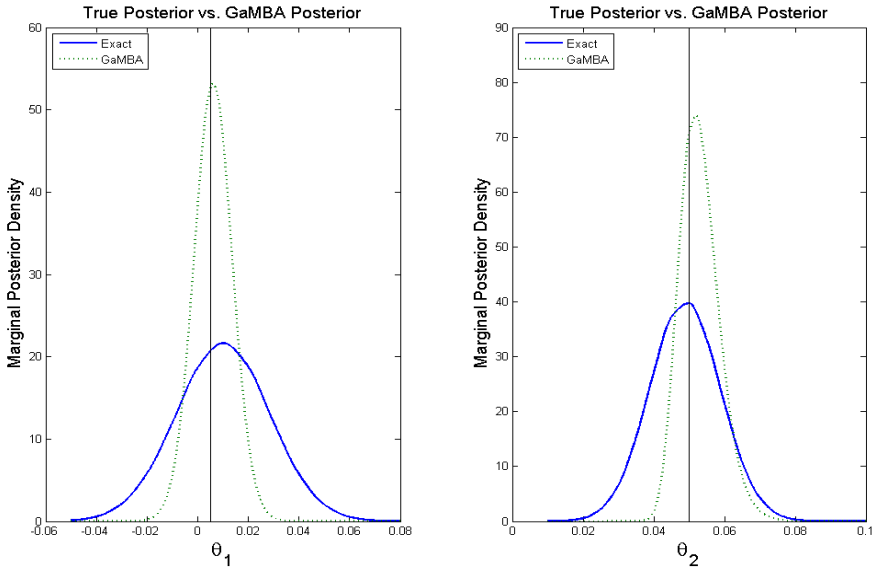


Figure 4.4: GBM: Marginal True posterior (line) vs. posterior using GaMBA $M = 5$ (dotted line). Vertical lines denote the true values $\theta_1 = 0.005$ and $\theta_2 = 0.05$.

parameter θ_2. A smaller absolute value of θ_2 means that the process will revert back to 0 less often, while a large absolute value of θ_2 implies that the process will revert back to 0 more often.

This two parameter Ornstein-Uhlenbeck process is the solution to the stochastic differential equation,

$$dX_t = -\theta_2 X_t\, dt + \theta_3\, dW_t \quad X_0 = x_0 \tag{4.23}$$

with $\theta_2 \in \Re$, and $\theta_3 \in \Re^+$.

As described in Section 2.2.2, the transition density for this process is Gaussian and hence true posterior distribution can be determined. The exact solution of the above SDE is given by

$$X_t = x_0 e^{-\theta_2 t} + \theta_3 \int_0^t e^{-\theta_2 (t-u)}\, du. \tag{4.24}$$

A dataset of 500 observations was simulated using parameter values $\theta_2 = 0.1$, $\theta_3 = 0.25$ and starting value of $Y_0 = 0$. This dataset was simulated using Euler's method with $M = 10$ and $\delta_{tau} = 0.1$ (i.e $\Delta_t = 1$). Figure 4.5 shows the simulated data. The parameter space $\Xi^* = \Xi_2^* \times \Xi_3^*$ was chosen based on prior knowledge as $\Xi^* = [-0.1, 0.3] \times [0.1, 0.4]$, and a grid $\mathcal{G}_{\Xi^*}$ was considered with $\Delta\Xi_2 = 0.01$ and $\Delta\Xi_3 = 0.01$. GaMBA was implemented on the 1271 points thus sampled from Ξ^* and marginal posteriors distributions were obtained.

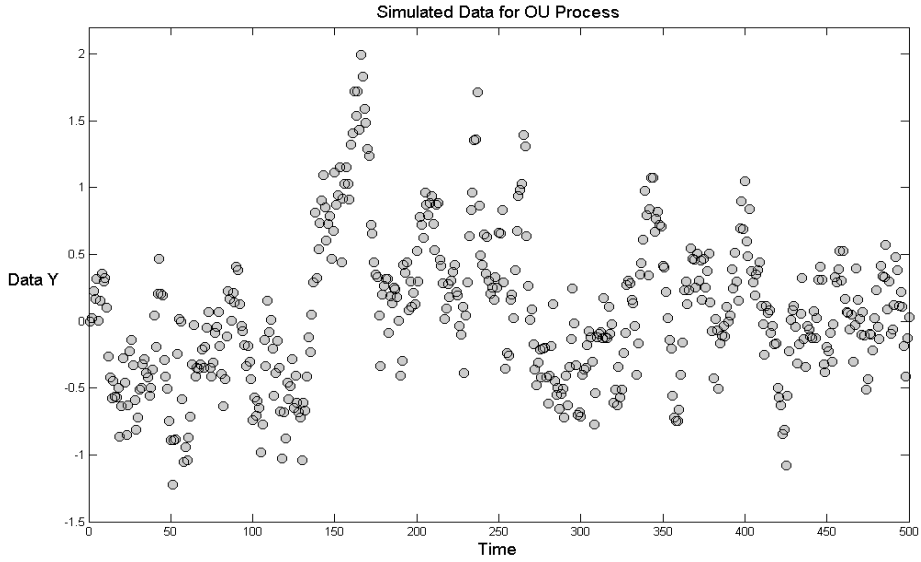


Figure 4.5: Simulated data for O-U process.

In addition, GaMBA-I was also implemented on the same $\mathcal{G}_{\Xi^*}$ as above for various M and K. Table 4.1 lists the mean squared error (MSE) obtained for each of the parameters using different methods for inference. Asymptotic properties of the GaMBA-I posterior can be seen. Figure 4.6 plots the marginal posteriors obtained using GaMBA-I (for [$M = 5$,$K = 10$] and [$M = 10$,$K = 20$]) along with those obtained

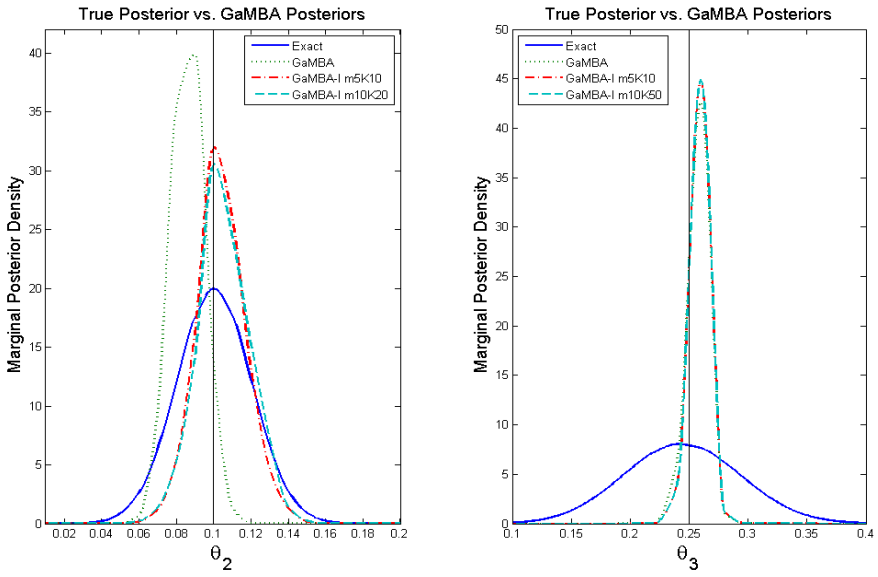


Figure 4.6: O-U: Marginal True posterior (line) vs. posterior using GaMBA $M = 5$ (dotted line). Vertical lines denote the true values $\theta_2 = 0.1$ and $\theta_3 = 0.25$.

used GaMBA and the true likelihood. It shows that for θ_2, GaMBA-I works much better than GaMBA, and also that as M and K increases posteriors obtained using GaMBA-I become more accurate for θ_2, but this effect is not seen for θ_3.

Thus, it can be said that though the marginal posteriors obtained using GaMBA-I contain the true parameter values for both θ_2 and θ_3, the spread of these distributions does not match very closely with the true marginal posterior distributions, specially for θ_3.

4.6.3 Cox-Ingersoll-Ross (CIR) Process

This process was introduced in Section 3.6.3. This process is also non-linear and it is important to see how accurate the posteriors obtained using GaMBA and GaMBA-I

Table 4.1: O-U: MSE obtained for θ_2 & θ_3 for the posteriors obtained using different methods

$Method$	M	K	MSE θ_2	MSE θ_3
$TrueLikelihood$	-	-	0.004851	0.000356
$GaMBA$	5	-	0.006775	0.000158
$GaMBA-I$	5	10	0.001102	8.88×10^{-5}
$GaMBA-I$	5	20	0.000987	7.93×10^{-5}
$GaMBA-I$	10	10	0.001052	8.01×10^{-5}
$GaMBA-I$	10	20	0.000826	6.86×10^{-5}

are for this process.

A dataset of 100 observations was simulated using parameter values $\theta_1 = 1$, $\theta_2 = 0.5$, $\theta_3 = 0.2$ and starting value of $y_0 = 2.5$. This dataset was simulated using Euler's method with $M = 10$ and $\delta_{tau} = 0.1$ (i.e $\Delta_t = 1$). Figure 4.7 shows the simulated data. The parameter space $\Xi^* = \Xi_1^* \times \Xi_2^* \times \Xi_3^*$ was chosen based on prior knowledge as $\Xi^* = [0.1, 1.9] \times [0.1, 0.8] \times [0.1, 0.3]$, and a grid $\mathcal{G}_{\Xi^*}$ was considered with $\Delta\Xi_1 = 0.1$, $\Delta\Xi_2 = 0.05$ and $\Delta\Xi_3 = 0.02$. GaMBA was implemented on the 3135 points thus sampled from Ξ^* and marginal posteriors distributions were obtained.

Table 4.2: CIR: MSE obtained for θ_1, θ_2 & θ_3 for the posteriors obtained using different methods

$Method$	M	K	MSE θ_1	MSE θ_2	MSE θ_3
$GaMBA$	5	-	0.11623	0.034306	0.000576
$GaMBA-I$	5	5	0.05767	0.01799	0.000575
$GaMBA-I$	5	20	0.04538	0.015	0.000405
$GaMBA-I$	10	5	0.0455	0.01394	0.000401
$GaMBA-I$	10	20	0.0421	0.01312	0.000378

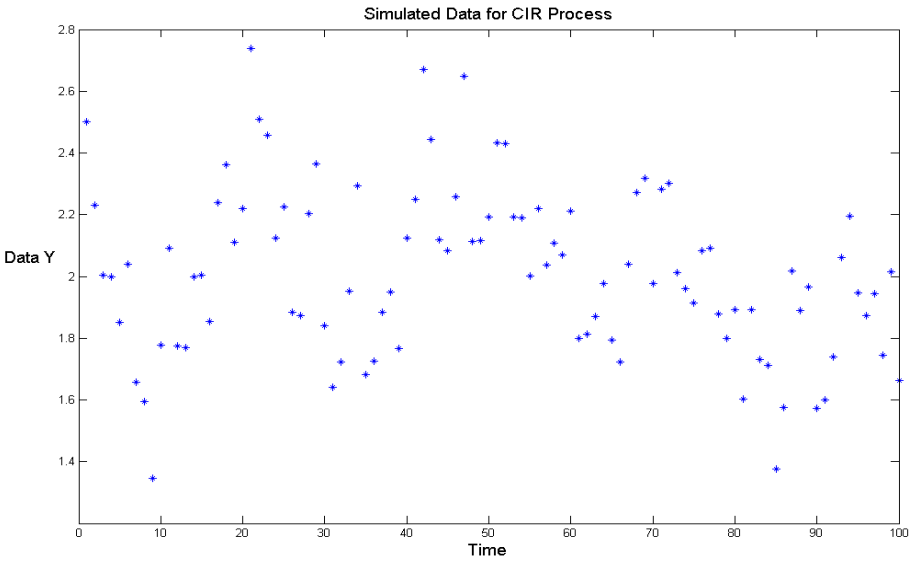


Figure 4.7: Simulated data for CIR process.

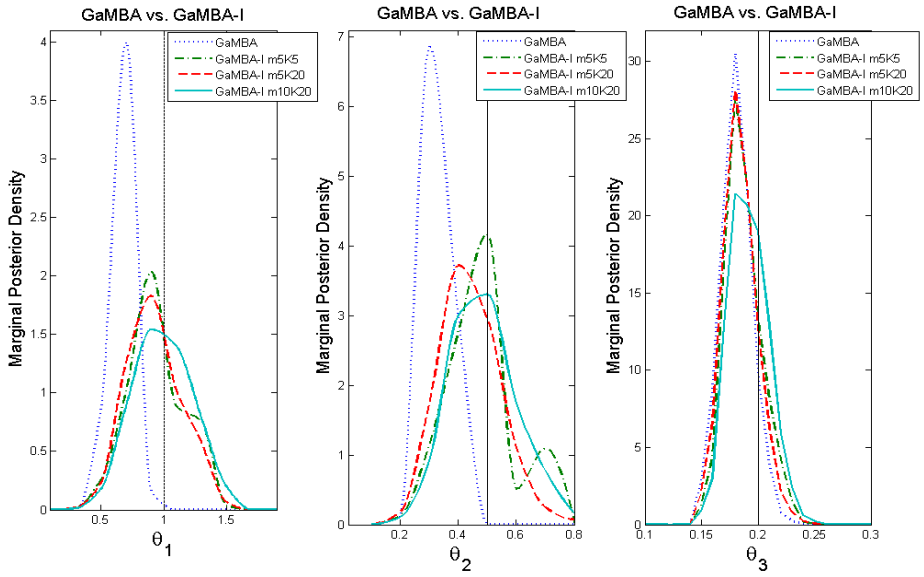


Figure 4.8: CIR: Marginal posterior using GaMBA for $M = 5$ vs. posteriors obtained using GaMBA-I for $[M = 5, K = 5]$, $[M = 5, K = 20]$ and $[M = 10, K = 20]$ respectively. Vertical lines denote the true values $\theta_1 = 1$, $\theta_2 = 0.5$ and $\theta_3 = 0.2$.

Table 4.2 lists the mean squared error (MSE) obtained for each of the parameters

using different methods for inference. Asymptotic properties of the GaMBA-I posterior can be seen. Posteriors obtained using GaMBA are not accurate for the CIR process. GaMBA-I does provide better approximation and that these approximations become more accurate as M and K increase. Figure 4.8 plots some of these posteriors.

Example: Euro-Dollar interest rate data

Recall the Euro-Dollar interest rate data analysed using the MCMC method in Section 3.6.3. This data is now analysed using GaMBA-I with $M = 5$ and $K = 10$. The parameter space $\Xi^* = \Xi_1^* \times \Xi_2^* \times \Xi_3^*$ was chosen based on prior knowledge as $\Xi^* = [0.00001, 0.00401] \times [0.00002, 0.02] \times [0.001, 0.0181]$, and a grid $\mathcal{G}_{\Xi^*}$ was considered with $\Delta\Xi_1 = 0.0003$, $\Delta\Xi_2 = 0.003$ and $\Delta\Xi_3 = 0.001$. GaMBA was implemented on the 1764 points thus sampled from Ξ^* and marginal posteriors distributions were obtained.

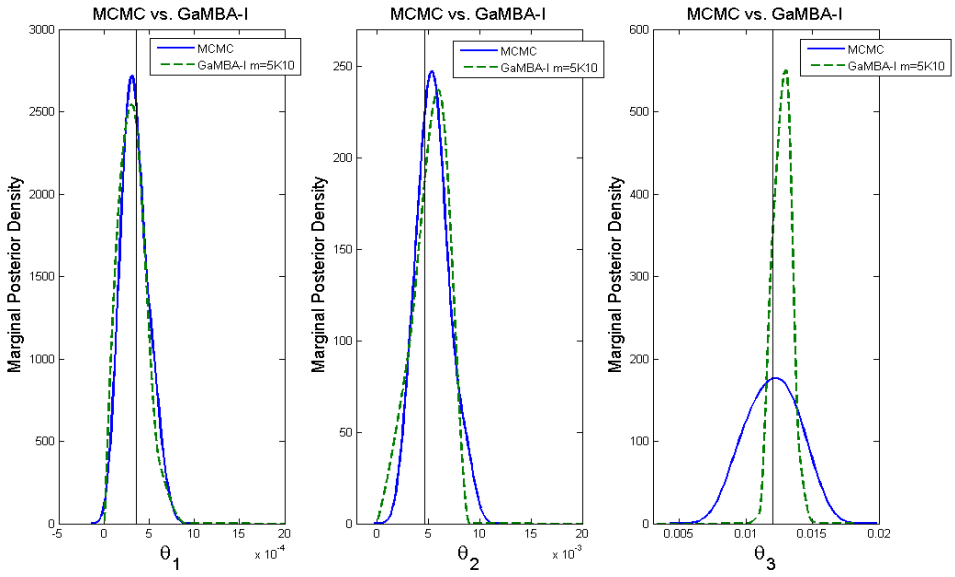


Figure 4.9: Euro-Dollar data: Marginal posteriors using MCMC vs. posteriors obtained using GaMBA-I for $[M = 5, K = 10]$. Vertical lines denote the true values $\theta_1 = 00036$, $\theta_2 = 0.0047$ and $\theta_3 = 0.012$.

Figure 4.9 shows the GaMBA-I posterior thus obtained plotted over the MCMC posterior of Figure 3.5 obtained in Section 3.6.3. It can be seen that the marginal posteriors using these two methods for θ_1 and θ_2 are quite close. For θ_3, posterior obtained using GaMBA-I has much smaller variance than the one obtained using MCMC, but it still covers the true value of the parameters. GaMBA-I takes only 6 minutes to implement in this case, and as described in the Section 4.8, this time can be further reduced multi-fold by using multiple parallel processing computing units. Further as seen in the earlier example, if desired, accuracy of GaMBA-I can be improved by increasing M and K.

4.7 Where GaMBA and GaMBA-I do not work

As described earlier, the mean of the MBB density is the linear interpolation between the two observed points. As a consequence GaMBA and GaMBA-I would yield efficient and accurate inference when the data are observed close enough so that the path taken by the process between any two consecutive observations can be considered to be approximately linear.

Let Δ_t denote the time difference between the consecutive data points. For all the examples considered in Section 4.6, the data was observed at every unit time interval, i.e $\Delta_t = 1$. In this section, it is illustrated using simulated examples that in the extreme cases where $\Delta_t >> 1$, the 95% probability interval for $P_E(y_t, y_{t+\Delta_t}, \mathbf{X}|\Theta)$ and $P_{MBB}(\mathbf{X}|y_t, y_{t+\Delta_t}, \Theta)$, where y_t and $y_{t+\Delta_t}$ are the observed data, do not overlap for processes with highly non-linear paths. These examples illustrate that for processes with highly non-linear paths and where the data is sparsely observed, GaMBA and GaMBA-I would not yield an efficient and accurate inference.

4.7.1 GBM process

The GBM process models exponential growth the rate of which is determined by parameter θ_1. A smaller value for θ_1 would indicate a slower rate of growth.

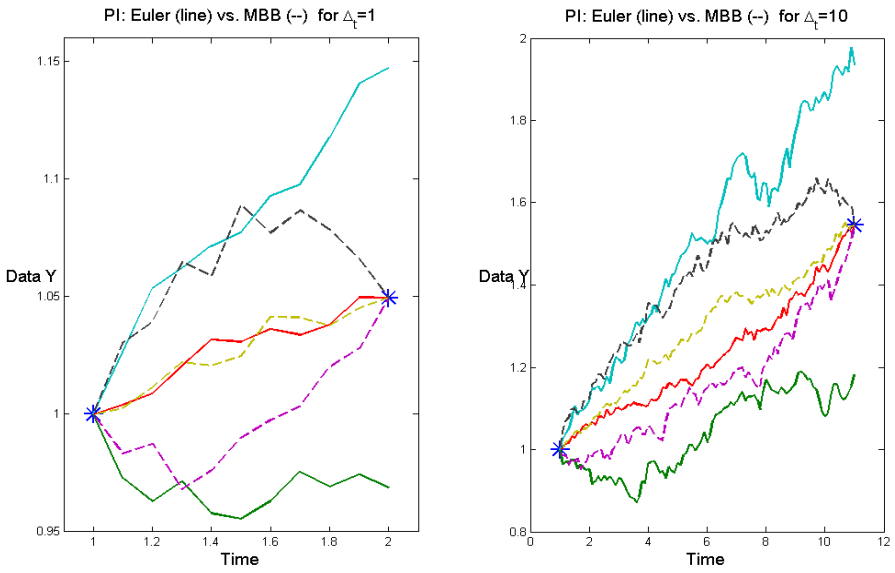


Figure 4.10: Median and 95% probability intervals based on a set of simulations, for GBM with $\theta_1 = 0.05$ and $\theta_2 = 0.05$ using P_{MBB} (dashed lines) vs. P_E (lines) for $\Delta_t = 1$ (left) and $\Delta_t = 10$ (right). Observed data are denoted by the asterisk.

Consider a GBM process with a slower rate of growth, $\theta_1 = 0.05$. Figure 4.10 shows that even when $\Delta_t = 10$, the 95% probability interval for $P_E(y_t, y_{t+\Delta_t}, \mathbf{X}|\Theta)$ and $P_{MBB}(\mathbf{X}|y_t, y_{t+\Delta_t}, \Theta)$ do overlap, as they do when $\Delta_t = 1$. However, for a GBM process with a faster rate of growth $\theta_1 = 0.25$, Figure 4.11 shows that if $\Delta_t = 10$, then the 95% probability interval for $P_E(y_t, y_{t+\Delta_t}, \mathbf{X}|\Theta)$ and $P_{MBB}(\mathbf{X}|y_t, y_{t+\Delta_t}, \Theta)$ do not overlap. Note that, this is not the case when $\Delta_t = 1$.

4.7.2 OU process

As described earlier, OU process is a *mean-reverting* process with the rate of reversion controlled by the parameter θ_2. A larger value of θ_2 implies that the process will revert back more often.

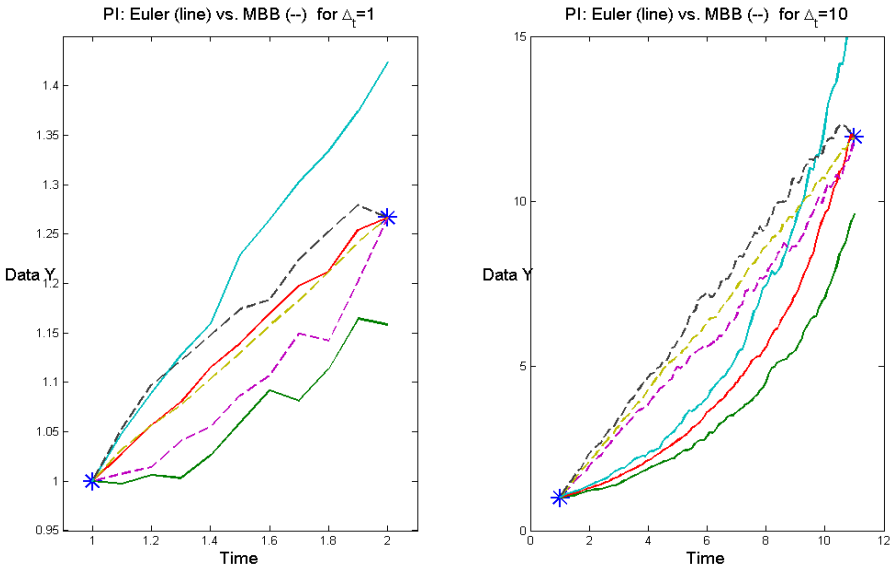


Figure 4.11: Median and 95% probability intervals based on a set of simulations, for GBM with $\theta_1 = 0.25$ and $\theta_2 = 0.05$ using P_{MBB} (dashed lines) vs. P_E (lines) for $\Delta_t = 1$ (left) and $\Delta_t = 10$ (right). Observed data are denoted by the asterisk.

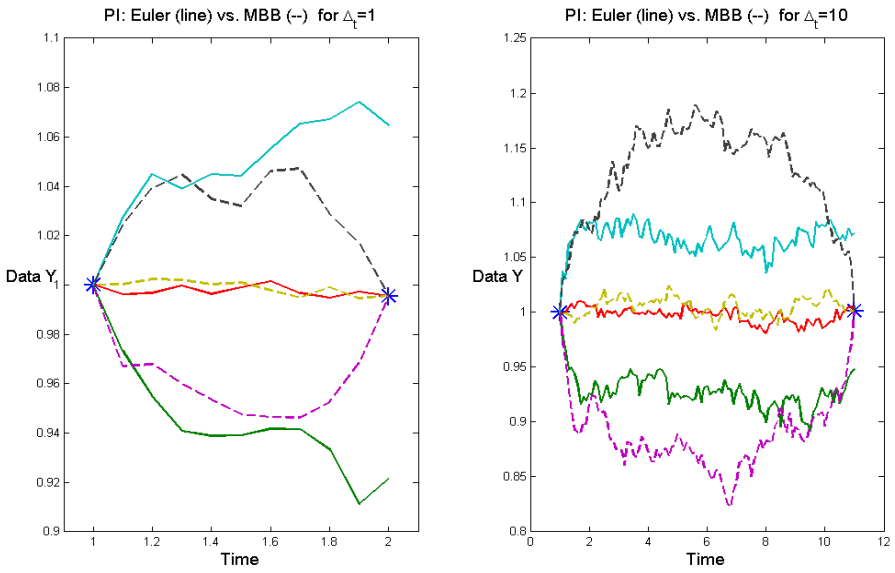


Figure 4.12: Median and 95% probability intervals based on a set of simulations, for OU with $\theta_1 = 1$, $\theta_2 = 1$ and $\theta_3 = 0.05$ using P_{MBB} (dashed lines) vs. P_E (lines) for $\Delta_t = 1$ (left) and $\Delta_t = 10$ (right). Observed data are denoted by the asterisk.

Consider an OU process with a rate of reversion, $\theta_2 = 1$. Figure 4.12 shows that even when $\Delta_t = 10$, the 95% probability interval for $P_E(y_t, y_{t+\Delta_t}, \mathbf{X}|\Theta)$ and $P_{MBB}(\mathbf{X}|y_t, y_{t+\Delta_t}, \Theta)$ do overlap, as they do when $\Delta_t = 1$. However, for an OU process with a much slower rate of reversion $\theta_2 = 0.1$, Figure 4.13 shows that if $\Delta_t = 10$, then the 95% probability interval for $P_E(y_t, y_{t+\Delta_t}, \mathbf{X}|\Theta)$ and $P_{MBB}(\mathbf{X}|y_t, y_{t+\Delta_t}, \Theta)$ do not overlap. Note that, this is not the case when $\Delta_t = 1$.

4.7.3 CIR process

The CIR process also has a *mean-reverting* property with the rate of reversion controlled by the parameter θ_2. A larger value of θ_2 implies that the process will revert back more often.

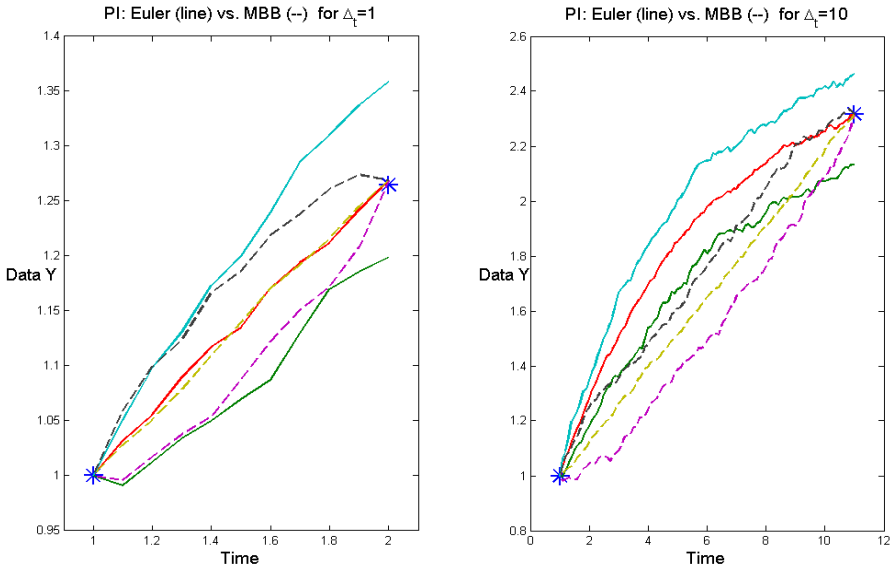


Figure 4.13: Median and 95% probability intervals based on a set of simulations, for OU with $\theta_1 = 1$, $\theta_2 = 0.1$ and $\theta_3 = 0.05$ using P_{MBB} (dashed lines) vs. P_E (lines) for $\Delta_t = 1$ (left) and $\Delta_t = 10$ (right). Observed data are denoted by the asterisk.

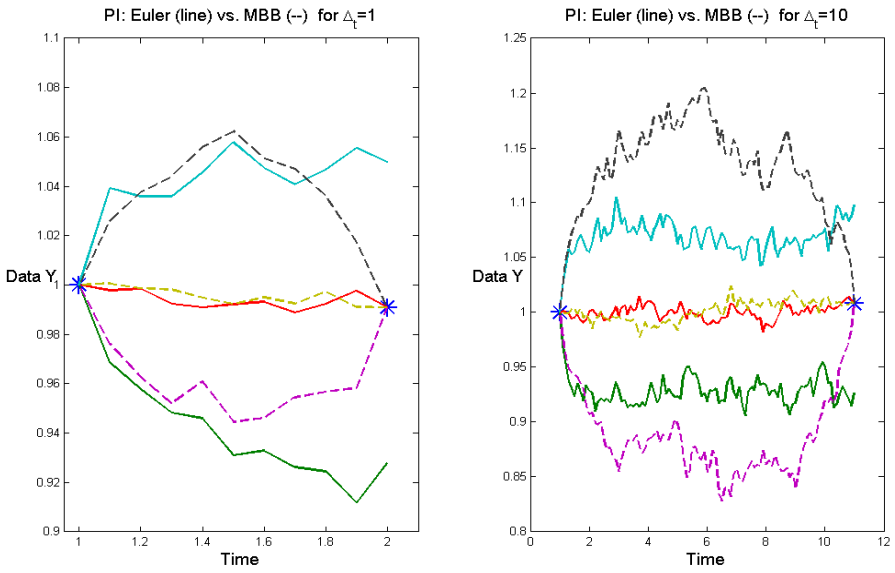


Figure 4.14: Median and 95% probability intervals based on a set of simulations, for CIR with $\theta_1 = 1$, $\theta_2 = 1$ and $\theta_3 = 0.05$ using P_{MBB} (dashed lines) vs. P_E (lines) for $\Delta_t = 1$ (left) and $\Delta_t = 10$ (right).Observed data are denoted by the asterisk.

Consider a CIR process with a rate of reversion, $\theta_2 = 1$. Figure 4.14 shows that even when $\Delta_t = 10$, the 95% probability interval for $P_E(y_t, y_{t+\Delta_t}, \mathbf{X}|\Theta)$ and $P_{MBB}(\mathbf{X}|y_t, y_{t+\Delta_t}, \Theta)$ do overlap, as they do when $\Delta_t = 1$. However, for a CIR process with a much slower rate of reversion $\theta_2 = 0.25$, Figure 4.15 shows that if $\Delta_t = 10$, then the 95% probability interval for $P_E(y_t, y_{t+\Delta_t}, \mathbf{X}|\Theta)$ and $P_{MBB}(\mathbf{X}|y_t, y_{t+\Delta_t}, \Theta)$ do not overlap. Note that, this is not the case when $\Delta_t = 1$.

Remark: Non-linearity of the paths is also governed by the diffusion coefficient. The illustrations shown above have been selected as *extreme* cases with very small diffusion coefficients (0.05). A larger diffusion coefficient will yield wider 95% probability intervals for both P_E and P_{MBB} and thus will likely increase the efficiency of GaMBA-I on sparse data.

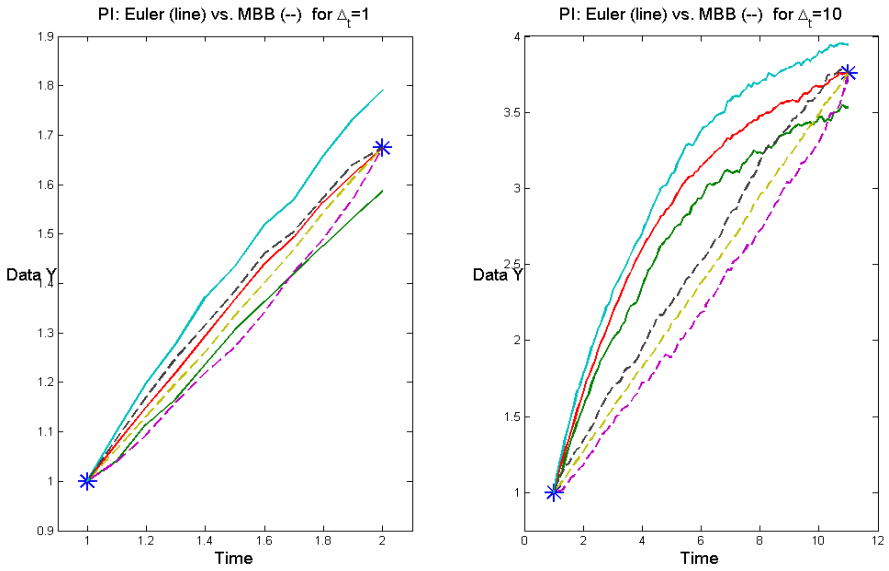


Figure 4.15: Median and 95% probability intervals based on a set of simulations, for CIR with $\theta_1 = 1$, $\theta_2 = 0.25$ and $\theta_3 = 0.05$ using P_{MBB} (dashed lines) vs. P_E (lines) for $\Delta_t = 1$ (left) and $\Delta_t = 10$ (right). Observed data are denoted by the asterisk.

4.8 Discussion

Two algorithms have been introduced in this chapter to obtain computationally efficient Bayesian inference on SDE models. Various aspects of these algorithms are discussed below.

4.8.1 Which is more appropriate: GaMBA or GaMBA-I?

Both these methods use the MBB density to sample the latent variables $\mathbf{X}$. However while GaMBA is a deterministic approach, GaMBA-I is stochastic. There are also important differences between them regarding the computatinal efficienty, consistency

and applicability of these two methods. Some guidelines are provided here to help the practitioner decide which method is more appropriate.

The advantage that GaMBA has over GaMBA-I is that, in principle, GaMBA is K times faster than GaMBA-I, because it simply chooses the modal value of the MBB transition density at each time point unlike GaMBA-I which draws K samples of $\mathbf{X}$ instead.

GaMBA-I has two clear advantages over GaMBA: firstly, unlike GaMBA, GaMBA-I provides the possibility to study the consistency properties of its posterior. Some consistency results are already available for GaMBA-I as discussed in Section 4.5.2 and illustrated in Section 4.6. Secondly, from the illustrations in Section 4.7 it is clear that if the data are sparsely observed then GaMBA-I will provide better inference than GaMBA and is therefore more widely applicable.

Therefore to summarise, it would be advisable to use GaMBA only if data are fairly closely observed so that $\Delta_t \leq 1$, and when the computational resources are scarce. In all other circumstances GaMBA-I would be more preferable than GaMBA. Illustrations in Section 4.7 can be used to decide the if the data are too sparsely observed for the concerned process for GaMBA-I to be no longer efficient.

4.8.2 Practical Considerations

Because of its asymptotic properties, and wider applicability GaMBA-I is the preferred algorithm and is therefore the focus of the following discussion. However, unless otherwise stated, the same discussion is applicable to GaMBA as well.

Sparsely observed data: Stramer and Yan (2007) have illustrated that their convergence results take effect only for a large value of M even for $\Delta_t = 1$. All the datasets analysed in Section 4.6 were simulated with $\Delta_t = 1$. Further in Section 4.7 it has been effectively illustrated why GaMBA-I could become very inefficient (depending on the non-linearity of the process) for sparsely observed data where $\Delta_t = 10$. Also, as stated above, it would be advisable to use GaMBA only if data are fairly closely observed so that $\Delta_t \leq 1$, and when the computational resources are scarce.

Choosing M : Section 4.6 has illustrated that reasonably accurate results could be obtained for values of M as small as 5 or 10. However, in many inference problems, the accuracy achieved in Section 4.6 may not be enough and higher M will have to be used to get better results.

Choosing K : Stramer and Yan (2007) prove that $K = M^2$ is the computationally optimal value for SDEs with unit diffusion coefficient. As can be seen from the examples considered in Section 4.6, a smaller value of K (between 5 and 20) is often enough to provide efficient inference. The results obtained for untransformed CIR process are in line with those obtained for Stramer and Yan (2007). However, in many inference problems, the accuracy achieved in Section 4.6 may not be enough and higher K will have to be used to get better results.

Choosing $\Delta\Xi$: $\Delta\Xi$ governs how fine (or coarse) the grid constructed on the support Ξ^* is. Because GaMBA-I produces discretised posteriors, a smaller $\Delta\Xi$ implies a finer grid and a better approximation to the true continuous posterior. As illustrated in the examples, $\Delta\Xi$ will be different for every parameter, and its value will depend on the length of the support for that parameter. As a rule-of-thumb,

$\Delta\Xi$ should be small enough so that the grid selects at least 10 values from the support for each parameter. This means if there are 3 parameters, $|\mathcal{G}_{\Xi^*}| > 1000$. It is very important to note that this rule-of-thumb strategy advised above is applicable only to symmetric uni-modal posterior distributions. If there are reasons to believe that the posterior distribution may be highly skewed or multi-modal in at least one of the dimensions then much finer $\mathcal{G}_{\Xi^*}$ might be required.

No dependency between X and Θ : As described in Section 4.4, GaMBA (and also GaMBA-I) get around the problem of dependency between the latent process $\mathbf{X}$ and the parameters Θ. This means that there is no need to re-parameterise an SDE before inference and thus GaMBA can be applied to a wider class of SDE models; for example: to multivariate SDE's.

Ease of Implementation : Unlike the MCMC based methods, GaMBA-I does not need the SDE to be re-parameterised – this is illustrated from the results obtained for un-transformed CIR process in Section 4.6. Further unlike the simulated likelihood methods, GaMBA-I does not involve use of numerical methods. This makes GaMBA-I both mathematically as well as computationally simpler to implement and more widely applicable.

4.8.3 Computational Efficiency

- Computational effort required for implementing the MCMC based method described in Section 3.6.2 is directly proportional to M and the number (R) of the samples needed to be drawn using the Markov chains. Computational effort required for implementing GaMBA-I is directly proportional to M, the cardinality of $\mathcal{G}_{\Xi^*}$ and K.

- Computational effort required by GaMBA-I (with $K = 1$) for *one* value of Θ is only slightly smaller than the effort required by MCMC method specified. This slight advantage is due to the fact that unlike MCMC (which samples new Θ from $P(\cdot|\mathbf{X}, \mathbf{Y})$), GAMBA-I takes selects Θ from a pre-determined set of values.

- Let $\mathcal{T}_M$ and $\mathcal{T}_{GI}$ be the computational time required by the MCMC method and GaMBA-I respectively. Then, ignoring the small difference mentioned in the previous point, and assuming that same M is used for both MCMC and GaMBA-I, the computational advantage that GaMBA-I has over MCMC is

$$\mathcal{T}_M - \mathcal{T}_{GI} \propto R - |\mathcal{G}_{\Xi^*}| \times K.$$

- As seen from the Euro-Dollar data example, accurate inference using MCMC requires a large R. Further, a smaller M might be sufficient for inference using GaMBA-I, as seen in that example. Thus the computational advantage mentioned above is expected to be significant for many SDE models.

- In addition to the above computational advantage, the computational effort required for GaMBA-I can be reduced multi-fold using multiple parallel processing computing units. Because multi-core processors have now become a standard in personal computers, this does not require the user to have access to external servers. As discussed in Section 2.4.3.6, it is not straightforward to implement MCMC based methods using parallel-processing capabilities.

- Computational effort required for GaMBA is same as the computational effort required for GaMBA-I with $K = 1$.

Note that parallel processing has not been implemented in any of the examples mentioned in this thesis. However, because this technology has a significant potential to improve the computational efficienty of GaMBA-I, presented below is the algorithm

of how it could be implemented.

Algorithm for parallel processing of GaMBA-I:

Assume that P parallel processing units are available.

1. Identify Ξ^* (as described in Section 4.2.1.)
2. Define a discrete grid $\mathcal{G}_{\Xi^*}$ on Ξ^*.
3. Divide the total number of points on $\mathcal{G}_{\Xi^*}$ on Ξ^* into P subsets (mutually exclusive and exhaustive) and assign each subset to a different parallel processing unit.
4. For each of the subsets, say $\mathcal{G}_{\Xi^*}^{(p)}$, for $p = 1, \cdots, P$, implement the following steps $5(a)$ and $5(b)$ simultaneously (using the parallel processing units) for each of the subsets
5. For each point on the subset $\Theta_j \in \mathcal{G}_{\Xi^*}^{(p)}$
   (a) for $k = 1, \cdots, K$,
   Sample $\mathbf{X}_k \sim P_{MBB}(\mathbf{X}_k|\Theta_j, \mathbf{Y})$ as described in Section 4.3.2
   (b) Evaluate
   $$P_{GaMBA-I}(\Theta_j|\mathbf{Y}) \propto \frac{1}{K}\sum_{k=1}^{K}\frac{P_E(\mathbf{Y}, \mathbf{X}_k|\Theta_j)\cdot P(\Theta_j)}{P_{MBB}(\mathbf{X}_k|\Theta_j, \mathbf{Y})}$$
   .
6. Normalise to obtain $P(\Theta|\mathbf{Y})$ over $\mathcal{G}_{\Xi^*}$.

where:

- $P_E(\mathbf{Y}, \mathbf{X}|\Theta_j)$ is the Euler's density as in Equation (4.7);

- $P(\Theta_j)$ a suitable prior density;
- $P_{MBB}(\mathbf{X}|\mathbf{Y}, \Theta_j)$ is the MBB density as in Equation (4.17).

When GaMBA-I is implemented using parallel processing, it is expected that its computational advantage over the MCMC would be

$$\mathcal{T}_M - \mathcal{T}_{GI} \propto R - \frac{|\mathcal{G}_{\Xi^*}| \times K}{P}.$$

4.8.4 Limitations

Ξ^* is not known : In its present form, GaMBA-I assumes that the support Ξ^* is known due to prior knowledge or otherwise. This is a very important limitation, since this assumption may not be valid for many practical modeling problems. As mentioned in Section 4.2.1, it may be possible to develop a methodology to objectively identify Ξ^*, however this is a separate research problem in itself at present remains a open problem.

Dimensionality of the parameter space Ξ : While implementing GaMBA-I, the parameter values are sampled from the space Ξ^* by constructing a regular grid on this space. Though this may possibly be the simplest way to sample, it may not be the most efficient one. The grid sampling method works well when the parameter space is five-dimensional or less, but beyond that it very rapidly becomes computationally too expensive. Thus, GaMBA-I may not be computationally efficient for SDE models with five or more parameters.

Sparsely observed data : Section 4.7 illustrates that GaMBA-I will not be efficient when data are too sparsely observed. Further, how closely the data need to

observed also depends on the non-linearity of the process. In general, if the data are not observed closely enough, larger M and K would be required to obtain accurate inference and GaMBA-I may not be as computationally efficient in this case. This clearly limits the potential of GaMBA-I to be applicable in situations where the data is sparsely observed.

Chapter 5

Modeling Dynamic Force using Stochastic Differential Equations

This Ph.D project was motivated by the need to develop stochastic process models to understand the dynamics of the accumulation of damage to a road surface. The factors which cause this degradation are the forces exerted by the vehicles, weather conditions and the materials used to construct the road. The extent of degradation at a given point in time also varies spatially. Eventually, given a road surface, it would be desired to be able to predict the distribution of its time to failure. Further it is also hoped that resulting statistical modeling will provide a better understanding of the uncertainties involved at various stages and thus will eventually also help in building better roads.

The dynamic force exerted by vehicles on the road surface is a very important factor in road degradation, and investigating the relationship between the force exerted by the vehicle and the mass of the vehicle is known as the Weight-in-Motion (WIM) problem in the engineering literature. Typically the force sensors are placed in the road surface on a small patch of the road, and the forces exerted by each axle of the vehicle are measured as the vehicle travels over these sensors at the usual speed. The problem

is then to estimate the mass of each of the vehicle axles using the corresponding set of measured forces.

In the statistics literature, a hierarchical Bayesian model was used to model these dynamics and the Bayesian inference was implemented using the standard MCMC methods by Wilson et al. (2008). The existing engineering models used to capture these dynamics are differential equation models based on Newton's second law. One such model is briefly reviewed in Section 5.1. The motivation behind this work was to check if better models could be built by using stochastic differential equations (SDE) instead, where the inference can be derived using Bayesian methodology. Since the SDE models capture the inherent uncertainty associated with physical processes, it was desired to investigate if an SDE model based on a simpler engineering model, can sufficiently capture these dynamics.

This work is in collaboration with Prof. Eugene O'Brien of the School of Architecture Landscape and Civil Engineering at University College Dublin. The engineering models used in this research are provided by Prof. O'Brien's team.

5.1 Background

This section provides a brief description of a few concepts necessary to model the damage to the road surface. A detailed background could be found in Tedesco et al. (1999), Cebon (1999),Harris (2007), and Tegegn (2007).

5.1.1 Spatial Repeatability (SR)

One of the principal factors that causes road damage is the dynamic force imposed on the road surface by heavy vehicles, principally trucks. It has been observed both experimentally, as well as numerically (Ervin (1983),Mitchell (1987),Huhtala et al. (1992)), that the pattern of dynamic tyre forces applied by the truck axles to a road surface is similar for repeated runs at similar speeds. This phenomenon is called *Spatial Repeatability (SR)*. This basically implies that a road surface is likely to be affected more at particular areas and less affected at others along its length.

5.1.1.1 Statistical Spatial Repeatability (SSR)

This is an extension to the concept of *spatial repeatability*. *SSR* states that the mean of many patterns of dynamic tyre forces applied to a pavement surface is similar for a fleet of trucks of a given type. It has been shown experimentally (O'Connor et al. (2000)) that the mean pattern of the forces exerted is similar for many trucks of the same type.

5.1.2 Models capturing Road-Vehicle interaction

Existing engineering models used to capture the road-vehicle dynamics are based on Newton's second law whereby the resulting force F is described as

$$\begin{aligned} F &= m \cdot a \\ F &= m \cdot u'' \end{aligned} \tag{5.1}$$

where m is the mass, a is the acceleration, u is the displacement, and u'' is its second derivative with respect to time t. In general, if there are multiple forces (say k different

forces), then we have

$$F_1 + F_2 + \cdots + F_k = m \cdot u''. \tag{5.2}$$

The number of masses considered in a model determine the 'degree of freedom' (DOF) for that model. The higher the degree of freedom, the more accurately does the model capture the true dynamics. However, such models also become increasingly more complicated. Here, only the simplest model has been considered.

5.1.2.1 Single DOF model

This is the most basic of the model. It represents the vehicle as a system consisting of a single mass, a spring and a viscous damping on a fixed road surface. It does not perfectly capture the dynamics of the force exerted by a vehicle, but is very easy to evaluate. Figure 5.1 illustrates the Model.

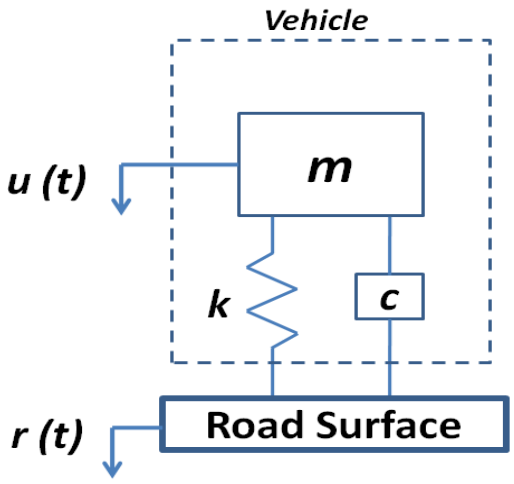


Figure 5.1: Single DOF Model

The vehicle is excited by the pavement roughness $r(t)$ measured in terms of road

elavation. The equations of motion for a one DOF model are given by (Tedesco et al. (1999), pg. 129)

$$m \cdot u''(t) + c \cdot (u'(t) - r'(t)) + k \cdot (u(t) - r(t)) = 0 \tag{5.3}$$

where $u'(t)$ and $u''(t)$ represent the first and the second derivatives of the vertical displacement $u(t)$, and $r'(t)$ represents the first derivative of the external excitation (road surface) at t.

Note that, Equation (5.3) can be written as

$$m \cdot u''(t) = -(G(t) + F(t)), \tag{5.4}$$

where

$$G(t) = c \cdot (u'(t) - r'(t)) \tag{5.5}$$

is the (absorbed) force due to damping c, and

$$F(t) = k \cdot (u(t) - r(t)) \tag{5.6}$$

is the resulting dynamic force exerted on the road surface.

5.2 Modeling Dynamic Forces

A small patch of the road (a few meters in length), is fitted with sensors which can measure the force exerted by every vehicle as it traverses over the sensors. Thus, if there are p sensors, then for every vehicle, the forces are measured at p different locations. Figure 5.2 illustrates how sensors are located. Thus data consists of the observed forces captured using these sensors.

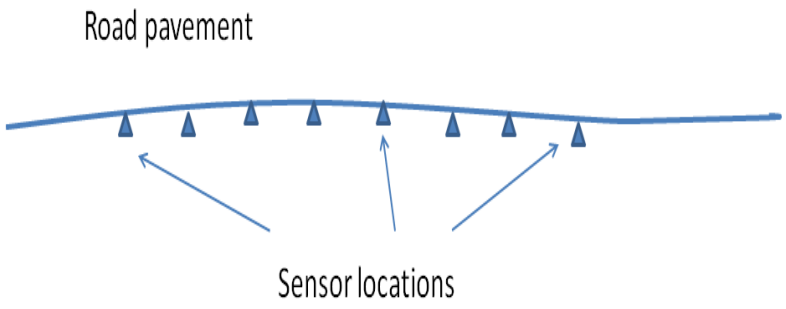


Figure 5.2: Sensors to measure the force

Consider Equation (5.3) corresponding to the single DOF model, which provides the relationship between the vertical displacement $u(t)$ of an axle and the known pavement roughness $r(t)$. The solution of this differential equation is the vertical displacement $u(t)$. However what can be observed in practice, are the forces $F(t)$ and not the displacements $u(t)$. In order to use this model to model dynamic forces, the model needs to be reparameterised, so that the solution of the model is now the dynamic force $F(t)$ instead.

5.2.1 Single DOF model for dynamic force

This force $F(t)$ can be expressed using Equation (5.6), and we have

$$u(t) = F(t)/k + r(t) \tag{5.7}$$

$$u'(t) = F'(t)/k + r'(t) \tag{5.8}$$

$$u''(t) = F''(t)/k + r''(t). \tag{5.9}$$

Substituting Equations (5.7), (5.8) and (5.9) in Equation (5.3), results in the fol-

lowing differential equation

$$m \cdot F''(t) + c \cdot F'(t) + k \cdot F(t) = -m \cdot r''(t) \cdot k. \tag{5.10}$$

whose solution is the force $F(t)$. Equation (5.10) can now be used to model the dynamic forces measured using the sensors. The road excitation $r(t)$ are known.

5.2.2 Building an SDE model

The equation of Force obtained by solving this single DOF model is given by (Tedesco et al. (1999), pg. 152)

$$F(t) = F_0(t) \cdot [1 - \exp^{-\mu\omega_n t}(\cos(\omega_d t) + \frac{\mu}{\sqrt{(1-\mu^2)}}\sin(\omega_d t))] \tag{5.11}$$

where $F_0(t) = -m \cdot r''(t) \cdot k$, $\omega_n = \sqrt{k/m}$, $c_r = 2\sqrt{m/k}$, $\mu = \frac{c/k}{c_r}$, and $\omega_d = \omega_n\sqrt{(1-\mu^2)}$.

Equation (5.11) provides a deterministic solution for the dynamic force. It is possible to use this equation to build a stochastic differential equation (SDE) model for this dynamic force.

Such a model will have the general form

$$d\,F(t) = F'(t)\,dt + g(\cdot, t)\,dW(t) \tag{5.12}$$

where

$$F'(t) = \frac{d\,F(t)}{dt}$$

and $g(\cdot, t)$ is some function which is believed to capture the uncertainty in the process.

The objective of this chapter is to explore if an SDE based model such as Equation (5.12) could be used to model the dynamic force exerted by the vehicles. However for such a modeling exercise to be meaningful, there has to be enough justification to believe that the vehicle-force interaction is indeed stochastic in nature. After discussions with the collaborators, it emerged that though they believe this interaction to be largely deterministic, but the uncertainty could stem from the spring stiffness coefficient k.

It could be possible to capture this uncertainty in the SDE model using a suitable diffusion term $g(\cdot, t)$. One way to do this is to define g as a linear function of the spring stiffness coefficient k. Thus, we have

$$g(\cdot, t) = \theta \cdot k. \tag{5.13}$$

Using Equations (5.12) and (5.13), the SDE for the dynamic force is given by

$$d\,F(t) = F'(t)\;dt + \theta \cdot k\;dW(t). \tag{5.14}$$

Note that, Equation (5.14) provides one way of modeling the force exerted by the vehicles on the road surface using an SDE. Also note that, the above SDE is linear and both its drift and diffusion coefficients are deterministic - the randomness only comes from the Wiener process components. Thus, this is not a particularly challenging SDE to infer and infering this process using GaMBA and MCMC is mainly of academic interest.

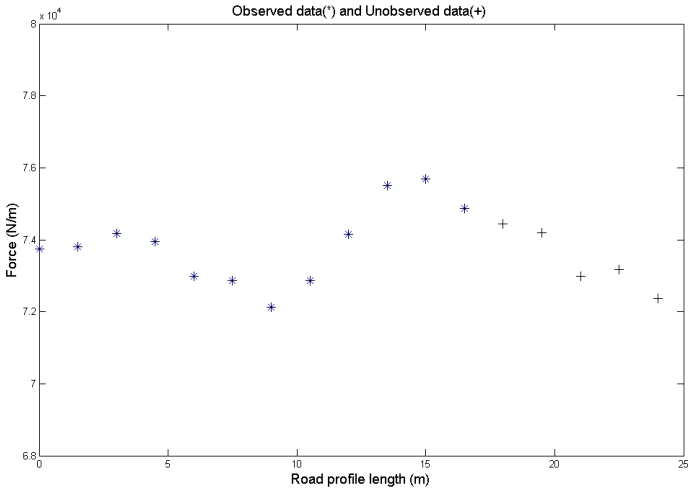


Figure 5.3: Simulated data: Observed (*) and unobserved (+)

5.3 SDE Modeling for the Simulated Data

The aim of this exercise was to infer vehicle properties — namely the mass and the stiffness of the vehicle — having observed the forces exerted by it on a given road surface. The SDE of Equation (5.14) was used to model the relationship between the force and the vehicle properties. It was assumed that a weak solution to Equation (5.14) exists. The inference on the parameters of Equation (5.14) would be done using the Bayesian methodology. Both MCMC and GaMBA were used to implement Bayesian inference on these parameters to compare the speed and the accuracy of these two methods.

The data were simulated using the Q-C model (a higher order deterministic differential equation model) with added Gaussian noise. The road surface was 24.5 meters in length and sensors were placed at every 1.5 meters; thus there were 17 data points (one corresponding to each sensor) in all. As shown in Figure 5.3, the first 12 data points were considered as 'observed', and the last 5 were used to check the accuracy of the prediction intervals.

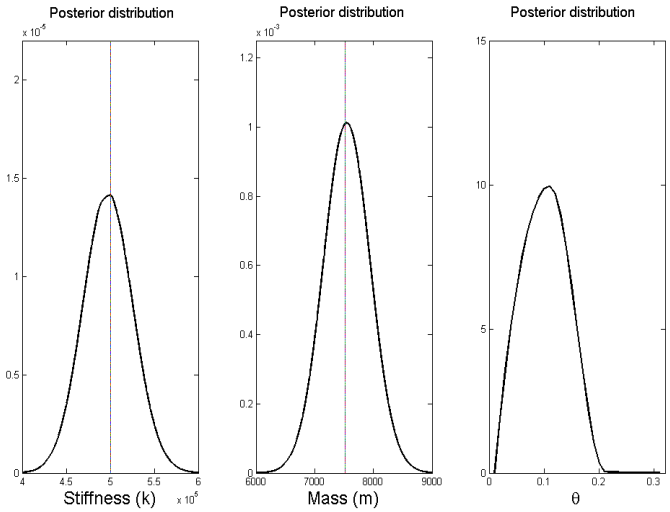


Figure 5.4: Posterior distributions obtained using GaMBA along with the true values for k and m shown by the vertical lines.

Among all vehicular traffic, the main contributor to the road degradation are the fully loaded large trucks (averaging in weight at about 40 tonnes each). The axle used in the above set-up belonged to a 5-axle truck of this type. Therefore, it was possible and reasonable to use an informative prior for the mass. This prior was chosen to be a Gaussian distribution with mean 8000 Kg and the standard deviation of 400, i.e. $N(8000, 400)$. Similarly, it is possible to choose an informative prior on the stiffness k, and was chosen as $N(550 \times 10^3, 15 \times 10^3)$. However, there is no background information on θ, and therefore the prior for θ was chosen to be $U(0.01, 0.31)$.

Based on this prior knowledge, the parameter space was chosen as $\Xi^* = [6000, 9000] \times [400 \times 10^3, 600 \times 10^3] \times [0.01, 0.31]$. The grid $\mathcal{G}_{\Xi^*}$ was constructed with $\Delta\Xi_1 = 20 \times 10^3$, $\Delta\Xi_2 = 100$ and $\Delta\Xi_3 = 0.1$. GaMBA was implemented on the $1,367$ points thus sampled from Ξ^* and marginal posteriors distributions were obtained.

Figure 5.4 shows the marginal posterior distributions obtained using GaMBA. The true values of the parameters were later revealed to be $m = 7524$, and $k = 500 \times 10^3$ and are depicted using vertical lines. *Note that there is no true value for θ, and that it has been used as a nuisance parameter to capture the uncertainty regarding the true track of travel.*

Sampling the parameter values from the joint posterior distribution, and then simulating the SDE forward in time lends the posterior predictive distribution for the 'unobserved' (data points 13 through 17) sensors. Figure 5.5 shows the 95% bounds for this predictive distribution and also its median. It can be seen that both the observed as well as the unobserved data compare well against the median of the predictive distribution indicating a good fit of the model.

The MCMC described in Section 3.6 was implemented on this data. Visual check of the MCMC trace plot along with the correlograms were used to assess stationarity. First $5,000$ samples were discarded as the 'burn-in' period, and the next $5,000$ samples were chosen as the correlated draws from the stationary distribution. Figure 5.6 shows the MCMC trace plots along with the correlograms. The MCMC posteriors were plotted along with the GaMBA posteriors and are shown in Figure 5.7, where the vertical lines depict the true values of stiffness k and mass m. In order to assess how closely the results from GaMBA agree with those obtained using MCMC, the posterior distribution functions (CDFs) obtained using the two methods were plotted together and are shown in Figure 5.8.

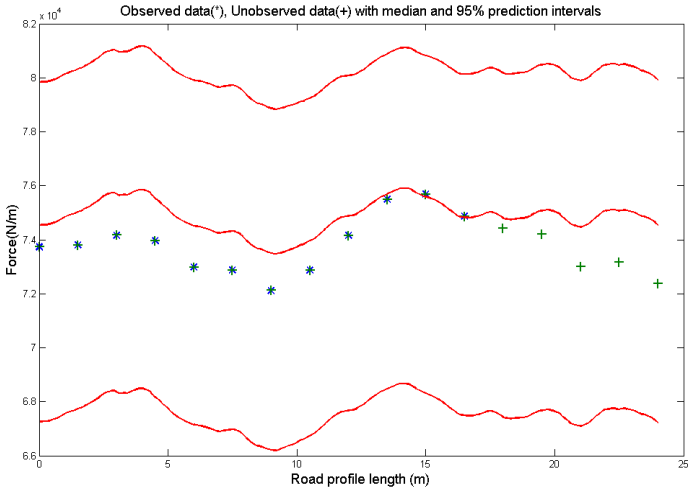


Figure 5.5: Simulated data with 95% Prediction intervals and the median prediction level using GaMBA.

Figure 5.9 shows the prediction plot obtained using MCMC posteriors. It can be seen that GaMBA posteriors correctly identify the true value - however the dispersion of GaMBA posteriors is considerably different from those obtained using MCMC. But while GaMBA takes 30 seconds, MCMC takes more than 4 minutes.

Thus, GaMBA turns out to be nearly 8 times faster compared to a standard MCMC scheme in this case.

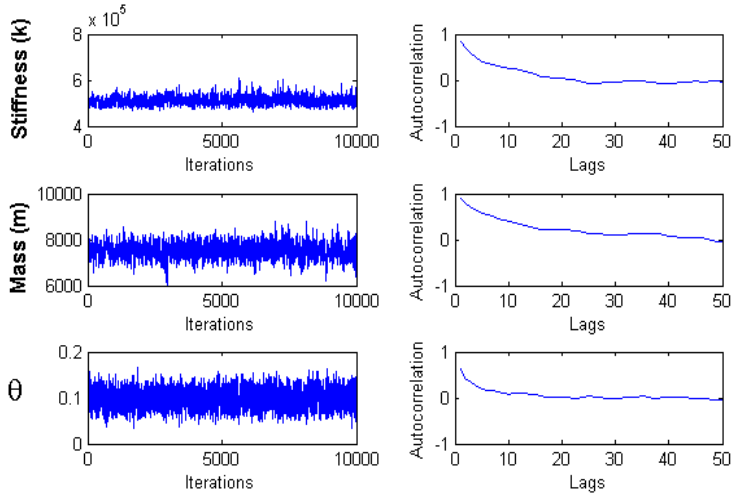


Figure 5.6: MCMC trace plots along with their correlograms.

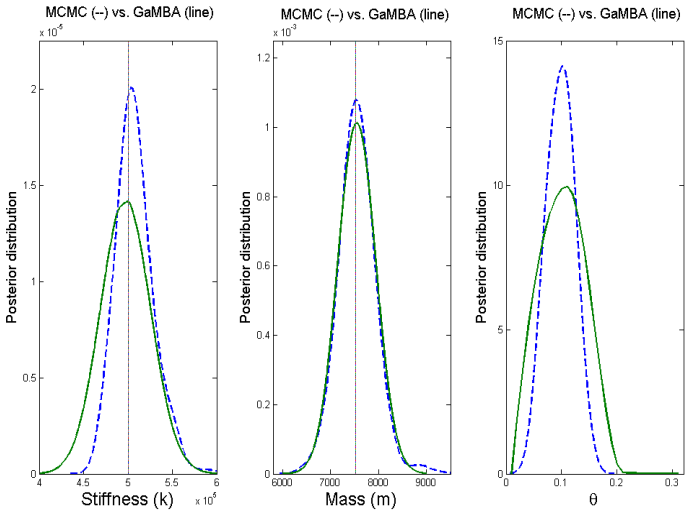


Figure 5.7: MCMC posteriors (–) plotted over GaMBA posteriors (line) along with the true values for k and m shown by the vertical lines.

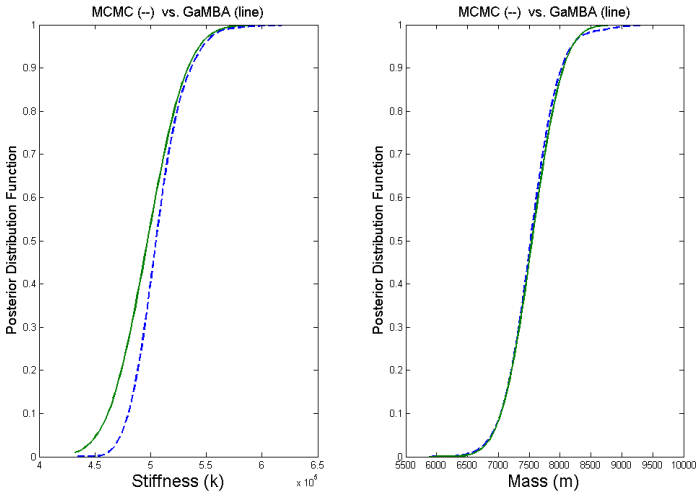


Figure 5.8: Distribution functions using MCMC(red) plotted over GaMBA(blue) for (a) k and (b) m.

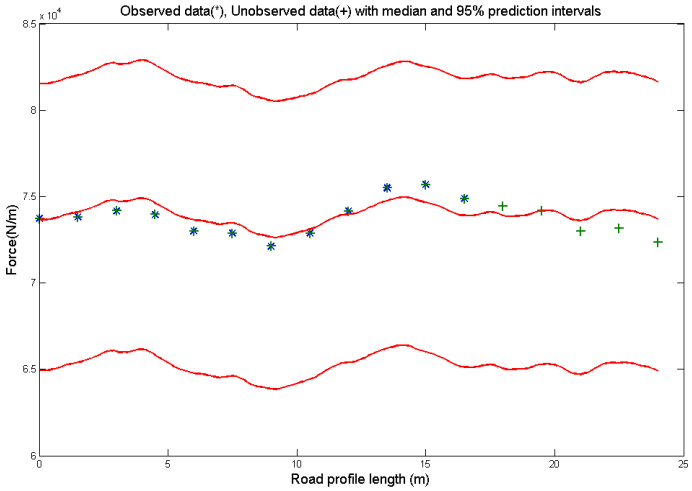


Figure 5.9: Simulated data with 95% Prediction intervals and the median prediction level using MCMC.

5.4 Discussion

This Ph.D. work was motivated by the need to develop better understanding of the dynamic relationship between the vehicle properties and their effect on a road surface. Specifically the purpose was to achieve better statistical inference on vehicle properties such as mass and stiffness having observed the forces exerted by the vehicle on the road surface.

This chapter first reviews the necessary engineering concepts, then develops a stochastic differential equation model to model the relationship between the vehicle properties and the force exerted by the vehicles. The authors are not aware of any other work where SDE's have been used to model this relationship. This model building process has been exploratory in nature and the model built is mainly of academic interest only. It is however shown, that inference on this SDE model can be obtained at 8 times less computational cost using GaMBA.

Chapter 6

Conclusions & Further Work

6.1 Conclusions

A new approach for Bayesian inference on stochastic differential equation (SDE) models has been proposed. This approach is *not MCMC based* and is inspired from the work of Rue et al. (2009) on the Integrated Nested Laplace Approximation (INLA) for Gaussian Markov Random Field (GMRF) models. This thesis introduces two new methods to implement this approach. These methods have been named as the Gaussian Modified Bridge Approximation (GaMBA) and its extension GaMBA- Importance sampling (GaMBA-I). This thesis provides an easy to use algorithm for both these methods, discusses their consistency properties, describes examples where these methods provide efficient inference and also illustrates situations where these methods would not yield efficient and accurate inference.

More importantly, this thesis provides a general framework that can be used for Bayesian inference on SDE models rather than using the MCMC based methods. As the research progresses and better diffusion bridge approximations become available, better computational methods are discovered, they can be incorporated into this framework

to make feasible efficient Bayesian inference for a wider class of SDE models. Further, since this approach is computationally cheaper than the MCMC based methods, its use has the potential to make possible the inference on several highly complex processes using SDE models.

This research has also attempted to model the dynamic force exerted by the vehicles on the road surface using SDE models. As far as the author and the collaborators are aware, this has not been done so far. An SDE model based on one of the existing differential equation models was used to fit a simulated force data using GaMBA. This was considered as a *'proof of concept'* work to investigate if the SDE modeling of this problem is feasible.

6.2 Further Work

While developing GaMBA, several questions and several new ideas emerged leading to following possible areas for future research.

1. There remains the need to develop a diffusion bridge construct which would be easy to implement, applicable on a wide range of SDE models, computationally cheaper but whose paths would converge in distribution to the true diffusion. As discussed in Chapter 4, the MBB density though easy to use, is not very useful when data are too sparsely observed.

   One approach to overcoming the limitations induced by MBB might be to develop a diffusion bridge construct with an inflated variance and this option needs to be explored.

2. It is needed to develop more objective and widely applicable methods for choosing Ξ^*. For all the examples carried out in this thesis, Ξ^* was chosen based on prior

knowledge, which can be subjective. It might be possible to develop a method to do this based on Rue et al. (2009). This and other options need to be further explored.

3. In its current form, the parameter values are sampled from the space Ξ^* by constructing a regular grid on this space. Though this may possibly be the simplest way to sample, it may not be the most efficient one. The grid sampling method works well when the parameter space is five-dimensional or less, but beyond that it very rapidly becomes computationally too expensive.

   By developing more efficient methods to sample from Ξ^*, it might be possible to get an even faster inference using GaMBA. Also, it might allow GaMBA to be used on SDE models with more unknown parameters.

4. Another possible extension could be to devise efficient ways to implement GaMBA on models such as the *stochastic volatility* model, where there is a sequential dependence between the different SDEs. Efficient methods for sampling from Ξ^* may also need to be used in conjunction.

5. So far, GaMBA has been developed assuming that the data has been observed without error. It is desirable to extend GaMBA to include the cases when the data are observed with errors.

6. Finally, it is important to note that diffusion processes and diffusion bridges are continuous in both state space and time. Since GaMBA is based on a set-up which involves discretising the time between the two observed data points, even when a diffusion bridge with the desirable asymptotic convergence properties is used, it will still be subject to the discretisation error. Exact methods which sample continuous diffusion bridges (see for example, Beskos et al. (2008)) are already available, but involve MCMC based methods.

Since the objective was to develop a method computationally much cheaper than the MCMC based methods, it is not immediately clear if these exact methods could be used within the GaMBA framework, and if yes, then how. Therefore, this remains an area open for further work.

Bibliography

Ait-Sahalia, Y. (2002). Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica 70*(1), pp. 223–262.

Bernardo, J. M. and A. F. M. Smith (2000). *Bayesian Theory* (First ed.). Chester: Wiley.

Beskos, A., O. Papaspiliopoulos, G. Roberts, and P. Fernhead (2006). Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *Royal Stat. Soc. B 68*, 333–382.

Beskos, A., G. Roberts, A. Stuart, and J. Voss (2008). MCMC methods for diffusion bridges. *Stochastics and Dynamics 8*, 319–350.

Bladt, M. and M. Sorensen (2010). Simple simulation of diffusion bridges with application to likelihood inference for diffusions. Working paper, University of Copenhagen. Available at 'http://www.math.ku.dk/ michael/papers.

Brandt, M. W. and P. Santa-Clara (2002). Simulated likelihood estimation of diffusions with application to exchange rate dynamics in an incomplete market. *Journal of Financial Economics. 63*, 161–210.

Brockwell, A. (2006). Parallel markov chain monte carlo simulation by pre-fetching. *Journal of Computational and Graphical Statistics 15*, pp. 246 – 261.

Cappé, O., E. Moulines, and T. Rydén (2005). *Iinference in Hidden Markov Models* (First ed.). USA: Springer.

Cebon, D. (1999). *Handbook of Vehicle-Road Interaction* (First ed.). NY 10016, USA: Taylor & Francis.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician 49*, 327–335.

Chib, S., M. Pitt, and N. Shephard (2006). Likelihood based inference for diffusion driven models. Olin School of Business, Washington University.

Chib, S. and N. Shephard (2002). Comment on Garland B. Durham and A. Ronald Gallant 'numerical techniques for maximum likelihood estimation of continuous-time diffusion processes'. *J.Busi.& Eco. Statis. 20*, 325–327.

Cowles, M. and B. Carlin (1996). Markov chain monte Carlo convergence diagnostics: a comparitive review. *J.American Stat. Ass. 91*, 883–904.

Cox, D. (2006). *Principles of Statistical Inference* (First ed.). New York, USA: Cambridge.

Del Moral, P., J. Jacod, and P. Protter (2002). The monte carlo method for filtering with discrete-time observations. *Probability Theory and Related Fields 120*, pp. 346 – 368.

Delyon, B. and H. Ying (2006). Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Processes and their Applications 116*, 1660–1675.

Dudley, R. (2003). *Real Analysis and Probability* (Second ed.). New York, USA: Cambridge.

Durett, R. (1996). *Stochastic Calculus: A Practical Introduction.* Boca Raton: CRC Press.

Durham, G. B. and R. Gallant (2002). Numerical techniques for maximum likelihood estimation of continuous time diffusion processes. *J.Busi.& Eco. Statis. 20*, 297–316.

Elerian, O., S. Chib, and N. Shephard (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica 69*, 959–993.

Ervin, R. (1983). Influence of truck size and weight variables on the stability and control properties of heavy trucks. Technical Report UMTRI-83-10/2, University of Michigen Transport Research Institute.

Fernhead, P., O. Papaspiliopoulos, G. Roberts, and A. Stuart (2010). Random-weight particle filtering of continuous time processes. *Royal Stat. Soc. B 72*, 497–512.

Friedman, A. (1975). *Stochastic Differential Equations and Applications: Volume* 1 (First ed.). London: Academic Press.

Gelman, A., J. B. Carlin, H. Stern, and D. Rubin (2003). *Bayesian Data Analysis* (Second ed.). London, UK: Chapman & Hall.

Gelman, A. and D. Rubin (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7(4), 457–472.

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica 57*(6), pp. 1317–1339.

Gilks, W., N. Best, and K. Tan (1994). Adaptive rejection metropolis sampling within gibbs sampling. *Journal of the Royal Statistical Society. Series C 44*, 455–472.

Gilks, W., S. Richardson, and D. Spiegelhalter (1996). *Markov Chain Monte Carlo in Practice* (First ed.). London, UK: Chapman & Hall.

Golightly, A. and D. Wilkinson (2006). Bayesian sequential inference for nonlinear multivariate diffusions. *Statist. Comput. 16*, 323–338.

Golightly, A. and D. Wilkinson (2007). Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis 52*, 1674–1693.

Grimmett, G. and D. Stirzaker (2001). *Probability and Random Processes* (Third ed.). Oxford, UK: Oxford.

Harris, N. (2007). Characterisation of factors affecting dynamic heavy vehicle infrastructure interaction. Ph.D. Thesis, School of Architecture, Landscape & Civil Engineering, University College Dublin.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57*(1), 97–109.

Heyde, C. C. (1997). *Quasi-Likelihood and Its Applications: A General Approach to Optimal Parameter Estimation*. New York: Springer-Verlag.

Huhtala, M., J. Pihlajamak, and P. Halonen (1992). WIM and dynamic loading on pavements. In: D.Cebon et al. eds. Third International Symposium on Heavy Vehicle Weights and Dimensions, Cambridge, UK.

Jacob, P., C. Robert, and M. Smith (2010). Using parallel computation to improve independent metropolis-hastings based estimation. *available from Cornell University Library http://arxiv.org/abs/1010.1595*.

Kessler, M. (2000). Simple and explicit estimating functions for a discretely observed diffusion process. *Scand. J. Stat. 27*, pp. 65–82.

Kloeden, P. and E. Platen (1992). *Numerical Solution of Stochastic Differential Equations* (First ed.). New York, USA: Springer.

Koralov, L. and Y. Sinai (2007). *Theory of Probability and Random Processes* (Second ed.). New York, USA: Springer.

Kutoyants, Y. A. (2004). *Statistical Inference for Ergodic Diffusion Processes* (First ed.). London: Springer.

Lacus, S. M. (2008). *Simulation and Inference for Stochastic Differential Equations* (First ed.). New York, USA: Springer.

Lyons, T. and W. Zheng (1990). On conditional diffusion processes. *Proc. Roy. Soc. Edinburgh Sect. A 115*, 243–255.

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*(6), 1087–1092.

Mitchell, C. (1987). The effect of the design of goods vehicle suspension on loads and bridges. Project Report 115, Transport Research Laboratory, United Kingdom.

Neal, P. and G. Roberts (2006). Optimal scaling for partially updating MCMC algorithms. *Ann. of Applied Probability 16*, 475–515.

O'Connor, T., E. O'Brien, and B. Jacob (2000). An experimental investigation of spatial repeatability. *International Journal of Heavy Vehicle Systems 7*, 64–81.

Oksendal, B. (2007). *Stochastic Differential Equations: An introduction with applications* (Sixth ed.). New York, USA: Springer.

Pedersen, A. R. (1995a). Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli. 1*, 257– 279.

Pedersen, A. R. (1995b). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete approximations. *Scand. J. Stat. 22*, 55–71.

Polson, N. G. and G. O. Roberts (1994). Bayes factors for discrete observations from diffusion processes. *Biometrika 81*(1), pp. 11–26.

Prakasa Rao, B. (1999). *Statistical Inference for Diffusion Type Processes*. New York, USA: Oxford University Press.

Robert, P. C. and G. Casella (2004). *Monte Carlo Statistical Methods* (Second ed.). USA: Springer.

Roberts, G., A. Gelman, and W. Gilks (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. of Applied Probability 7*, 110–120.

Roberts, G. and J. Rosenthal (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science 16*, 351–367.

Roberts, G. and J. Rosenthal (2004). General state space Markov chains and MCMC algorithms. *Probability Surveys 1*, 20–71.

Roberts, G. and J. Rosenthal (2006). Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains. *Ann. of Applied Probability 16*, 2123–2139.

Roberts, G. and O. Stramer (2001). On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika 88*, 603–621.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of Royal Statistical Society, Series B 71*, 1–35.

Santa-Clara, P. (1995). Simulated likelihood estimation of diffusion with application to the short term interest rate. *Anderson Graduate School of Management, University of California at Los Angeles.*.

Sarkka, S. and T. Sottinen (2008). Application of girsanov theorem to particle filtering of discretely observed continuous time non-linear systems. *Bayesian Analysis 3*, pp. 555– 584.

Sorensen, H. (2004). Parametric inference for diffusion processes observed at discrete points in time: a survey. *Int. Stat. Rev. 72*, 337–354.

Sorensen, M. (1999). On asymptotics of estimating functions. *Brazilian Journal Probability and Statistics 13*, 111–136.

Stirzaker, D. (2005). *Stochastic Processes & Models* (First ed.). Oxford, UK: Oxford.

Stramer, O. and J. Yan (2007). Asymptotics of an efficient monte carlo estimation for the transition density of diffusion processes. *9*, 483–496.

Tedesco, J. W., W. G. McDougal, and C. A. Ross (1999). Strucural dynamics - theory and applications.

Tegegn, A. B. (2007). Spatial repeatability of heavy vehicle axle forces and the implications for pavement degradation. Ph.D. Thesis, School of Architecture, Landscape & Civil Engineering, University College Dublin.

Tierney, L. (1996). Introduction to general state-space Markov chain theory. In: Markov chain Monte carlo in practice, Chapman & Hall, Chapter 4.

Uhlenbeck, G. E. and L. S. Ornstein (1930, Sep). On the theory of the Brownian Motion. *Phys. Rev. 36*(5), 823–841.

Vasicek, O. (1977). An equilibrium characterisation of the term structure. *Journal of Financial Economics 5*, 177–188.

Wilson, S., N. Harris, and E. O'Brien (2008). The use of Bayesian statistics to predict patterns of spatial repeatability. *Journal of Sound and Vibration To Appear*.