THE UNIVERSITY OF WARWICK

Library Declaration and Deposit Agreement

1. STUDENT DETAILS

Please complete the following:

Full name: ……

University ID number: ……

2. THESIS DEPOSIT

2.1 I understand that under my registration at the University, I am required to deposit my thesis with the University in BOTH hard copy and in digital format. The digital version should normally be saved as a single pdf file.

2.2 The hard copy will be housed in the University Library. The digital version will be deposited in the University's Institutional Repository (WRAP). Unless otherwise indicated (see 2.3 below) this will be made openly accessible on the Internet and will be supplied to the British Library to be made available online via its Electronic Theses Online Service (EThOS) service.
[At present, theses submitted for a Master's degree by Research (MA, MSc, LLM, MS or MMedSci) are not being deposited in WRAP and not being made available via EthOS. This may change in future.]

2.3 In exceptional circumstances, the Chair of the Board of Graduate Studies may grant permission for an embargo to be placed on public access to the hard copy thesis for a limited period. It is also possible to apply separately for an embargo on the digital version. (Further information is available in the *Guide to Examinations for Higher Degrees by Research*.)

2.4 *If you are depositing a thesis for a Master's degree by Research, please complete section (a) below. For all other research degrees, please complete both sections (a) and (b) below:*

(a) Hard Copy

I hereby deposit a hard copy of my thesis in the University Library to be made publicly available to readers (please delete as appropriate) EITHER immediately OR after an embargo period of ……..... months/years as agreed by the Chair of the Board of Graduate Studies.

I agree that my thesis may be photocopied. YES / NO (*Please delete as appropriate*)

(b) Digital Copy

I hereby deposit a digital copy of my thesis to be held in WRAP and made available via EThOS.

Please choose one of the following options:

EITHER My thesis can be made publicly available online. YES / NO (*Please delete as appropriate*)

OR My thesis can be made publicly available only after…..[date] (Please give date)
YES / NO (*Please delete as appropriate*)

OR My full thesis cannot be made publicly available online but I am submitting a separately identified additional, abridged version that can be made available online.
YES / NO (*Please delete as appropriate*)

OR My thesis cannot be made publicly available online. YES / NO (*Please delete as appropriate*)

3. GRANTING OF NON-EXCLUSIVE RIGHTS

Whether I deposit my Work personally or through an assistant or other agent, I agree to the following:

Rights granted to the University of Warwick and the British Library and the user of the thesis through this agreement are non-exclusive. I retain all rights in the thesis in its present version or future versions. I agree that the institutional repository administrators and the British Library or their agents may, without changing content, digitise and migrate the thesis to any medium or format for the purpose of future preservation and accessibility.

4. DECLARATIONS

(a) I DECLARE THAT:

- I am the author and owner of the copyright in the thesis and/or I have the authority of the authors and owners of the copyright in the thesis to make this agreement. Reproduction of any part of this thesis for teaching or in academic or other forms of publication is subject to the normal limitations on the use of copyrighted materials and to the proper and full acknowledgement of its source.

- The digital version of the thesis I am supplying is the same version as the final, hard-bound copy submitted in completion of my degree, once any minor corrections have been completed.

- I have exercised reasonable care to ensure that the thesis is original, and does not to the best of my knowledge break any UK law or other Intellectual Property Right, or contain any confidential material.

- I understand that, through the medium of the Internet, files will be available to automated agents, and may be searched and copied by, for example, text mining and plagiarism detection software.

(b) IF I HAVE AGREED (in Section 2 above) TO MAKE MY THESIS PUBLICLY AVAILABLE DIGITALLY, I ALSO DECLARE THAT:

- I grant the University of Warwick and the British Library a licence to make available on the Internet the thesis in digitised format through the Institutional Repository and through the British Library via the EThOS service.

- If my thesis does include any substantial subsidiary material owned by third-party copyright holders, I have sought and obtained permission to include it in any version of my thesis available in digital format and that this permission encompasses the rights that I have granted to the University of Warwick and to the British Library.

5. LEGAL INFRINGEMENTS

I understand that neither the University of Warwick nor the British Library have any obligation to take legal action on behalf of myself, or other rights holders, in the event of infringement of intellectual property rights, breach of contract or of any other right, in the thesis.

Please sign this agreement and return it to the Graduate School Office when you submit your thesis.

Student's signature: .....…… Date: .....

THE UNIVERSITY OF
WARWICK

AUTHOR: **Daniel Peavoy** DEGREE: **Ph.D.**

TITLE: **Methods of Likelihood Based Inference for Constructing Stochastic Climate Models.**

DATE OF DEPOSIT: .....

I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: .....

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

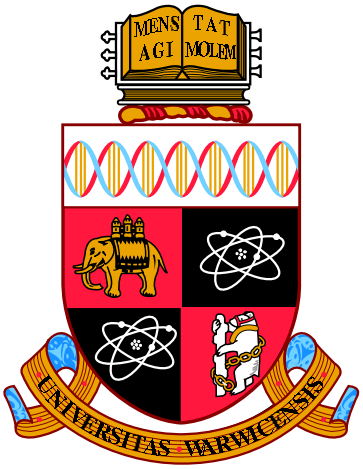
.....

.....

.....

.....

.....



Methods of Likelihood Based Inference for Constructing Stochastic Climate Models.

by

Daniel Peavoy

Thesis

Submitted to the University of Warwick

for the degree of Doctor of Philosophy

Doctor of Philosophy

Centre for Complexity Science

December 2012

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

SDE Stochastic Differential Equation 3

MCMC Markov Chain Monte Carlo 3

ODE Ordinary Differential Equation 7

FP Fokker-Planck 14

PDE Partial Differential Equation 14

LFV Low Frequency Variability 24

PDF Probability Density Function 25

EOF Empirical Orthogonal Function 25

GCM General Circulation Model 26

HMM Hidden Markov Model 26

QG Quasi-Geostrophic 26

OPP Optimal Persistence Pattern 27

PIP Principal Interaction Pattern 27

POP Principal Oscillation Pattern 28, 40

CLT Central Limit Theorem 29

MTV Stochastic Mode Reduction Strategy of Majda et al. [1999], Majda et al. [2001], Majda et al. [2009], Majda et al. [2002], Majda et al. [2003] 35

OU Ornstein-Uhlenbeck 40

ENSO El-Nino Southern Oscillation 40

List of Symbols

$\int_0^t \sigma(X_s, s) \circ dB_s$ Stratonovich integration of function $f(X, s)$ with respect to Brownian motion B_s. 12

$\boldsymbol{\mu}(\boldsymbol{X}_t, t)$ The drift function of a stochastic differential equation for $\boldsymbol{X}$. 12

$\boldsymbol{a}(\boldsymbol{X}_t, t)$ The diffusion function of a stochastic differential equation for $\boldsymbol{X}$. 13

$\boldsymbol{\theta}$ The parameter vector entering the drift and diffusion function. 13

$O(dt)$ The order of a quantity dt. 13

$L^2(\mathbb{P})$ Space of square integrable functions with respect to probability space $(\Omega, \mathcal{F}, \mathbb{P})$. 13

$\boldsymbol{\Sigma}(\boldsymbol{x}, t)$ The covariance matrix of the process $\boldsymbol{x}$ at time t. 14

mb Shorthand for millibar. One thousandth of the unit of atmospheric pressure, the Bar=100,000 pascals. 25

ϵ Small parameter quantifying time scale separation. 29

$\cdot$ Inner product between vectors, $\boldsymbol{a} \cdot \boldsymbol{b} = \sum_i a_i b_i$. 31

: Inner product between matrices, $\boldsymbol{A} : \boldsymbol{B} = \sum_{ij} A_{ij} B_{ij}$. 31

$\otimes$ Outer product between matrices, where $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m \times d} \Rightarrow \boldsymbol{A} \otimes \boldsymbol{B} \in \mathbb{R}^{m \times m}$ satisfies $(\boldsymbol{A} \otimes \boldsymbol{B})\, c = \boldsymbol{A}\boldsymbol{B}^T c$ 34

hPa Hecto-Pascal = 100 Pa. 41

Δt_k Forward observation interval $\Delta t_k = t_{k+1} - t_k$. 62

Δ Constant inter-observation time. 64

m Partition of interval into $m - 1$ sub-intervals. 70

δ Constant sub-interval time $\delta = \Delta / m$. 70

$\boldsymbol{\mu}_k$ Shorthand for the drift function $\boldsymbol{\mu}(\boldsymbol{\xi}_k, \boldsymbol{\theta})$. 70

$\boldsymbol{\Sigma}_k$ Shorthand for the covariance function $\boldsymbol{\Sigma}(\boldsymbol{\xi}_k, \boldsymbol{\theta})$. 70

Acknowledgments

Many thanks to my two supervisors Dr Christian Franzke at the British Antarctic Survey (BAS) and Professor Gareth Roberts of Warwick Statistics. Also I would like to express my appreciation for the training provided by the Complexity Science Doctoral Training Centre and the friendly support of the students, lecturers and administrators. Add to this Warwick's Centre for Scientific Computing, whose facilities I used extensively during this work, particularly the high performance cluster Minerva. Many thanks to the EPSRC for providing the main funding for this research. Thanks also to support from NERC, who provided my travel and accommodation expenses to visit Dr Franzke at BAS. Special thanks to my wife Charlotte for all her moral support and patience during the hectic write up period.

Declarations

The work presented here is my own, except where stated otherwise. This thesis has been composed by myself and has not been submitted for any other degree or professional qualification.

- The majority of computer code for this thesis was written by myself. Examples of this code can be found in the Appendix and all parts can be requested by email. I acknowledge Koev and Edelman [2006] for making available their code for computing the Hypergeometric Function of a Matrix Argument (Written: May 2004). The program also relies upon the GNU scientific library. Additional code was provided by Dr Franzke to simulate the quasi-geostrophic model.

- All data in this thesis were produced by simulations conducted by myself.

- The analytical results in Chapter 3 are known, but all derivations were performed independently by myself.

- The findings of Chapter 8 regarding best practice in stochastic climate modelling including the use of latent noise processes will be submitted for publication.

- The study of improved proposal distributions for missing data in Chapter 5 will be used as supporting methodology in this publication.

Abstract

This thesis is about the construction of low dimensional diffusion models of climate variables. It assesses the predictive skill of models derived from a principled averaging procedure and a purely empirical approach. The averaging procedure starts from the equations for the original system then approximates the "weather" variables by a stochastic process. They are then averaged with respect to their invariant measure. This assumes that they equilibriate much faster than the climate variables. The empirical approach argues for a very general model form, then parameters are estimated using likelihood based inference for Stochastic Differential Equations. This is computationally demanding and relies upon Markov Chain Monte Carlo methods. A large part of this thesis is focused upon techniques to improve the efficiency of these algorithms.

The empirical approach works well on simple one dimensional models but performs poorly on multivariate problems due to the rapid increase in unknown parameters. The averaging procedure is skillful in multivariate problems but is sensitive to lack of complete time scale separation in the system. In conclusion, the averaging procedure is better and can be improved by estimating parameters in a principled way based on the likelihood function and by including a latent noise process in the model.

Chapter 1

Introduction

The Earth's climate is a complex system consisting of several coupled sub-components such as the atmosphere, oceans, biosphere and cryosphere (glaciers, sea ice and snow), which evolve on different time scales. The deterministic equations governing the physics of these systems are derived from the classical laws of mechanics, thermodynamics and fluid flow. In the case of the atmosphere, the dynamics are governed by the non-linear Navier-Stokes equations for compressible flow on a rotating sphere. Together with the equation of state, and conservation of mass and energy, they determine the changes in velocity, temperature, pressure and density, as well as the amount of water vapour in the air. Fundamentally, it is these equations which must be solved to provide weather predictions or climate simulations. However, the Navier-Stokes equations are much too detailed for climate prediction as they resolve processes with length scales $\Delta x = O(10^{-3}m)$ and time scales as small as $\Delta t = O(10^{-1}s)$. They include a range of processes from sound waves, with time scales of milliseconds, to the thermohaline circulation of the ocean.

Whether simulating the full climate system or making short term weather predictions, whatever the time scale of interest, approximations are made, which express the fields of interest as composed of an average component and small high frequency perturbations from this balanced state. The model is simplified by filtering out the high frequency variability. Due to the interaction between different scales in the system the averaged equations are not closed with respect to the high frequency fields. Closed equations are obtained by introducing a parametrisation: a law that specifies the effects of the unresolved processes on the large scale dynamics. Parametrisations could be based on a physical or empirical relation. Examples include the representation of clouds, sub-grid scale turbulence and radiation processes. They are an important component of a climate model and potentially a

major source of error in a simulation.

A parametrisation can be considered a statistical mechanics treatment of the system. Macroscopic quantities arise as the most likely state given the ensemble of microstates which are distributed according to some stationary probability distribution. More usefully one can employ a mesoscopic treatment of the sub-grid scale processes. Then we allow the probability distribution of microstates to evolve in time governed by some PDE. Now our macroscopic quantity obeys a stochastic process akin to a Brownian particle buffeted by invisible fluid particles.

Stochastic parametrisation is a method of including model uncertainty in our predictions. Alternatively, we can consider the initial condition uncertainty due to our imperfect observations of the system. One can consider the initial conditions as random variables with a known probability distribution. This randomness then propagates into the solution. This has implications for predictability since the atmosphere is a highly non-linear chaotic system. Systems evolving from different initial conditions can have diverging solutions and so one simulation may not capture the full range of possible dynamics given our knowledge of the initial state. It is therefore useful to perform an ensemble of simulations with varying initial conditions. This is done routinely at the European Centre for Medium Range Weather Forecasting.

Given the importance of climate science in preparing humanity for a changing future and the role that weather prediction plays in everything from insurance claims to the price of energy it is vital that Earth system research continues to make improvements in the predictive skill of models. Improvements in parametrisations, finer grid resolutions and the use of satellite data to initialise weather simulations have made great progress. However, there is a growing appreciation that, due to the scale invariance of the system, sub grid scale processes will always be important and the error can not be reduced to zero. This motivates further study of stochastic parametrisations and stochastic modelling in climate science with a view towards the probabilistic Earth-System Simulator. At least with this we will have an accurate estimate of our prediction uncertainty.

1.1 Aims of this Thesis

This thesis is about low dimensional stochastic modelling of atmospheric systems. The closure problem discussed above introduces stochastic terms and unknown parameters into these models. This thesis focuses on the problem of statistical estimation of these parameters. Given a functional form for a model of $1-10$ dimensions

we research and develop suitable methodology to estimate the unknown parameters from time series observations of the system. Specifically we work with Stochastic Differential Equations (SDEs).

Although this empirical approach to constructing SDEs for atmospheric processes has been done in the literature in several ways we focus on the difficult statistical problem of likelihood based inference. This is challenging because non-linearity of the models forces an approximation of the likelihood function. Fundamental results show that this approximation converges to the true likelihood as the observation interval goes to zero. However, this is not necessarily obtained by a real data sampling strategy.

In order for the methods to be useful in practice we consider the scenario of infrequent observations. In this case the literature suggests augmenting the data by repeatedly simulating additional points between observations, adopting a Monte-Carlo strategy to integrate over the missing data. Maximum likelihood estimation is difficult in this case due to the noise introduced by the Monte Carlo method leading to a non-convex optimisation. Instead we use a Bayesian approach. We aim to estimate the posterior distribution of the unknown parameters given the observations and the additional uncertainty introduced by the missing data. We now have an integration rather than a maximisation problem.

One of the aims of this thesis is to continue the research into efficient Markov Chain Monte Carlo (MCMC) methods for this problem. Specifically, we investigate the performance of missing data sampling strategies as the dimension of the system increases. We also consider the problem of poor mixing of MCMC due to the dependency between missing data and diffusion parameters. We review the literature of methods proposed to tackle this problem and compare them with other MCMC strategies.

One issue is the computational effort required in the inference of multi-dimensional systems. We aim to implement efficient code written in a low level language such as C/C++. We also assess the performance gains from using massively parallel computation with Graphics Processor Units (GPUs).

One problem encountered early in the research was that the subset of parameter space leading to a stable SDE becomes small as the dimension of the system increases. A lot of the posterior mass was on parameter values which lead to solutions exploding to infinity in finite time. Predictions from these models are obviously not useable. The Bayesian approach is useful because we can include prior information about parameters which restricts them to the subspace leading to stable solutions. An aim of this thesis is to investigate how this prior information can be

included and how it affects the inference strategy.

We aim to assess a current stochastic modelling strategy which predicts the functional form for the reduced model but introduces several parameters which must be estimated from data. We apply our inference algorithm to these unknown parameters. A crucial working assumption of this method is that there exists time scale separation between resolved and unresolved modes of the system. We aim to assess the performance when there is imperfect separation of time scales. To do this we use a series of toy models, starting with those where the time scale separation is explicit and known, then moving to more sophisticated models of geophysical dynamics where time scale separation is an assumption. We aim to use consistent measures of predictive skill to determine the ability of reduced models to reproduce the statistics of the full. Finally we aim to apply our methods to data from a sophisticated atmospheric model.

1.2 Outline of this Thesis

This thesis is broadly divided into methods (Chapters 4 and 5-7) and applications (Chapters 3 and 8) although, firstly, in **Chapter 2** we present some theory of Stochastic Differential Equations (SDEs). We briefly recap some properties of Brownian motion and diffusion processes. We state some useful results regarding the existence and stability of solutions of SDEs, which will be used later to restrict the parameter space through prior information. We also discuss the Girsanov change of measure theorem which is crucial to understanding likelihood based inference for SDEs and the problems that arise. We introduce bridge processes which will be used as part of the inference methods in Chapters 4 and 5.

In **Chapter 3**, we review the existing literature on stochastic modelling in climate science. Here we discuss how the field developed from Hasselmann's seminal work of 1976 and the successful application of these ideas to the understanding of the El Nino system. We also review more recent work on low dimensional modelling of the atmosphere and the methods of statistical inference that have been employed. In this Chapter we also introduce some of the mathematical theory of averaging and homogenisation which underpins the stochastic mode reduction strategy that motivates this thesis. We then present the three toy problems with which we will work, and derive reduced models for each case.

Chapter 4 builds upon the theory of Chapter 2 applied to the inference problem. We discuss the literature, briefly mentioning non-likelihood based approaches and other algorithms that have been developed for our problem. We focus on some

key contributions from the literature regarding likelihood based inference and in particular on the Bayesian approach and Markov Chain Monte Carlo (MCMC) methods. We demonstrate the potential problems encountered with naive algorithms and review more sophisticated methods. We argue for a particular flexible algorithm, taken from the literature, as suitable for our applications and we give details of the implementation.

In **Chapter 5** we focus on improving the efficiency of MCMC methods applied to our particular class of model. We introduce the use of the multivariate non-time homogeneous linear bridge as an efficient method to propose missing data. We discuss sampling diffusion parameters and the difficulty associated with models having low dimensional noise. We present a Gibbs sampler for the drift parameters and investigate the computational improvements gained from using Graphics Processing Units for sampling diffusion parameters.

SDEs driven by red noise are one possibility for modelling systems with lack of time scale separation. This introduces latent, unobserved processes into the SDE model. Inference methods for imputing latent processes are derived in **Chapter 6**.

In **Chapter 7** we discuss the problem of restricting the parameter space in order to obtain stable SDEs and we present one method of solving this problem. From this arises the problem of inference for positive definite matrices. We present several algorithms which tackle this problem, one of which is based on a novel use of the non-central Wishart distribution.

In **Chapter 8** we apply our methods to a range of toy problems which, to varying degree, represent the type of non-linear dynamics with time scale separation one could expect from the atmosphere. We start with double well dynamics, coupled to the chaotic Lorenz system. We compare the cubic models, with parameters inferred empirically, with the theoretically motivated model resulting from the homogenisation procedure. The next step is to consider a bivariate model. For this we consider a triad model coupled to a the Burgers equation. We then move onto a model with more realistic features of atmospheric flow, namely the Quasi-Geostrophic Model on the Beta Plane with mean flow. In this case the time scale separation is not explicitly known. In each case we compare the stationary probability density and autocorrelation functions of the reduced model with the full. We summarise our findings in Chapter 9.

Chapter 2

Stochastic Differential Equations

In this thesis we work extensively with Stochastic Differential Equations (SDEs). In this Chapter we will collect some definitions and results needed to work with SDE models and that will be required for the rest of the thesis. The Chapter is largely based upon the books by Øksendal [2007], Gardiner [2004] and Kloeden and Platen [1992].

SDEs are a widely used modelling framework. They continue to be extensively used in mathematical finance since the seminal work of Black and Scholes [1973] on option pricing and have been used in equilibrium economics as models of interest rates [Cox et al., 1985]. Techniques for fitting nonlinear models have been developed and applied to the Eurodollar exchange rate [Elerian et al., 2001] and stock prices [Bibby and Sorensen, 2001]. Stochastic volatility models have become popular to capture the time dependent noise in stocks; methods for fitting models with latent, unobserved processes have therefore been developed [Eraker, 2001]. Rigorous treatment of topics in mathematical finance, including option pricing and optimal control, is given by Karatzas and Shreve [1997]. The extension to modelling markets with jumps (diffusions with discontinuous paths) is given in Øksendal and Sulem [2007].

In physics, SDEs have been used in the development of non-equilibrium statistical mechanics since the early 20th century. They are used to describe the time dependence of fluctuations in macroscopic quantities such as pressure and energy in a system with an enormous number of variables [van Kampen, 1997]. Einstein derived an equation to describe the old problem of Brownian motion of a particle in a fluid from a probabilistic view while Langevin took an approach based on the mechanics

of individual particles. The resulting Langevin equation (or the Ornstein-Uhlenbeck model [Uhlenbeck and Ornstein, 1930]) has now been generalised for nonlinear models [van Kampen, 1981; Ramshaw, 1985]. Specific applications include molecular dynamics [Feller et al., 1995; Gordon et al., 2009; Pokern et al., 2009; Hegger and Stock, 2009], chemical reaction dynamics [Gillespie, 2000], quantum mechanics [Ford et al., 1988; Olavo et al., 2012], neuron firing [Ota et al., 2009], nuclear fission [Abe et al., 1996] and turbulence [POPE, 1994].

Applications in medicine and biology include modelling gene regulatory networks [Golightly and Wilkinson, 2005, 2008], molecular reaction networks [Sjberg et al., 2005], nonlinear models in epidemiology [Chen and Bokka, 2005], modelling the growth of blood vessels in tumours [Capasso and Morale, 2009] and population genetics [Fearnhead, 2006].

Examples in Earth Sciences include the work of Ditlevsen [1999] on modelling sudden climate change observed in ice core data; modelling drought and flood risk using SDEs [Unami et al., 2010]; stochastic modelling of soil salinity [Suweis et al., 2010]; stochastic parametrisations of unresolved processes in climate models [Wilks, 2008] and hedging climate risk exposure using financial markets [Chaumont et al., 2006]. A lot of work has been done on modelling fast chaotic processes in the atmosphere as noise, resulting in an SDE model for the slow variables (see for example Franzke et al. [2005]). This is closely related to the work in this thesis and the associated literature will be reviewed in Chapter 3.

An SDE is an extension of an Ordinary Differential Equation (ODE) to include a random component. Consider extending a d dimensional ODE system to include a random component as in

$$\frac{d\boldsymbol{X}_t}{dt} = \boldsymbol{\mu}(\boldsymbol{X}_t, t) + \boldsymbol{a}(\boldsymbol{X}_t, t)\boldsymbol{W}_t, \;\; \boldsymbol{X}_0 = \boldsymbol{x}_0, \;\; \boldsymbol{X} \in \mathbb{R}^d \tag{2.1}$$

where $\boldsymbol{\mu} : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^d$, $\boldsymbol{a} : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}^{d\times m}$ and $\boldsymbol{W}_t \in \mathbb{R}^m$ is a standard Gaussian white noise. Solutions to this equation can be written, formally, as

$$\boldsymbol{X}_t = \boldsymbol{x}_0 + \int_0^t \boldsymbol{\mu}(\boldsymbol{X}_s, s)ds + \int_0^t \boldsymbol{a}(\boldsymbol{X}_s, s)\boldsymbol{W}_s ds\,. \tag{2.2}$$

In Section 2.2 we discuss the meaning of the second integral, but we first recall some mathematical preliminaries.

2.1 Some Mathematical Preliminaries

Here we recall some concepts related to random variables and stochastic processes, fixing the notation for the thesis. For further details refer to Øksendal [2007]. If Ω is a set then a σ-algebra $\mathcal{F}$ on Ω is a collection of subsets with the following properties

- $\emptyset \in \mathcal{F}$
- $F \in \mathcal{F} \Rightarrow F^C \in \mathcal{F}$, where F^C is the compliment of F
- $A_1, A_2, \ldots \in \mathcal{F} \Rightarrow A = \cup_{i=1}^{\infty} A_i \in \mathcal{F}$

The pair $(\Omega, \mathcal{F})$ is called a measurable space. A probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathbb{P} : \mathcal{F} \to [0, 1]$ such that

- $\mathbb{P}(\emptyset) = 0$
- $\mathbb{P}(\Omega) = 1$
- If $A_1, A_2, \ldots \in \mathcal{F}$ and $\{A_i\}_{i=1}^{\infty}$ is disjoint then

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)\,.$$

The triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space.

A d dimensional random variable $\mathbf{X}$ is a function from the probability space to the d dimensional real numbers $\boldsymbol{X} : \Omega \to \mathbb{R}^d$. To denote an observation of the random variable we use the informal notation $\boldsymbol{X}(\omega) = \mathbf{x}$, where $\omega \in \Omega$ is an event. For each Borel set $U \subset \mathbb{R}^d$ a random variable induces a probability measure, defined by

$$\pi_X(U) = \mathbb{P}(\boldsymbol{X}^{-1}(U))\,.$$

$\pi_X(U)$ is called the distribution of $\boldsymbol{X}$. The expectation of a function is defined

$$E[f(\boldsymbol{X})] = \int_{\mathbb{R}^d} f(\boldsymbol{x}) d\pi_X(\boldsymbol{x})$$

In this thesis we work with probability measures that have a density $p(\boldsymbol{x})$ with respect to Lebesgue measure so that

$$E[f(\boldsymbol{X})] = \int_{\mathbb{R}^d} f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}\,.$$

A **stochastic process** $\{\mathbf{X}_t\}_{t \in T} \in \mathbb{R}^d$, on probability space $(\Omega, \mathcal{F}, \mathbb{P})$, is a collection of vector valued random variables indexed by time $t \in T$. For each fixed

time t we have an observation function $\mathbf{x}_t : \omega \to \mathbf{X}_t(\omega), \omega \in \Omega$. For fixed $\omega \in \Omega$ we call the function $\mathbf{X}_t : t \to X_t(\omega)$ the path of the stochastic process.

Relevant for later theory in Section 2.5 and the literature review in Chapter 4 are the concepts of a **filtered probability space** and a **martingale**. A filtration, on measurable space $(\Omega, \mathcal{F})$, is an increasing family of σ-algebras $\mathcal{F}_t \subset \mathcal{F}$ so that

$$0 \leq s < t \Rightarrow \mathcal{F}_s \subset \mathcal{F}_t \,.$$

We use the notation $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\})$ to refer to a filtered probability space. A d-dimensional stochastic process $\{\boldsymbol{X}_t\}$ on $(\Omega, \mathcal{F}, \mathbb{P})$ is a martingale with respect to filtration $\mathcal{F}_t$ if

- $\boldsymbol{X}_t$ is $\mathcal{F}_t$-measurable for all t
- $E[|\boldsymbol{X}_t|] < \infty$ for all t
- $E[\boldsymbol{X}_t|\boldsymbol{X}_s] = \boldsymbol{X}_s$ for all $t \geq s$,

where the expectations are taken with respect to $\mathbb{P}$.

Associated with a stochastic process is a transition probability density, defined by

$$\mathbb{P}(\boldsymbol{X}_t \in A|\boldsymbol{X}_s = \boldsymbol{x}_s) = \int_{\boldsymbol{y}\in A} p(t, \boldsymbol{y}|s, \boldsymbol{x}) d\boldsymbol{y} \,.$$

In general a stochastic process can depend upon its full path history so that the transition probability density to be in state $\boldsymbol{x}_1$ at time t_1, is written $p(\boldsymbol{x}_1, t_1|\boldsymbol{x}_2, t_2; \boldsymbol{x}_3, t_3, \cdots)$, where $t_1 > t_2 > t_3 > \cdots$. We say that the process is **Markov** if

$$p(\boldsymbol{x}_1, t_1|\boldsymbol{x}_2, t_2; \boldsymbol{x}_3, t_3, \cdots) = p(\boldsymbol{x}_1, t_1|\boldsymbol{x}_2, t_2) \,, \tag{2.3}$$

i.e the transition density only depends upon the current state. In real systems, observations at fine intervals are likely to depend upon some of the recent history. However, a Markov process may be appropriate on the time scale we are interested in and is a useful modelling framework.

In this thesis we will only consider stochastic processes with continuous sample paths. This excludes models with jumps that are gaining popularity in finance. A sample path is continuous if it satisfies the **Lindeberg condition**

$$\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|\boldsymbol{x}-\boldsymbol{z}|>\epsilon} d\boldsymbol{x} p(\boldsymbol{x}, t + \Delta t|\boldsymbol{z}, t) = 0 \,, \tag{2.4}$$

where $\epsilon > 0$. This states that the probability for $\boldsymbol{x}$ to be finitely different from $\boldsymbol{z}$ goes to 0 faster than Δt.

2.2 Brownian Motion and the Ito Integral

Now that we have introduced notation for stochastic processes we can consider an important example. **Brownian motion** was first proposed as a model for the movement of pollen grains undergoing "random" movements. This has been studied mathematically as a stochastic process and generalised to d dimensions. Here, we recap some of the theoretical properties of mathematical Brownian motion that will allow is to understand and evaluate integrals such as the one in Eq. (2.2). With regards to Eq. (2.2) we introduce the notation $d\boldsymbol{B}_t = \boldsymbol{W}_t dt$. The stochastic process given by

$$\boldsymbol{B}_t = \int_0^t d\boldsymbol{B}_s$$

is known as standard Brownian motion and has the following properties

1. $\boldsymbol{B}_0 = 0$
2. Almost surely continuous paths $\boldsymbol{B}_t$
3. Independent, stationary increments
4. $\boldsymbol{B}_t - \boldsymbol{B}_s \sim \mathcal{N}(0, t-s)\,, 0 \leq s \leq t$

Sometimes referred to as the Wiener process, its existence was proved by Wiener [Wiener et al., 1966]. We state some of the key facts that allow one to integrate with respect to Brownian motion.

The first integral in Eq. (2.2) is to be understood in the usual Riemann-Stieltjes sense. To appreciate why the second integral can not be treated this way consider the one dimensional problem of computing

$$\int_0^1 B_s dB_s\,. \tag{2.5}$$

First define the step function as any function that can be written as

$$f(x) = \sum_{i=0}^{n} \alpha_i I_{A_i}(x)\,,$$

where α_i are real numbers, A_i are intervals and I is the indicator function ($I_{[t_1,t_2]}(t) = 1$ if $t_1 \leq t < t_2, 0$ otherwise). Approximating the integrand in Eq. (2.5) as the step function over the interval $[0,1]$ as

$$B(t,\omega) \approx f_1^{(n)}(t,\omega) = \sum_{j=0}^{2^n-1} B_{j2^{-n}} I_{[j2^{-n},(j+1)2^{-n}]}(t)\,, \tag{2.6}$$

the integral has expected value

$$\mathrm{E}\left[\int_0^1 f_1^{(n)}(s,\omega)dB_s(\omega)\right] = \sum_{j=0}^{2^n-1} \mathrm{E}[B_{j2^{-n}}(B_{(j+1)2^{-n}} - B_{j2^{-n}})] = 0\,.$$

Here we have used the independence of the increments of Brownian motion. Alternatively, if the integrand is approximated as

$$B(t,\omega) \approx f_2^{(n)}(t,\omega) = \sum_{j=0}^{2^n-1} B_{(j+1)2^{-n}} I_{[j2^{-n},(j+1)2^{-n}]}(t)\,, \tag{2.7}$$

then we get

$$\begin{aligned}
\mathrm{E}\left[\int_0^1 f_2^{(n)}(s,\omega)dB_s(\omega)\right] &= \sum_{j=0}^{2^n-1} \mathrm{E}[B_{(j+1)2^{-n}}(B_{(j+1)2^{-n}} - B_{j2^{-n}})] \\
&= \sum_{j=0}^{2^n-1} \mathrm{E}[(B_{(j+1)2^{-n}} - B_{j2^{-n}})^2] \\
&= \sum_{j=0}^{2^n-1} 2^{-n} = 1\,,
\end{aligned}$$

where we have used properties 3 and 4 of standard Brownian motion. This example shows that the integral depends upon which point of the interval $[j2^{-n}, (j+1)2^{-n})$ we choose to approximate the function, unlike the Riemann integral which converges regardless of the point chosen. This phenomenon is due to the large increments of the Brownian motion path; it can be shown that Brownian motion is nowhere differentiable [Øksendal, 2007]. Also, the **total variation** of almost all Brownian motion sample paths over an interval $[s,t]$ is unbounded, i.e

$$\lim_{\Delta t_k \to 0} \sum_{s \le t_k < t} |B_{t_{i+1}} - B_{t_i}| = \infty\,, \tag{2.8}$$

Since the integrator dB_t is not a bounded variation process the Riemann-Stieltjes interpretation of the integral does not necessarily exist [Øksendal, 2007]. However, Brownian motion has finite **quadratic variation**, given by

$$\lim_{\Delta t_i \to 0} \sum_{s \le t_i < t} (B_{t_{i+1}} - B_{t_i})^2 = (t-s) \ \text{ in } L^2\,. \tag{2.9}$$

Therefore, the sums can be shown to converge in L^2 and so the integral is well

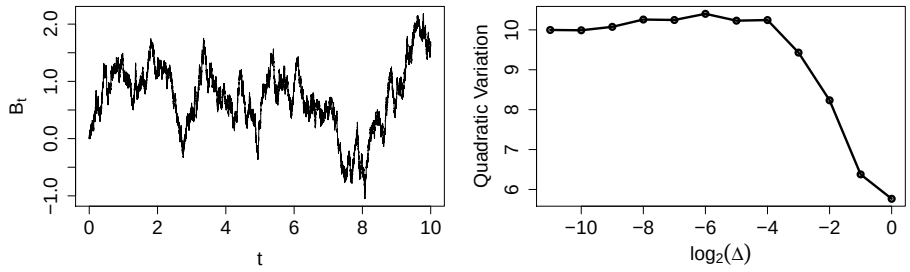


Figure 2.1: Brownian Motion. Sample path of the process (left) and quadratic variation as a function of log time interval (right).

defined even though different for each choice of approximation. Figure 2.1a shows a sample path of Brownian motion over the time interval $[0, 10]$. Figure 2.1b shows the quadratic variation converging to the value in Eq. (2.9) as the discretisation interval goes to zero.

Integration with respect to Brownian motion depends upon the point where the integrand is approximated. Choosing the left point $t_j^* = t_j$, as in Eq. (2.6), leads to the **Ito integral**, denoted

$$\int_0^t \sigma(X_s, s) dB_s = \lim_{n\to\infty} \sum_j f(X_{t_j}, t_j)(B_{j+1} - B_j)\,,$$

Another common choice is to use $t_j^* = (t_{j+1} - t_j)/2$, the mid point of the interval. This is called the **Stratonovich Integral** and is written

$$\int_0^t \sigma(X_s, s) \circ dB_s = \lim_{n\to\infty} \sum_j f(X_{(t_{j+1}-t_j)/2}, (t_{j+1} - t_j)/2)(B_{j+1} - B_j)\,.$$

There are different cases where each integral is more appropriate and there exist relations between the two.

In this thesis we write Eq. (2.1) in the standard notation for SDEs

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, t)dt + \boldsymbol{a}(\boldsymbol{X}_t, t)d\boldsymbol{B}_t \;\;, \boldsymbol{X}_0 = \boldsymbol{x}_0\,, \tag{2.10}$$

where in general the Brownian motion may be of different dimension to $\boldsymbol{X}$, so that $\boldsymbol{X} \in \mathbb{R}^d, \boldsymbol{B} \in \mathbb{R}^m$, $\boldsymbol{\mu} : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^d$ and $\boldsymbol{a} : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^{d\times m}$. It is understood that the second term is integrated in the sense of Ito. Eq. (2.10) sets the notation used throughout the thesis: $\boldsymbol{\mu}(\boldsymbol{X}_t, t)$ is referred to as the **drift**

function and $\boldsymbol{a}(\boldsymbol{X}_t, t)$ as the **diffusion function**. We will usually work with autonomous SDEs, where there is no explicit time dependence in the drift and diffusion functions. It is also often useful to write the drift and diffusion function's dependence on a parameter vector $\boldsymbol{\theta}$ explicitly. Those cases where there is split between the components entering the drift and diffusion functions we write $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\sigma}\}$ and the SDE in Eq. (2.10) is written

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, \boldsymbol{\gamma})dt + \boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\sigma})d\boldsymbol{B}_t \;\; \boldsymbol{X}_0 = \boldsymbol{x}_0 \,. \tag{2.11}$$

In contexts where we want to emphasize a function's dependence upon the underlying probability space we write, for example, $\boldsymbol{a}(t, \omega)$ to mean $\boldsymbol{a}(\boldsymbol{X}_t(\omega), \boldsymbol{\sigma})$.

2.3 Ito's Formula

Ito's formula is the SDE analogue of the chain rule. It is a key tool when working with SDEs and in particular is needed to integrate equations like Eq. (2.11). Ito's formula is a rule for changing variables when working with SDEs. It is used to determine the governing equation for a smooth function $\boldsymbol{f} : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^p$.

Let $\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}_t, t)$ then expanding the differential to second order we have

$$dY_k = \frac{\partial f_k}{\partial t}(\boldsymbol{X}_t, t)dt + \sum_i \frac{\partial f_k}{\partial x_i}(\boldsymbol{X}_t, t)dX_i + \frac{1}{2}\sum_{i,j} \frac{\partial^2 f_k}{\partial x_i \partial x_j}(\boldsymbol{X}_t, t)dX_i dX_j \,. \tag{2.12}$$

In the usual chain rule only the first two terms are of order $O(dt)$. The third term is $O(dt^2)$ and not included. In the case of SDEs, if we substitute the expressions for dX_i from Eq. (2.11) into the third term we get terms of the form $dB_i dB_j$. These are retained as they are of order $O(dt)$. To see this consider property 4 of Brownian motion: the variance of an increment ΔB_j is equal to the time difference Δt_j. One can derive rigorously that

$$\sum_j f(X_j, t_j)(\Delta B_j)^2 \to \int_0^t f(X_s, s)ds \quad \text{in } L^2(\mathbb{P}) \text{ as } \Delta t_j \to 0 \,.$$

and the rules: $dB_i dB_j = \delta_{ij} dt$, $dB_i dt = 0$ (see Øksendal [2007]). Applying these to

calculate the $O(dt)$ terms from $dX_i dX_j$ in Eq. (2.12) gives

$$dY_k = \left(\frac{\partial f_k}{\partial t}(\boldsymbol{X}_t, t)dt + \sum_i \frac{\partial f_k}{\partial x_i}(\boldsymbol{X}_t, t)\mu_i(\boldsymbol{X}_t, t) + \frac{1}{2}\sum_{i,j,l} \frac{\partial^2 f_k}{\partial x_i \partial x_j}(\boldsymbol{X}_t, t)a_{il}a_{jl} \right) dt$$
$$+ \sum_{i,j} \frac{\partial f_k}{\partial x_i}(\boldsymbol{X}_t, t)a_{ij}(\boldsymbol{X}_t, t)dB_j \,. \tag{2.13}$$

Ito's formula can be used to compute integrals such as Eq. (2.5). Using the change of variables $Y_t = B_t^2$ we have

$$dB_s^2 = 2B_s dB_s + \frac{1}{2}2ds \,.$$

Then

$$\int_0^t B_s dB_s = \int_0^t (dB_s^2 - ds) = B_t^2 - t \,, \tag{2.14}$$

so that the integral differs from the equivalent finite variation process by the factor $-t$.

2.4 The Fokker-Planck Equation

The Fokker-Planck (FP) is a Partial Differential Equation (PDE) for the time evolution of the transition density of a SDE (see Gardiner [2004]). In Chapter 3 it is used to derive the reduced dimensional climate model.

Consider SDE of the form Eq. (2.11). If $\boldsymbol{\Sigma}(\boldsymbol{x}, t) = \boldsymbol{a}^T(\boldsymbol{x}, t)\boldsymbol{a}(\boldsymbol{x}, t)$ is the **covariance matrix** of the process then the Fokker-Planck equation for the transition density is

$$\frac{\partial p(\boldsymbol{x}, t|\boldsymbol{z}, t')}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}\left(\mu_i(\boldsymbol{x}, t)p(\boldsymbol{x}, t|\boldsymbol{z}, t')\right) + \frac{1}{2}\sum_{i,j} \frac{\partial^2}{\partial x_i \partial x_j}\left(\Sigma_{ij}(\boldsymbol{x}, t)p(\boldsymbol{x}, t|\boldsymbol{z}, t')\right) \,, \tag{2.15}$$

where we have

$$\mu_i(\boldsymbol{x}, t) = \lim_{\Delta t \to 0} \int_{\mathbb{R}^d} (z_i - x_i)p(\boldsymbol{z}, t + \Delta t|\boldsymbol{x}, t)d\boldsymbol{z} \tag{2.16}$$

and

$$\Sigma_{i,j}(\boldsymbol{x}, t) = \lim_{\Delta t \to 0} \int_{\mathbb{R}^d} (z_i - x_i)(z_j - x_j)p(\boldsymbol{z}, t + \Delta t|\boldsymbol{x}, t)d\boldsymbol{z} \,. \tag{2.17}$$

Solutions of the equation are **diffusion processes**. This equation defines the **drift** and **diffusion** coefficients of the SDE in Eq. (2.11) as the functions in Eqns (2.16)

and (2.17) respectively, connecting the stochastic differential equation and diffusion descriptions of the same system. The Fokker-Planck equation is known in mathematics as the **Forward Kolmogorov** equation.

For insight, and to demonstrate an application of Ito's formula, we provide a derivation of the Fokker-Planck equation for a 1 dimensional process. Using Ito's formula, consider the time evolution of the expectation of an arbitrary twice continuously differentiable function $f(X(t))$, where $X(t)$ is the process in Eq. (2.11) with initial condition $X(0) = y$. Taking expectations with respect to the transition density we have

$$\begin{aligned}
\frac{d\mathbb{E}[f(x(t))]}{dt} &= \int dx f(x) \frac{\partial p(x,t|y,t_0)}{\partial t} \\
&= \mathbb{E}\left[\frac{df(x(t))}{dt}\right] \\
&= \mathbb{E}\left[\mu(x,t)\frac{\partial f}{\partial x} + \frac{1}{2}a(x,t)^2\frac{\partial^2 f}{\partial x^2}\right] \\
&= \int dx \left(\mu(x,t)\frac{\partial f}{\partial x} + \frac{1}{2}a(x,t)^2\frac{\partial^2 f}{\partial x^2}\right) p(x,t|y,t_0) \\
&= \int dx f(x) \left(-\frac{\partial[\mu(x,t)p(x,t|y,t_0)]}{\partial x} + \frac{1}{2}\frac{\partial^2[a(x,t)^2 p(x,t|y,t_0)]}{\partial x^2}\right),
\end{aligned}$$

where we have used integration by parts, discarding surface terms. Since $f(x)$ is arbitrary we have, with the appropriate initial condition $p(x,t_0|y,t_0) = \delta(x-y)$, the Fokker-Planck equation in 1 dimension

$$\frac{\partial p(x,t|y,t_0)}{\partial t} = -\frac{\partial[\mu(x,t)p(x,t|y,t_0)]}{\partial x} + \frac{1}{2}\frac{\partial^2[a(x,t)^2 p(x,t|y,t_0)]}{\partial x^2}. \tag{2.18}$$

It is sometimes easier to work with the **backward Fokker-Planck** equation [Gardiner, 2004]:

$$\frac{\partial p(\boldsymbol{x},t|\boldsymbol{y},t_0)}{\partial t} = -\sum_i \mu_i(\boldsymbol{y},t_0)\frac{\partial[p(\boldsymbol{x},t|\boldsymbol{y},t_0)]}{\partial y_i} - \frac{1}{2}a_{ij}(\boldsymbol{y},t_0)^2\frac{\partial^2[p(\boldsymbol{x},t|\boldsymbol{y},t_0)]}{\partial y_i \partial y_j}. \tag{2.19}$$

The difference is that now the variables $\boldsymbol{x}$ at time t are fixed and $\boldsymbol{y}$ varies. Eq. (2.19) is also referred to as the **backward Kolmogorov equation**. We will use it several times in this thesis.

2.5 Girsanov's Change of Measure Theorem

Intuitively Girsanov's theorem states that if we change the drift of an Ito diffusion then the law of the new process will be absolutely continuous with the law of the original process. Girsanov's theorem is extensively used for option pricing in mathematical finance. It is also central to likelihood based inference for diffusion processes, discussed in detail in Chapter 4.

Let $\boldsymbol{X}_t$ be an Ito process, on filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\})$, of the form

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(t,\omega)dt + d\boldsymbol{B}_t\,. \tag{2.20}$$

Let M_t be given by

$$M_t(\omega) = \exp\left(-\int_0^t \boldsymbol{\mu}(s,\omega)d\boldsymbol{B}_s - \frac{1}{2}\int_0^t \boldsymbol{\mu}^T(s,\omega)\boldsymbol{\mu}(s,\omega)ds\right)\,.$$

The **Novikov condition**

$$\mathrm{E}\left[\exp\left(\frac{1}{2}\int_0^t \boldsymbol{\mu}^T(s,\omega)\boldsymbol{\mu}(s,\omega)ds\right)\right] < \infty\,,$$

is sufficient to guarantee that this is a martingale with respect to the filtration $\mathcal{F}_t$ (see Øksendal [2007]). If $\mathbb{P}$ is the law associated with process (2.20) then the Girsanov transformation gives a new measure on $\mathcal{F}_t$ by

$$d\mathbb{Q}(\omega) = M_t(\omega)d\mathbb{P}(\omega)\,.$$

and the process $\boldsymbol{X}_t$ is a Brownian motion with respect to $\mathbb{Q}$. The theorem implies that for all sets $F_1, \ldots, F_k \subset \mathbb{R}$ and all $t_1, \ldots, t_k \leq t$ we have

$$\mathbb{Q}(\boldsymbol{X}_{t_1} \in F_1, \ldots, \boldsymbol{X}_{t_k} \in F_k) = \mathbb{P}(\boldsymbol{B}_{t_1} \in F_1, \ldots, \boldsymbol{B}_{t_k} \in F_k)\,.$$

Equivalently we can say that $\mathbb{Q}$ is absolutely continuous with respect to $\mathbb{P}$, written $\mathbb{P} \ll \mathbb{Q}$. Then we can write

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = M_t \;\text{ on } \mathcal{F}_t\,.$$

and call M_t the **Radon-Nikodym derivative** of $\mathbb{Q}$ with respect to $\mathbb{P}$.

For the law $\mathbb{P}_a$ of a SDE with general diffusion function

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(t,\omega)dt + \boldsymbol{a}(t,\omega)d\boldsymbol{W}_t\,,$$

the Radon-Nikodym derivative is

$$\begin{aligned}\frac{d\mathbb{Q}_a}{d\mathbb{P}_a}(\omega) &= \exp\left(-\int_0^t \boldsymbol{\mu}^T(s,\omega)\boldsymbol{a}^{-1}(s,\omega)d\boldsymbol{W}_s - \frac{1}{2}\int_0^t \boldsymbol{\mu}^T(s,\omega)\boldsymbol{a}^{-1}(s,\omega)\boldsymbol{\mu}(s,\omega)ds\right) \\ &= \exp\left(-\int_0^t \boldsymbol{\mu}^T(s,\omega)\boldsymbol{a}^{-1}(s,\omega)d\boldsymbol{X}_s + \frac{1}{2}\int_0^t \boldsymbol{\mu}^T(s,\omega)\boldsymbol{a}^{-1}(s,\omega)\boldsymbol{\mu}(s,\omega)ds\right).\end{aligned} \tag{2.21}$$

This ratio serves as the likelihood function for the parameters entering the drift function. However, the Radon-Nikodym derivative between measures induced by diffusions with differing diffusion functions does not exist. This is because they do not have the same sets of measure zero: they are **mutually singular**. As discussed further in Chapter 4 one can not use Eq. (2.21) as the likelihood for inferring parameters in the diffusion function as there is no common dominating measure. Therefore, it is sometimes useful to transform a SDE to one of unit diffusion (so all unknown parameters are in the drift function) before performing inference. This can be done using the **Lamperti Transform**. Consider the one dimensional SDE

$$dX_t = \mu(X,t)dt + a(X_t)dB_t\,,\; X_{t_0} = x_0$$

and let

$$Y_t = g(X_t) = \int^{X_t} \frac{du}{a(u)}\,. \tag{2.22}$$

Then, using Ito's formula, Y_t satisfies the SDE

$$dY_t = \left(\frac{\mu(g^{-1}(Y_t),t)}{a(g^{-1}(Y_t))} - \frac{1}{2}\frac{\partial a}{\partial x}(g^{-1}(Y_t))\right)dt + dB_t\,,\; Y_{t_0} = g(x_0)\,,$$

which has unit diffusion. However, this transformation can not be performed for general multivariate diffusion. The change of variable g is the solution of $\nabla g(x) = a^{-1}(x)$. A solution exists if the inverse of a is a gradient, i.e

$$\frac{\partial[a^{-1}]_{ij}(x)}{\partial x_k} = \frac{\partial[a^{-1}]_{ik}(x)}{\partial x_j} \tag{2.23}$$

for all triples $i,j,k = 1,\ldots,n$, where n is the dimensionality [Ait-Sahalia, 2008].

2.6 Existence, Uniqueness and Stochastic Stability

Here we state some qualitative results about the solution to SDEs such as Eq. (2.11). Although analytic solutions can only be found in a few cases one can determine

global properties such as the stability of the solutions or the existence of a limiting invariant measure for the system. In applications we can decide a priori that our model should have these properties. One of the aims of this thesis is to show how the parameter space can be restricted in order to enforce the stability property. This is achieved through the use of prior information in a Bayesian setting and is demonstrated in Chapter 7. Here we state some definitions and introduce stability criteria for Ito diffusions that will be used later in the thesis.

The existence and uniqueness criteria for SDEs are analogous to those for ODEs. In order for a solution to the SDE (2.11) to exist and be unique within the time interval $[t_0, T]$ it is sufficient for the drift and diffusion functions to satisfy

1. **Lipschitz Condition**: there exists a constant C such that

$$|\boldsymbol{\mu}(\boldsymbol{x}, t) - \boldsymbol{\mu}(\boldsymbol{y}, t)| + |\boldsymbol{a}(\boldsymbol{x}, t) - \boldsymbol{a}(\boldsymbol{y}, t)| \leq C|\boldsymbol{x} - \boldsymbol{y}|$$

   for all x and y and all $t \in [t_0, T]$.

2. **Growth Condition**: there exists a constant K such that for all $t \in [t_0, T]$

$$|\boldsymbol{\mu}(\boldsymbol{x}, t)| + |\boldsymbol{a}(\boldsymbol{x}, t)| \leq K(1 + |\boldsymbol{x}|)\,. \tag{2.24}$$

For time invariant systems the Lipschitz condition implies the growth condition. The above conditions are very restrictive and are often violated in practice, meaning that the solution may explode to infinity. However, these global Lipschitz conditions can be weakened to local ones. Then one can use results from Lyapunov stability theory to ensure sufficient conditions for global existence and uniqueness of solutions. Lyapunov theory for SDEs proceeds similarly to that for ODEs (see Thygesen [1997]). Here we consider only the time homogeneous case.

Associated with Eq. (2.11) we define the differential operator mapping twice differentiable functions of coordinate space $V : X \to \mathbb{R}$ as

$$LV(\boldsymbol{x}) = \frac{\partial V}{\partial \boldsymbol{x}}(\boldsymbol{x}) \cdot \boldsymbol{\mu}(\boldsymbol{x}) + \frac{1}{2}\text{tr } \boldsymbol{a}'(\boldsymbol{x}) \frac{\partial^2 V}{\partial \boldsymbol{x}^2}(\boldsymbol{x}) \boldsymbol{a}(\boldsymbol{x})\,. \tag{2.25}$$

The operator L is known as the **Infinitesimal Generator** associated with SDE (2.11)

$$LV(\boldsymbol{x}) = \lim_{t \downarrow 0} \frac{E^{\boldsymbol{x}}[V(\boldsymbol{X}_t)] - V(\boldsymbol{x})}{t}\,,$$

where the expectation $E^{\boldsymbol{x}}[V(\boldsymbol{X}_t)]$ is for initial condition $\boldsymbol{x}$.

A function $V : \boldsymbol{X} \to \mathbb{R}$ is **proper** if it satisfies

$$a(|\boldsymbol{x}|) \leq V(\boldsymbol{x}) \leq b(|\boldsymbol{x}|)$$

for some strictly increasing functions a and b for which $a(0) = b(0) = 0$ and $a(\boldsymbol{x}) \to \infty$ for $\boldsymbol{x} \to \infty$. A Lyapunov function is a proper continuous, twice differentiable function of $\boldsymbol{X}$. Given that our process obeys local Lipschitz conditions we would like to know what further conditions guarantee that finite escape times occur with probability 0. The following theorem is from Thygesen [1997]

Theorem 1. *Let there exist a proper, twice differentiable function V and numbers $K > 0$, $c > 0$ and $\epsilon \geq 0$ such that for $|\boldsymbol{x}| > K$ we have $LV \leq cV + \epsilon$. Then, with probability 1 the sample paths do not converge to ∞ in finite time.*

2.7 Ergodicity and Stationarity

Intuitively **stationarity** implies that a process has settled to a steady state. We say that a stochastic process $\boldsymbol{X}$ is stationary if $\boldsymbol{X}(t)$ and $\boldsymbol{X}(t + \Delta)$ have the same statistics for all Δ and denote the stationary probability density $p_s(\boldsymbol{x})$. This is equivalent to saying that all joint probability distributions are invariant to time translation

$$p(\boldsymbol{x}_1, t_1; \boldsymbol{x}_2, t_2; \ldots \boldsymbol{x}_n, t_n) = p(\boldsymbol{x}_1, t_1 + \Delta; \boldsymbol{x}_2, t_2 + \Delta; \ldots \boldsymbol{x}_n, t_n + \Delta)\,.$$

Conditional probabilities only depend upon the time difference. In practice we make estimates of the stationary statistics of a process by recording successive measurements in time.

Now consider calculating a time average

$$\bar{\boldsymbol{X}}(T) = \frac{1}{2T} \int_{-T}^{T} dt\, \boldsymbol{X}_t\,. \tag{2.26}$$

Clearly the expectation of this quantity is the average of the process $\boldsymbol{X}_t$. The variance of Eq. (2.26) can be shown to be

$$\mathbb{E}[\bar{\boldsymbol{X}}(T)^2] - \mathbb{E}[\boldsymbol{X}_t]^2 = \frac{1}{4T^2} \int_{-2T}^{2T} dt\, R(t)(2T - t)\,, \tag{2.27}$$

where $R(t)$ is the stationary **autocorrelation function**. A sufficient condition for

the variance to go to zero is

$$\int_0^\infty dt\,|R(t)| < \infty\,.$$

This is satisfied if $R(t) \sim \exp(-t/t_c)$, where t_c is the characteristic decay time. This form of autocorrelation is often met asymptotically in practice and so the variance of $\bar{X}(T)$ goes to zero and we say that the estimate converges in mean square. With similar reasoning one can also show that the autocorrelation function can be measured from time averages.

If we wanted to estimate the stationary distribution $p_s(x)$ from successive times then this is essentially measuring the time average of the indicator function for a grid of intervals (x_1, x_2). This would give an estimate of $\int_{x_1}^{x_2} p_s(x)dx$. We follow the same reasoning as for the time average of the process and find that a sufficient condition is that the limit

$$\lim_{t\to\infty} p(x,t|x_0,t_0) = p_s(x) \tag{2.28}$$

is approached sufficiently rapidly. Under this condition all statistics of the process can be estimated using time averages and we will call the process **ergodic**.

2.8 Some Exact Solutions

Consider the following SDE

$$dX_t = \gamma X_t dt + \sigma X_t dB_t\,. \tag{2.29}$$

The solution of this equation is known as **Geometric Brownian motion** and can be obtained as follows. First make the change of variables $f(x) = \log(X)$ then using Ito's formula

$$\begin{aligned} df_t &= \frac{\partial f}{\partial x}dX_t + \frac{1}{2}\frac{\partial^2 f}{\partial x^2}(dX_t)^2 \\ &= \frac{1}{X}dX_t - \frac{1}{2X^2}(dX_t)^2 \\ &= \frac{1}{X}(\gamma X_t dt + \sigma X_t dB_t) - \frac{1}{2X^2}\sigma^2 X^2 dt \\ &= \left(\gamma - \frac{\sigma^2}{2}\right)dt + \sigma dB_t\,. \end{aligned}$$

This implies that

$$f_t = \left(\gamma - \frac{\sigma^2}{2}\right) t + \sigma B_t \,.$$

Changing back to the original variables gives

$$X_t = \exp\left(\left(\gamma - \frac{\sigma^2}{2}\right) t + \sigma B_t\right) \,. \tag{2.30}$$

Solutions of this equation are shown in Figure 2.2.

Another soluble process, which we shall use, is the **multivariate Ornstein-Uhlenbeck process** [Gardiner, 2004]. It is given by

$$dX_t = -AX_t dt + C dB_t \,, \tag{2.31}$$

with solution

$$X_t = \exp(-At)X_0 + \int_0^t \exp(-A(t-s))C dB_s \,. \tag{2.32}$$

This is a Gaussian process with mean

$$\mathbb{E}[X_t] = \exp(-At)\mathbb{E}[X_0] \tag{2.33}$$

and covariance

$$\begin{aligned}\mathrm{Cov}(X_t, X_s) &= \exp(-At)\mathrm{Var}(X_0)\exp(-As) \\ &\quad + \int_0^{\min(t,s)} \exp(-A(t-r))CC^T \exp(-A^T(s-r))\, dr \,.\end{aligned} \tag{2.34}$$

This matrix can be calculated simply if A and C commute. Otherwise it can be computed component wise.

2.9 The Ito-Taylor Expansion

Taylor expansions are a useful tool for theoretical and numerical studies of smooth deterministic functions. In this section we present the analogous methodology for Ito SDEs, which is based upon the repeated application of the Ito formula. We will use these tools in later sections to both simulate an SDE and construct approximations of its transition density. Consider the time homogeneous SDE

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t)dt + \boldsymbol{a}(\boldsymbol{X}_t)d\boldsymbol{B}_t \,, \boldsymbol{X}_0 = \boldsymbol{x}_0 \,, \tag{2.35}$$

with formal solution

$$\boldsymbol{X}_t = \boldsymbol{X}_0 + \int_0^t \boldsymbol{\mu}(\boldsymbol{X}_s)ds + \int_0^t \boldsymbol{a}(\boldsymbol{X}_s)d\boldsymbol{B}_s\,. \tag{2.36}$$

Applying Ito's formula to the functions $\boldsymbol{\mu}(\boldsymbol{x})$ and $\boldsymbol{a}(\boldsymbol{x})$ gives

$$\begin{aligned}\boldsymbol{X}_t = \boldsymbol{X}_{t_0} &+ \int_{t_0}^t \left[\boldsymbol{\mu}(\boldsymbol{X}_{t_0}) + \int_{t_0}^s \left(\boldsymbol{\mu}^T\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{x}}(\boldsymbol{X}_r) + \frac{1}{2}\boldsymbol{a}^T\boldsymbol{a}\frac{\partial^2\boldsymbol{\mu}}{\partial\boldsymbol{x}^T\partial\boldsymbol{x}}(\boldsymbol{X}_r)\right)dr\right.\\
&\left. + \int_{t_0}^s \boldsymbol{a}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{x}}(\boldsymbol{X}_r)d\boldsymbol{B}_r\right]ds\\
&+ \int_{t_0}^t \left[\boldsymbol{a}(\boldsymbol{X}_{t_0}) + \int_{t_0}^s \left(\boldsymbol{\mu}\frac{\partial\boldsymbol{a}}{\partial\boldsymbol{x}}(\boldsymbol{X}_r) + \frac{1}{2}\boldsymbol{a}^T\boldsymbol{a}\frac{\partial^2\boldsymbol{a}}{\partial\boldsymbol{x}^T\partial\boldsymbol{x}}(\boldsymbol{X}_r)\right)dr\right.\\
&\left. + \int_{t_0}^s \boldsymbol{a}\frac{\partial\boldsymbol{a}}{\partial\boldsymbol{x}}(\boldsymbol{X}_r)d\boldsymbol{B}_r\right]d\boldsymbol{B}_s\,.\end{aligned} \tag{2.37}$$

If we discard the inner integrals we obtain the Euler approximating process

$$\boldsymbol{Y}_t^\delta = \boldsymbol{Y}_{t_0}^\delta + \boldsymbol{\mu}(\boldsymbol{Y}_{t_0}^\delta)(t-t_0) + \boldsymbol{a}(\boldsymbol{Y}_{t_0}^\delta)(\boldsymbol{B}_t - \boldsymbol{B}_{t_0})\,,$$

where $\delta = t - t_0$ is the time step. The difference between this approximation and the exact solution can be quantified as the expectation of the absolute difference at some time t. We say an approximation has **strong, pathwise convergence** of order α if

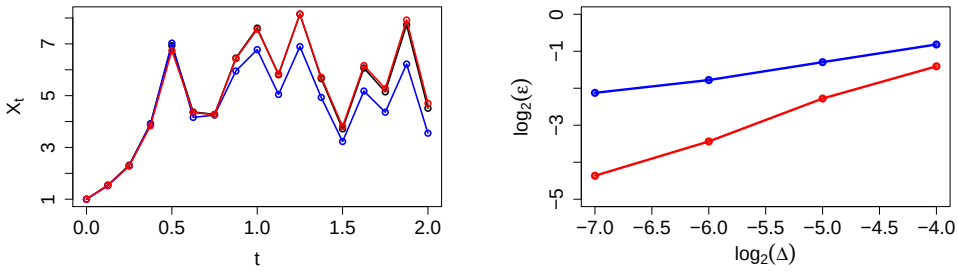
$$\mathbb{E}[|\boldsymbol{X}_t - \boldsymbol{Y}_t^\delta|] \leq C\delta^\alpha$$

and **weak convergence**, with respect to function $g \in \mathcal{G}$, with order β if

$$|\mathbb{E}[g(\boldsymbol{X}_t)] - \mathbb{E}[g(\boldsymbol{Y}_t^\delta)]| \leq C\delta^\beta\,,$$

where δ is the largest step size and C is a constant that does not depend upon δ. The Euler scheme has strong order $\alpha = 0.5$ and weak order $\beta = 1$ [Kloeden and Platen, 1992] essentially because the random increments are of order $\sqrt{\delta}$. If we retain the next order in Eq. (2.37) we have

$$\begin{aligned}\boldsymbol{Y}_t^\delta &= \boldsymbol{Y}_{t_0}^\delta + \boldsymbol{\mu}(\boldsymbol{Y}_{t_0}^\delta)(t-t_0)\\
&\quad + \boldsymbol{a}(\boldsymbol{Y}_{t_0}^\delta)(\boldsymbol{B}_t - \boldsymbol{B}_{t_0}) + \boldsymbol{a}(\boldsymbol{Y}_{t_0}^\delta)\frac{\partial\boldsymbol{a}}{\partial\boldsymbol{x}}(\boldsymbol{Y}_{t_0}^\delta)\int_{t_0}^t\int_{t_0}^s d\boldsymbol{B}_r d\boldsymbol{B}_s\\
&= \boldsymbol{Y}_{t_0}^\delta + \boldsymbol{\mu}(\boldsymbol{Y}_{t_0}^\delta)(t-t_0) + \boldsymbol{a}(\boldsymbol{Y}_{t_0}^\delta)(\boldsymbol{B}_t - \boldsymbol{B}_{t_0})\\
&\quad + \frac{\boldsymbol{a}(\boldsymbol{Y}_{t_0}^\delta)}{2}\frac{\partial a}{\partial\boldsymbol{x}}(\boldsymbol{Y}_{t_0}^\delta)((\boldsymbol{B}_t - \boldsymbol{B}_{t_0})^2 - (t-t_0))\,,\end{aligned} \tag{2.38}$$



(a) Sample path of Eq. (2.29) for parameters $\gamma = 1.5$ and $\sigma = 1.0$ for time $T = 2$ and $\delta = 2^{-6}$. The black line is the exact solution.

(b) Absolute error for the two schemes on a $\log_2$ scale. The Milstein error has a slope approximately 1 while the Euler's is 0.5.

Figure 2.2: Comparison of Euler and Milstein schemes for simulating the SDE in Eq. (2.29). The Euler simulations are in blue and the Milstein in red.

where we have used Eq. (2.14). This is known as the **Milstein** scheme and is of strong order $\alpha = 1$. From this we notice that the approximation will not be Gaussian if the diffusion coefficient has non-zero derivative. In Figure 2.2 we have compared the Euler and Milstein schemes applied to simulating Eq. (2.29). We used parameter values $\gamma = 1.5$ and $\sigma = 1.0$ and simulated for a total time of $T = 2.0$. Figure 2.2a compares a sample path of the exact solution with the Euler and Milstein approximations. The same underlying Brownian motion was used and the exact solution was obtained from Eq. (2.30). Notice that the Milstein scheme appears to stay close to the exact solution for a lot longer than the Euler scheme. The benefit from using the Milstein scheme can be quantified by calculating the absolute error at the final time of the simulation. The absolute error is given by

$$\epsilon = \mathbb{E}[|X_T - Y_T^\delta|]\,,$$

which we estimate using M repeated simulations

$$\hat{\epsilon} = \frac{1}{M}\sum_{i=1}^{M} |X_T^{(i)} - Y_T^{(i),\delta}|\,.$$

Figure 2.2b shows $\log_2(\hat{\epsilon})$ plotted against δ: the simulation time step. The slope of the Milstein error is approximately 1 and the Euler 0.5. Kloeden and Platen [1992] prove that these are the respective strong orders of these algorithms. In this thesis we use the Milstein scheme for simulation. In Chapter 6 we use higher order Ito-Taylor expansions, to propagate noise to all components, for likelihood inference.

Chapter 3

Stochastic Climate Modelling

In this Chapter we review previous work on stochastic modelling in climate science, particularly focussing on Low Frequency Variability (LFV) of the atmosphere. We discuss the various methods of reducing a large non-linear geophysical system to a low dimensional approximation. We review work on flow regimes and large scale persistent patterns, observed in the real atmosphere; the importance of which was highlighted by work on the simple geophysical model of Charney and De Vore [1979]. We give details of model reduction strategies based on the seminal work of Hasselmann [1976]. We discuss methods of multiple time scales: an important approach to understanding stochasticity in slow-fast systems.

Fast chaotic variables in a system can be approximated by a suitable random process. Formally this leads to a perturbation argument: either the theory of averaging or homogenisation. This is an important background for the work in this thesis so we present informal derivations of the procedure following Pavliotis and Stuart [2008]. We review the application of this theory to geophysical problems, particularly the work of Majda and coworkers [Majda et al., 1999, 2001, 2002, 2003]. We analyse the relative success of this approach in comparison to other methods that are more data driven. Here we discuss the approach taken in this thesis, which is a combination of analytical and empirical methods. Finally, we introduce the model problems that we will apply our methods to in Chapter 8.

3.1 Low Frequency Variability

We focus our attention on the type of problems that arise in modelling Low Frequency Variability (LFV) of the atmosphere. This encompasses time scales from 10 days to 6 months and is manifest in large scale patterns in pressure in the tropo-

sphere. This intraseasonal variability is most apparent in the wintertime extratropics (see e.g Pandolfo [1993]). LFV is the time scale at which numerical weather prediction starts to lose significant skill, due to the chaotic dynamics, and has therefore been the focus of much attention for extended range forecasts. Reduced dimensionality models work well at these time scales as LFV systems can often be well represented by just a few modes of variability [D'Andrea and Vautard, 2001; D'andrea, 2002].

A key property for a system to be predictable by a simple stochastic model is its persistence. Many investigations have been conducted into the predictability of persistent planetary flow regimes (e.g Dole and Gordan [1983]). It is essential to model these phenomena in order to understand atmospheric variability beyond the time scale of weather fluctuations. An early example of the identification of a flow regime was the phenomenon of blocking. Blocks are atmospheric pressure fields that are nearly stationary. They often have a region of high pressure known as the blocking high or blocking anticyclone. They can persist for weeks and cause a region to have the same weather for a significant time. Over the Atlantic a simple description of a block is that it is a breakdown in the usual strong westerly flow to a more cellular flow initiating a train of cyclonic and anticyclonic vortices [Rex, 1950]. They are associated with "heatwaves", droughts and severe winters. Significant progress was made by Charney and De Vore [1979] in explaining these modes using a simple barotropic model on the beta plane. They showed that, in this model, multiple equilibria could arise from the interaction of the topographic wave and the zonal mean flow.

Of particular interest for predictability are **teleconnection patterns**. These are correlations in the field of some meteorological variable separated over distance. They were originally defined by determining the correlation between a geographical location and all others and repeating this for each point. The teleconnections are then the centres of strongest negative correlation. Wallace and Gutzler [1981] document the teleconnections for winter 500mb height variability. They find at least four patterns: the North Atlantic and North Pacific Oscillations, a zonally symmetric see-saw in sea level pressure and the Pacific North American Oscillation. According to Barnston and Livezey [1987] this method of finding the most anti-correlated centres does not help in finding the most representative of patterns or their evolution. They suggest using rotated principal component analysis.

Multiple flow regimes have been investigated by Kimoto and Ghil [1993] who fitted bivariate Probability Density Functions (PDFs) to the first two Empirical Orthogonal Functions (EOFs) (principal components of the data) of Northern

Hemisphere wintertime 700 mb height anomalies. They argue that the existence of multiple nodes of the PDF indicates differing flow regimes. However, as stated by Franzke et al. [2008] long run integrations of General Circulation Models (GCMs) show very nearly Gaussian statistics precluding the possibility of multiple regime identification by analysing the PDF. They proceed by stating that multiple regimes could still exist but are unobserved variables. They suggest using Hidden Markov Models (HMMs) to identify regimes. HMMs have a prescribed number of Gaussian mixtures and transition probabilities between them. The eigenvalue structure of the Markov transition matrix is used to determine the existence of metastable regimes. The authors show that the barotropic quasi-geostrophic equation can have a different number of metastable regimes depending on the topographic height. They demonstrate that the nearly Gaussian PDF of the mean flow can be decomposed into the PDFs of three metastable states. They also find regime behaviour in the three level Quasi-Geostrophic (QG) model of Marshall and Molteni [1993] but not in a more realistic GCM. The method presented here is limited to the analysis of univariate time series. Application to multivariate time series would be an interesting extension.

Franzke et al. [2009] extend this work by using the regime identification method of Horenko et al. [2008]; Horenko [2009]. For a given number of clusters this method minimises the observed data from a cluster trajectory. It simultaneously identifies clusters and transitions between them, although they are not generally Markov in nature. Franzke et al. [2009] use the concept of embedding dimension to determine an effective Markov model before determining the metastable states. In this case the regime identification is applied to the full multivariate data set. For the barotropic model on a beta plane this analysis indicated three metastable regimes agreeing with the HMM analysis of Franzke et al. [2008]. They apply the method to the comprehensive National Center of Atmospheric Research (NCAR) Community Climate Model, version 0, which represents well LFV. They focus on a 100 dimensional subspace of the EOFs of 500mb height anomalies. They find that there is evidence for seven regimes, one of which corresponds to the Northern Annular mode. This is in contrast to Franzke et al. [2008] where no regimes were found for this model.

3.2 Model Reduction

To model low frequency dynamics one aims to retain the large scale features while approximating the small fast features, often by some stochastic process. The model

reduction procedure consists of two steps. First an optimal basis to represent the dynamics is chosen and then the system is truncated. Secondly a closure scheme is used to account for the affects of the unresolved variables on the retained modes. This split is often designed to separate the large scale, slow modes from the small, fast modes. The closure procedure could be based on fitting linear stochastic damping terms empirically [Selten, 1995] or predicting stochastic corrections using theory valid in the limit of complete time scale separation [Majda et al., 1999].

Crommelin and Majda [2004] investigated the importance of the choice of basis for the reduced system. Empirical Orthogonal Functions (EOFs) are often used as a basis. EOFs are constructed by finding the mode that accounts for the most variability in the system. Then subsequent modes account for the most variability subject to being orthogonal to the first and so on. EOFs are calculated by computing the eigenvectors of the covariance matrix. They can drastically reduce the number of dimensions in a system while retaining nearly all of the variance (see e.g Preisendorfer [1988]). However, they can fail to reproduce the correct dynamics, even if they account for 99% of the variance. This is particularly true in systems with bursty regime transitions, where low variability modes can be crucial in forcing the system between metastable states [Crommelin and Majda, 2004]. An alternative is to consider Optimal Persistence Patterns (OPPs) [DelSole, 2001]. In this case the basis is chosen to optimise persistence measures: either the integrated autocorrelation function or square integrated autocorrelation. This is a natural basis if one is aiming for long term predictive skill. Another choice are Principal Interaction Patterns (PIPs) [Kwasniok, 1996]. These take account of the dynamics of the system and so are a natural choice, although their calculation is more complicated. Basically, one minimises the integrated difference between the full system and the low dimensional system up to some final time to calculate the projection operator. Expressions for the gradient of the error can be calculated to facilitate the minimisation [Kwasniok, 1997]. A problem with the approach is that the calculation of PIPs can be sensitive to the final time chosen.

Crommelin and Majda [2004] calculated EOFs, OPPs and PIPs for the barotropic model on the beta-plane: the much studied model of Charney and De Vore [1979]. They studied the six dimensional truncated model and assessed the ability of the dimension reduction strategies to reproduce the regime transitting behaviour. They found that, even with five variables, the EOFs were not able to simulate the regime switching. OPPs also failed to produce the chaotic nature of the regime switching. Instead the five dimensional model produced periodic behaviour. The authors conclude that short time scale behaviour must be important to produce

regime switching, which OPPs fail to retain. The PIP models are able to reproduce the regime switching. However, the behaviour of the reduced model was noted to depend upon the final time for the "training". If a short time is used then PIP models can fail to reproduce the climate statistics; a long time and the variability of the reduced system can be too low. These problems with PIPs, applied to a semi-realistic model of the atmosphere, were noted by Kwasniok [2004].

Another related technique are Principal Oscillation Patterns (POPs) [Von Storch et al., 1995]. POPs are the normal modes of the linearised system and correspond to the unstable modes calculated from linear stability analysis. POP analysis includes both stages of the model reduction with the closure problem already solved by the resulting linear system. POP analysis can be used for prediction but the linearity of the model means extended forecasts have little skill.

In this thesis we will focus on the second stage of the model reduction: the closure problem. We focus on the type of problems that apply to LFV. We use the term climate to refer to the resolved modes of the system and occasionally refer to the fast modes as weather variables in agreement with terminology used in the literature.

As mentioned in the Introduction a characteristic of the climate system is its variability on multiple time scales. One way of explaining this variability has been to find external forcing factors driving the system at a range of frequencies such as some unknown solar forcing. In 1976 Hasselmann, with his seminal papers [Hasselmann, 1976; Frankignoul and Hasselmann, 1977], initiated a field of research aiming to explain this variability as part of the internal dynamics of the system. He considered the slow changes in climate to be the integrated response of rapid fluctuations in weather similar to the way a Brownian particle integrates the many collisions with faster moving fluid particles. The idea was to treat the fast deterministic motion as a stochastic process and then average the equations to leave an effective equation for the slow climate variables. This has been the starting point for much research into stochastic climate models. In this chapter we review some of this work but first we introduce some of the mathematical language as summarised by Arnold [2001].

Consider the full description of the climate given by the vector $\boldsymbol{z}$. A climate model starts with a set of deterministic equations

$$\frac{d\boldsymbol{z}}{dt} = \boldsymbol{h}(\boldsymbol{z})\,.$$

Hasselmann considered the case where there exist separate components of $\boldsymbol{z}$

$$\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$$

that evolve on different time scales. In this case $\boldsymbol{x}$ could represent climate variables with characteristic time $\tau_{\boldsymbol{x}}$ and $\boldsymbol{y}$ could be weather variables with characteristic time $\tau_{\boldsymbol{y}}$. In the atmosphere $\tau_{\boldsymbol{y}}$ would be the order of one day and $\tau_{\boldsymbol{x}}$ could be on the scale of weeks to months representing the intraseasonal variability associated with large scale teleconnections. To represent different response times we introduce a small scaling parameter ϵ and write the (non-dimensionalised) system as

$$\begin{aligned}\frac{d\boldsymbol{x}}{dt} &= \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\ , \boldsymbol{x}_0 = \boldsymbol{X} \in \mathbb{R}^n \\ \frac{d\boldsymbol{y}}{dt} &= \frac{1}{\epsilon}\boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})\ , \boldsymbol{y}_0 = \boldsymbol{Y} \in \mathbb{R}^m\end{aligned}$$

such that $\tau_{\boldsymbol{y}} \approx \epsilon \ll \tau_{\boldsymbol{x}} \approx 1$. We would like an approximate equation with solution $\boldsymbol{u}_t \in [0,T]$ such that $\lim_{\epsilon\to 0} \boldsymbol{x}_t^\epsilon = \boldsymbol{u}_t$. The simplest case where $\boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{g}(\boldsymbol{y})$ can be treated by the classical method of Averaging, which Hasselmann refers to as a Statistical Dynamical Model. In this case the forcing term for $\boldsymbol{x}$ is averaged over $\boldsymbol{y}$ giving the approximate equation

$$\frac{d\boldsymbol{u}}{dt} = \boldsymbol{F}(\boldsymbol{u})\,,$$

where

$$\boldsymbol{F}(\boldsymbol{x}) = \lim_{T\to\infty}\frac{1}{T}\int_0^T \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}_t)dt = \int \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\mu(d\boldsymbol{y})\,.$$

Here, μ is the unique invariant measure for $\boldsymbol{y}$ and ergodicity is assumed. Calculating $\boldsymbol{F}(\boldsymbol{x})$ is known as the closure problem. The next step is to consider the error in this approximation. It was shown by Khasminskii [1966] that if the fast variables are a stochastic process then on the interval $t \in [0,T]$, there is a Central Limit Theorem (CLT) such that

$$\boldsymbol{\xi}_t^\epsilon = \frac{1}{\sqrt{\epsilon}}(\boldsymbol{x}_t^\epsilon - \boldsymbol{u}_t)$$

has a limiting Gaussian distribution as $\epsilon \to 0$. Over longer time periods there are many phenomena that are not captured by the Method of Averaging or the Central Limit Theorem. These could include $\boldsymbol{x}_t$ hopping between stable attractors of the system. This could be described as a **Large Deviation** phenomena. The concepts of Averaging (a Law of Large Numbers), the CLT and Large Deviations are three fundamental concepts in asymptotic probability theory. In this thesis we will be focussing on approximations using the CLT.

At this point we have only considered the classical case where $\boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{g}(\boldsymbol{y})$. Arnold [2001] calls the generalisation "Hasselmann's case". Now the slow and fast variables are coupled and the situation is now much more complicated. The fast

dynamics now have invariant measures $\mu_{\boldsymbol{x}}(d\boldsymbol{y})$ which depend upon $\boldsymbol{x}$. If we consider $\boldsymbol{x}$ to be frozen, the solution operator of $\boldsymbol{y}$ maps the initial condition $\boldsymbol{y}_0 = \boldsymbol{Y}$ forward in time: $(t, \boldsymbol{Y}) \to \phi_t^{\boldsymbol{x}}(\boldsymbol{Y})$. Then we can write the averaged forcing as

$$\boldsymbol{F}_{\mu_{\boldsymbol{x}}}(\boldsymbol{x}) = \int_{\mathbb{R}^m} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \mu_{\boldsymbol{x}}(d\boldsymbol{y}) = \lim_{T\to\infty} \frac{1}{T} \int_0^T \boldsymbol{f}(\boldsymbol{x}, \phi_t^{\boldsymbol{x}}(\boldsymbol{Y}))dt\,.$$

Then we have the approximate equation

$$\frac{d\boldsymbol{u}}{dt} = \boldsymbol{F}_{\mu_{\boldsymbol{u}_t}}(\boldsymbol{u}_t)\ , \boldsymbol{u}_0 = \boldsymbol{X}$$

and $\lim_{\epsilon\to 0} \boldsymbol{x}_t^\epsilon(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{u}_t(\boldsymbol{X})$.

Averaging for ODEs is known as **Anosov's theorem** [Pavliotis and Stuart, 2008]. For the results to follow it is sufficient that the fast dynamics are a hyperbolic system [Kifer, 2001]. In this case one can also say something about the deviations from the average system. Kifer [1995] proved that the deviations from the averaged system are a Gaussian diffusion process. The problem is that the ergodicity and fast mixing assumption often fails for ODEs. It is easier to work with an SDE where there is a stochastic term entering into the equation for the fast dynamics as

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}),\ \boldsymbol{x}(0) = \boldsymbol{x}_0, \\ \frac{d\boldsymbol{y}}{dt} &= \frac{1}{\epsilon}\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{\sqrt{\epsilon}}\boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{y})\frac{d\boldsymbol{V}}{dt},\ \boldsymbol{y}(0) = \boldsymbol{y}_0\,, \end{aligned} \tag{3.1}$$

where $\boldsymbol{V}$ is a standard Brownian motion. Given certain conditions on the coefficients $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{y})$ it can be shown that the invariant measures for $\boldsymbol{y}$ have a density with respect to Lebesgue measure, $\mu_{\boldsymbol{x}}(d\boldsymbol{y}) = \rho_{\boldsymbol{x}}(\boldsymbol{y})d\boldsymbol{y}$. In simple cases this density is known explicitly.

3.3 Averaging and Homogenisation for SDEs

In this thesis we will start with the assumption that the fast dynamics are driven by a diffusion process such that the density $\rho_{\boldsymbol{x}}(\boldsymbol{y})$ can be calculated. This follows the stochastic mode reduction procedure of Majda et al. [2001] and related work. Therefore the first step in deriving a stochastic climate model is to approximate the non-linear fast dynamics by a diffusion process. This involves introducing some parameters that are to be determined empirically and motivates the inference problem we study later. For now we assume that we are given the form in Eq. (3.1) and we derive the averaged equation, following Pavliotis and Stuart [2008]. First define the

generators

$$\mathcal{L}_0 = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})\cdot\nabla_{\boldsymbol{y}} + \frac{1}{2}\boldsymbol{C}(\boldsymbol{x},\boldsymbol{y}) : \nabla_{\boldsymbol{y}}\nabla_{\boldsymbol{y}},$$
$$\mathcal{L}_1 = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\cdot\nabla_{\boldsymbol{x}}\,, \tag{3.2}$$

where $\boldsymbol{C}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\beta}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{\beta}(\boldsymbol{x},\boldsymbol{y})^T$ and : denotes the inner product between matrices. The operator $\mathcal{L}_0$ has a null space characterised by

$$\mathcal{L}_0^*\rho(\boldsymbol{y}|\boldsymbol{x}) = 0\,, \tag{3.3}$$

where $\mathcal{L}_0^*$ is the adjoint operator of $\mathcal{L}_0$. We work with the backward Kolmogorov equation (see Section 2.4). For an arbitrary function of state space $h(\boldsymbol{x}(t),\boldsymbol{y}(t))$, define

$$v(\boldsymbol{x},\boldsymbol{y},t) = \mathbb{E}\left(h(\boldsymbol{x}(t),\boldsymbol{y}(t))|\boldsymbol{x}(0)=\boldsymbol{x},\boldsymbol{y}(0)=\boldsymbol{y}\right).$$

Then the backward equation for SDE (3.1) is

$$\frac{\partial v}{\partial t} = \frac{1}{\epsilon}\mathcal{L}_0 v + \mathcal{L}_1 v\,.$$

We seek a multiscale solution to this equation

$$v = v_0 + \epsilon v_1 + O(\epsilon^2)$$

and equating powers of ϵ we get

$$O(1/\epsilon):\ \ \mathcal{L}_0 v_0 = 0, \tag{3.4}$$
$$O(1):\ \ \mathcal{L}_0 v_1 = -\mathcal{L}_1 v_0 + \frac{\partial v_0}{\partial t}\,. \tag{3.5}$$

Eq. (3.4) implies that v_0 is a function only of $(\boldsymbol{x},t)$. The Fredholm alternative (see Pavliotis and Stuart [2008, Theorem 2.42]) implies that

$$-\mathcal{L}_1 v_0 + \frac{\partial v_0}{\partial t}$$

is orthogonal to the null space of $\mathcal{L}_0^*$. Using (3.3) and (3.2) this implies that

$$\int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x})\left(\frac{\partial v_0}{\partial t}(\boldsymbol{x},t) - \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})\cdot\nabla_{\boldsymbol{x}} v_0(\boldsymbol{x},t)\right)d\boldsymbol{y} = 0\,,$$

where $\mathcal{Y}$ is the domain of $\boldsymbol{y}$. Since $\rho(\boldsymbol{y}|\boldsymbol{x})$ is a probability density we have

$$\frac{\partial v_0}{\partial t} - \left(\int_{\boldsymbol{Y}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \rho(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y} \right) \cdot \nabla_{\boldsymbol{x}} v_0(\boldsymbol{x}, t) = 0 \,. \tag{3.6}$$

Defining

$$\boldsymbol{F}(\boldsymbol{x}) = \int_{\boldsymbol{Y}} \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{y}) \rho(\boldsymbol{y}|\boldsymbol{x}) d\boldsymbol{y}$$

we get

$$\frac{\partial v_0}{\partial t} - \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} v_0 = 0 \,.$$

This is the backward equation (see Section 2.4) for

$$\frac{d\boldsymbol{X}}{dt} = \boldsymbol{F}(\boldsymbol{X}), \;\; \boldsymbol{X}(0) = \boldsymbol{x}_0 \,.$$

Therefore, up to times $O(1)$, $\boldsymbol{X}$ approximates the solution of Eq. (3.1).

Averaging can be considered **first order perturbation theory** or as a form of the law of large numbers. **Homogenisation** or **second order perturbation theory** is a form of the central limit theorem [Pavliotis and Stuart, 2008]. The homogenisation procedure describes the dynamics on the longer, diffusive time scale.

For generality we consider second order perturbation theory with three time scales. The linear operator has the form

$$\mathcal{L} = \frac{1}{\epsilon^2} \mathcal{L}_0 + \frac{1}{\epsilon} \mathcal{L}_1 + \mathcal{L}_2 \,. \tag{3.7}$$

Again following Pavliotis and Stuart [2008], consider the SDEs

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt} &= \frac{1}{\epsilon} \boldsymbol{a}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{\alpha}(\boldsymbol{x}, \boldsymbol{y}) \frac{d\boldsymbol{U}}{dt}, \;\; \boldsymbol{x}(0) = \boldsymbol{x}_0, \\ \frac{d\boldsymbol{y}}{dt} &= \frac{1}{\epsilon} \boldsymbol{\omega}(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{\epsilon^2} \boldsymbol{\gamma}(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{\epsilon} \boldsymbol{\beta}(\boldsymbol{x}, \boldsymbol{y}) \frac{d\boldsymbol{V}}{dt}, \;\; \boldsymbol{y}(0) = \boldsymbol{y}_0 \,, \end{aligned} \tag{3.8}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are standard Brownian motions. Then we can define the operators that enter Eq. (3.7) as

$$\begin{aligned} \mathcal{L}_0 &= \gamma \cdot \nabla_{\boldsymbol{y}} + \frac{1}{2} \boldsymbol{C} : \nabla_{\boldsymbol{y}} \nabla_{\boldsymbol{y}}, \\ \mathcal{L}_1 &= \boldsymbol{a} \cdot \nabla_{\boldsymbol{x}} + \boldsymbol{\omega} \cdot \nabla_{\boldsymbol{y}}, \\ \mathcal{L}_2 &= \boldsymbol{b} \cdot \nabla_{\boldsymbol{x}} + \frac{1}{2} \boldsymbol{A} : \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{x}}, \end{aligned} \tag{3.9}$$

where

$$\boldsymbol{A}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\alpha}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{\alpha}(\boldsymbol{x},\boldsymbol{y})^T,$$
$$\boldsymbol{C}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{\beta}(\boldsymbol{x},\boldsymbol{y})\boldsymbol{\beta}(\boldsymbol{x},\boldsymbol{y})^T\,.$$

Again the generator $\mathcal{L}_0$ is a differential operator in $\boldsymbol{y}$, where $\boldsymbol{x}$ enters as a parameter. has null space as in Eq. (3.3). We also assume that $\boldsymbol{a}(\boldsymbol{x},\boldsymbol{y})$ averages to zero under this measure. This is known as the centring condition

$$\int_Y a(x,y)\rho(y|x)dy = 0,\ \forall x \in X\,. \tag{3.10}$$

We seek a multiscale solution

$$v = v_0 + \epsilon v_1 + \epsilon^2 v_2 + \cdots$$

of

$$\frac{\partial v}{\partial t} = \left(\frac{1}{\epsilon^2}\mathcal{L}_0 + \frac{1}{\epsilon}\mathcal{L}_1 + \mathcal{L}_2\right) v,\ \text{for } (\boldsymbol{x},\boldsymbol{y},t) \in \mathcal{X}\times\mathcal{Y}\times\mathbb{R}^+\,.$$

Equating powers of ϵ we have

$$O(1/\epsilon^2):\ -\mathcal{L}_0 v_0 = 0 \tag{3.11}$$
$$O(1/\epsilon):\ -\mathcal{L}_0 v_1 = \mathcal{L}_1 v_0 \tag{3.12}$$
$$O(1):\ -\mathcal{L}_0 v_2 = -\frac{\partial v_0}{\partial t} + \mathcal{L}_1 v_1 + \mathcal{L}_2 v_0\,. \tag{3.13}$$

Eq. (3.11) implies that $v_0 = v_0(\boldsymbol{x},t)$. Solvability for Eq. (3.12) requires expectation zero with respect to the invariant measure on $\boldsymbol{y}$. Eq. (3.12) can be written

$$-\mathcal{L}_0 v_1 = \boldsymbol{a}(\boldsymbol{x},\boldsymbol{y})\cdot\nabla_{\boldsymbol{x}} v_0(\boldsymbol{x},t)\,. \tag{3.14}$$

Using Eq. (3.9) the general solution of Eq. (3.14) has the form

$$v_1(\boldsymbol{x},\boldsymbol{y},t) = \boldsymbol{\Phi}(\boldsymbol{x},\boldsymbol{y})\cdot\nabla_{\boldsymbol{x}} v_0(\boldsymbol{x},t) + \boldsymbol{\Phi}_1(\boldsymbol{x},t)\,.$$

The function $\boldsymbol{\Phi}_1$ plays no further role so it can be set to zero. Substituting $\boldsymbol{\Phi}$ into Eq. (3.14) we see that it solves the so called **cell problem**

$$-\mathcal{L}_0\boldsymbol{\Phi}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{a}(\boldsymbol{x},\boldsymbol{y}),\quad \int_{\mathcal{Y}} \boldsymbol{\Phi}(\boldsymbol{x},\boldsymbol{y})\rho(\boldsymbol{y}|\boldsymbol{x})d\boldsymbol{y} = 0, \tag{3.15}$$

which, by the Fredholm alternative for elliptic PDEs, has a solution if the centring

condition (Eq. (3.10)) holds. Substituting the expression for v_1 into Eq. (3.13) we have

$$-\mathcal{L}_0 v_2 = -\frac{\partial v_0}{\partial t} + \mathcal{L}_1 \left(\boldsymbol{\Phi} \cdot \nabla_{\boldsymbol{x}} v_0 \right) + \mathcal{L}_2 v_0 \,.$$

For solvability we need the right hand side to be in the null space of $\mathcal{L}_0$. Analogous to Eq. (3.6) this requires

$$\frac{\partial v_0}{\partial t} = \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \mathcal{L}_2 v_0(\boldsymbol{x}, t) dy + \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \mathcal{L}_1 \left(\boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla_{\boldsymbol{x}} v_0(\boldsymbol{x}, t) \right) d\boldsymbol{y} \,. \tag{3.16}$$

As in Pavliotis and Stuart [2008] we split the problem into separate integrals. Firstly, let

$$\begin{aligned} I_1 &= \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \mathcal{L}_2 v_0(\boldsymbol{x}, t) dy \\ &= \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \left(\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla_{\boldsymbol{x}} + \frac{1}{2} \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{y}) : \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} \right) v_0(\boldsymbol{x}, t) d\boldsymbol{y} \\ &= \boldsymbol{F}_1(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} v_0(\boldsymbol{x}, t) + \frac{1}{2} \boldsymbol{A}_1(\boldsymbol{x}) : \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} v_0(\boldsymbol{x}, t) \,, \end{aligned} \tag{3.17}$$

where

$$\boldsymbol{F}_1(\boldsymbol{x}) = \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \text{ and } \boldsymbol{A}_1(\boldsymbol{x}) = \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{A}(\boldsymbol{x}, \boldsymbol{y}) d\boldsymbol{y} \,.$$

Also

$$\begin{aligned} I_2 &= \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \mathcal{L}_1 \left(\Phi(\boldsymbol{x}, \boldsymbol{y}) \cdot \nabla_{\boldsymbol{x}} v_0(\boldsymbol{x}, t) \right) d\boldsymbol{y} \\ &= \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \left(\boldsymbol{a} \otimes \Phi : \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} + (\boldsymbol{a} \nabla_{\boldsymbol{x}} \Phi + \boldsymbol{\omega} \nabla_{\boldsymbol{y}} \boldsymbol{\Phi}) \cdot \nabla_{\boldsymbol{x}} \right) v_0(\boldsymbol{x}, t) d\boldsymbol{y} \\ &= \left(\boldsymbol{F}_0(\boldsymbol{x}) + \boldsymbol{G}_0(\boldsymbol{x}) \right) \cdot \nabla_{\boldsymbol{x}} v_0(\boldsymbol{x}, t) + \frac{1}{2} \boldsymbol{A}_0(\boldsymbol{x}) : \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} v_0(\boldsymbol{x}, t) \,, \end{aligned}$$

where

$$\boldsymbol{F}_0(\boldsymbol{x}) = \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{a}(\boldsymbol{x}, \boldsymbol{y}) \nabla_{\boldsymbol{x}} \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{y}) \, d\boldsymbol{y}, \;\; \boldsymbol{G}_0(\boldsymbol{x}) = \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{\omega}(\boldsymbol{x}, \boldsymbol{y}) \nabla_{\boldsymbol{y}} \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{y}) \, d\boldsymbol{y}$$

and

$$\boldsymbol{A}_0(\boldsymbol{x}) = 2 \int_{\mathcal{Y}} \rho(\boldsymbol{y}|\boldsymbol{x}) \boldsymbol{a}(\boldsymbol{x}, \boldsymbol{y}) \otimes \boldsymbol{\Phi}(\boldsymbol{x}, \boldsymbol{y}) \,.$$

Combining this with Eq. (3.17) gives

$$\frac{\partial v_0}{\partial t} = \boldsymbol{F}(\boldsymbol{x}) \cdot \nabla_{\boldsymbol{x}} v_0 + \frac{1}{2} \boldsymbol{A}(\boldsymbol{x}) \boldsymbol{A}(\boldsymbol{x})^T : \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} v_0 \,, \tag{3.18}$$

where

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{x}) &= \boldsymbol{F}_1(\boldsymbol{x}) + \boldsymbol{F}_0(\boldsymbol{x}) + \boldsymbol{G}_0(\boldsymbol{x}) \\ \boldsymbol{A}(\boldsymbol{x})\boldsymbol{A}(\boldsymbol{x})^T &= \boldsymbol{A}_1(\boldsymbol{x}) + \boldsymbol{A}_0(\boldsymbol{x})\,. \end{aligned} \tag{3.19}$$

Note that it can be shown that $\boldsymbol{A}(\boldsymbol{x})\boldsymbol{A}(\boldsymbol{x})^T, \boldsymbol{x} \in \mathcal{X}$ is positive definite [Pavliotis and Stuart, 2008]. Eq. (3.18) is the backward equation for the SDE

$$\frac{d\boldsymbol{X}}{dt} = \boldsymbol{F}(\boldsymbol{X}) + \boldsymbol{A}(\boldsymbol{X})\frac{d\boldsymbol{B}}{dt},\ \boldsymbol{X}(0) = \boldsymbol{x}_0\,.$$

Solutions to this equation then approximate solutions to Eq. (3.8) for times t of order $O(1)$. Note that this approximation is in the sense of measures.

3.3.1 Averaging and Homogenisation for Climate Modelling

The above theory has been applied to climate modelling by Andrew Majda and coworkers in a series of papers: Majda et al. [1999], Majda et al. [2001], Majda et al. [2009], Majda et al. [2002], Majda et al. [2003]. The authors refer to this **Stochastic Mode Reduction Strategy** as the MTV procedure and it is presented for the general case in Majda et al. [2009]. It is applied to simple toy models with explicit time scale separation in Majda et al. [1999] and Majda et al. [2002] and demonstrated on a simplified model of atmospheric flow in Majda et al. [2003]. Franzke et al. [2005] applied the method to a realistic barotropic model of climate.

The MTV procedure considers the case where we have a climate model, for state variable $\boldsymbol{z} \in \mathbb{R}^d$, of the following form

$$\frac{d\boldsymbol{z}}{dt} = \boldsymbol{f}(t) + \boldsymbol{L}\boldsymbol{z} + \boldsymbol{Q}(\boldsymbol{z}, \boldsymbol{z})\,, \tag{3.20}$$

so that we have a linear operator $\boldsymbol{L}$ and a quadratic operator $\boldsymbol{Q}$. As in Hasselmann [1976] we assume that there are two subsets of variables $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$ so that the resolved variables $\boldsymbol{x} \in \mathbb{R}^n$ evolve on the slow time scale and the unresolved variables $\boldsymbol{y} \in \mathbb{R}^m$ on the fast. Then we can write the model as

$$\begin{aligned}\frac{d\boldsymbol{x}}{dt} &= \boldsymbol{f}_1(t) + \boldsymbol{L}_{11}\boldsymbol{x} + \boldsymbol{L}_{12}\boldsymbol{y} + \boldsymbol{Q}^1_{11}(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{Q}^1_{12}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{Q}^1_{22}(\boldsymbol{y}, \boldsymbol{y}) \\ \frac{d\boldsymbol{y}}{dt} &= \boldsymbol{f}_2(t) + \boldsymbol{L}_{21}\boldsymbol{x} + \boldsymbol{L}_{22}\boldsymbol{y} + \boldsymbol{Q}^2_{11}(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{Q}^2_{12}(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{Q}^2_{22}(\boldsymbol{y}, \boldsymbol{y})\,. \end{aligned} \tag{3.21}$$

The authors of MTV then make the assumption that the non-linear self interaction of the unresolved variables $\boldsymbol{y}$ can be represented by an ergodic stochastic process.

As mentioned above this means that we can work with the better understood homogenisation theory for ergodic stochastic processes rather than that for deterministic systems. This stochastic approximation is

$$\boldsymbol{Q}_{22}^2(\boldsymbol{y},\boldsymbol{y}) \approx -\frac{\boldsymbol{\Gamma}}{\epsilon} + \frac{\boldsymbol{\Sigma}}{\sqrt{\epsilon}}\dot{\boldsymbol{B}}(t)\,,$$

where $\boldsymbol{\Gamma}, \boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$ are matrices and $\boldsymbol{B}$ is a m-dimensional Brownian motion. The authors assume that $\boldsymbol{\Gamma}$ and $\boldsymbol{\Sigma}$ are diagonal but this could be generalised. The parameter $\epsilon \ll 1$ is introduced here to represent our assumption that these terms are large and force the $\boldsymbol{y}$ variables to equilibriate quickly. If we coarse grain time in equations (3.21) so that $t \to \epsilon t$ we have

$$\begin{aligned} d\boldsymbol{x} &= \frac{1}{\epsilon}\left(\boldsymbol{f}_1\left(\frac{t}{\epsilon}\right) + \boldsymbol{L}_{11}\boldsymbol{x} + \boldsymbol{L}_{12}\boldsymbol{y} + \boldsymbol{Q}_{11}^1(\boldsymbol{x},\boldsymbol{x}) + \boldsymbol{Q}_{12}^1(\boldsymbol{x},\boldsymbol{y}) + \boldsymbol{Q}_{12}^1(\boldsymbol{y},\boldsymbol{y})\right)dt \\ d\boldsymbol{y} &= \frac{1}{\epsilon}\left(\boldsymbol{f}_2\left(\frac{t}{\epsilon}\right) + \boldsymbol{L}_{21}\boldsymbol{x} + \boldsymbol{L}_{22}\boldsymbol{y} + \boldsymbol{Q}_{11}^2(\boldsymbol{x},\boldsymbol{x}) + \boldsymbol{Q}_{12}^2(\boldsymbol{x},\boldsymbol{y})\right)dt - \frac{\boldsymbol{\Gamma}}{\epsilon^2}\boldsymbol{y}dt + \frac{\boldsymbol{\Sigma}}{\epsilon}d\boldsymbol{B}(t)\,. \end{aligned} \tag{3.22}$$

In order to derive a model for the climate variables we follow some assumptions of Majda et al. [2001]. We assume that there is damping on the climate time scale, i.e we add a term $-Dx$, and that the external forcing acts on the climate time scale so that

$$\frac{1}{\epsilon}\boldsymbol{f}_1\left(\frac{t}{\epsilon}\right) \to \boldsymbol{F}_1(t)\,.$$

We assume that the non-linear interaction of the climate variables is a slow time scale effect: $\boldsymbol{Q}_{11}^1(\boldsymbol{x},\boldsymbol{x})/\epsilon \to \boldsymbol{Q}_{11}^1(\boldsymbol{x},\boldsymbol{x})$. The operators $\boldsymbol{Q}_{22}^1, \boldsymbol{Q}_{11}^1, \boldsymbol{Q}_{11}^2$ are symmetric in their arguments. Finally we assume that the non-linear interactions in $\boldsymbol{y}$ have expectation zero with respect to the invariant measure of the fast process: $\mathbb{E}Q_{22}^1(y,y) = 0$. In particular, this will be the case if the diagonal terms are zero. Also for the derivation here we assume that there are no fast wave effects, i.e $\boldsymbol{L}_{11} = 0$. In terms of components our model is now written

$$
\begin{aligned}
dx_j &= F_j(t) + \sum_k \left(-D_{jk} x_k + \frac{1}{\epsilon} L^{12}_{jk} y_k \right) dt \\
&\quad + \sum_{k,l} \left(\frac{1}{2} Q^{111}_{jkl} x_k x_l + \frac{1}{\epsilon} Q^{112}_{jkl} x_k y_l dt + \frac{1}{2\epsilon} Q^{122}_{jkl} y_k y_l \right) dt \\
dy_j &= \sum_k \left(\frac{1}{\epsilon} L^{21}_{jk} x_k + \frac{1}{\epsilon} L^{22}_{jk} y_k \right) dt \\
&\quad + \sum_{k,l} \left(\frac{1}{2\epsilon} Q^{211}_{jkl} x_k x_l + \frac{1}{\epsilon} Q^{221}_{jkl} y_k x_l \right) dt - \frac{\gamma_j}{\epsilon^2} y_j + \frac{\sigma_j}{\epsilon} dB_j(t) \,.
\end{aligned}
$$

We can now define the operators as in Eq. (3.9):

$$
\begin{aligned}
\mathcal{L}_0 &= \sum_j \left(-\gamma_j y_j \frac{\partial}{\partial y_j} + \frac{\sigma_j^2}{2} \frac{\partial^2}{\partial y_j^2} \right) \\
\mathcal{L}_1 &= \sum_{j,k} \left(L^{12}_{jk} y_k + \frac{1}{2} \sum_l (2 Q^{112}_{jkl} x_k y_l + Q^{122}_{jkl} y_k y_l) \right) \frac{\partial}{\partial x_j} \\
&\quad + \sum_{j,k} \left(L^{21}_{jk} x_k + L^{22}_{jk} y_k + \frac{1}{2} \sum_l (Q^{211}_{jkl} x_k x_l + 2 Q^{221}_{jkl} y_k x_l) \right) \frac{\partial}{\partial y_j} \\
\mathcal{L}_2 &= \sum_j \left(F_j(s) - \sum_k D_{jk} x_k + \frac{1}{2} \sum_{k,l} Q^{111}_{jkl} x_k x_l \right) \frac{\partial}{\partial x_j} \,.
\end{aligned}
$$

Using the notation of Eq. (3.9) these operators imply

$$
\begin{aligned}
\lambda &= -\sum_j \gamma_j y_j \\
Q_{jj} &= \sigma_j^2 \\
a_j &= \sum_k \left(L^{12}_{jk} y_k + \frac{1}{2} \sum_l (2 Q^{112}_{jkl} x_k y_l + Q^{122}_{jkl} y_k y_l) \right) \\
\omega_j &= \sum_k \left(L^{21}_{jk} x_k + L^{22}_{jk} y_k + \frac{1}{2} \sum_l (Q^{211}_{jkl} x_k x_l + 2 Q^{221}_{jkl} y_k x_l) \right) \\
b_j &= F_j(s) - \sum_k D_{jk} x_k + \frac{1}{2} \sum_{k,l} Q^{111}_{jkl} x_k x_l \,.
\end{aligned}
$$

Firstly we find the solution to the cell problem, which is equivalent to inverting the

Ornstein-Uhlenbeck operator:

$$\sum_j \gamma_j y_j \frac{\partial}{\partial y_j}\Phi_i(\boldsymbol{x},\boldsymbol{y}) - \frac{1}{2}\sum_j \sigma_j^2 \frac{\partial^2}{\partial y_j^2}\Phi_i(\boldsymbol{x},\boldsymbol{y}) = \sum_j \left(L_{ij}^{12} y_j + \frac{1}{2}\sum_k Q_{ijk}^{122} y_j y_k \right) \tag{3.23}$$

$$+ \sum_j \sum_k Q_{ijk}^{112} x_j y_k \,. \tag{3.24}$$

We consider solutions of the form

$$\Phi_i(\boldsymbol{x},\boldsymbol{y}) = \sum_k A_{ik} y_k + \frac{1}{2}\sum_{k,l} C_{ikl} y_k y_l, \quad C_{ikl} = C_{ilk} \,.$$

Substituting into Eq. (3.24) gives

$$A_{ij} = (L_{ij}^{12} + \sum_k Q_{ikj}^{112} x_k)/\gamma_j, \quad C_{ijl} = \frac{Q_{ijl}^{122}}{\gamma_j + \gamma_l} \,.$$

By assumption $Q_{ijj} = 0$ and so $\boldsymbol{\Phi}$ satisfies the normalisation condition and is in fact a unique solution to Eq. (3.15). We now compute the integrals in Eq. (3.16) with respect to the stationary measure of the fast variables

$$\rho^\infty(\boldsymbol{y}|\boldsymbol{x}) = \prod_j \mathcal{N}\left(0, \frac{\sigma_j^2}{2\gamma_j}\right) \,. \tag{3.25}$$

Using the notation from Eq. (3.18)

$$
\begin{aligned}
\boldsymbol{F}_1(\boldsymbol{x}) &= F_j(s) - \sum_k D_{jk}x_k + \frac{1}{2}\sum_{k,l} Q^{111}_{jkl}x_k x_l \\
\boldsymbol{F}_0(\boldsymbol{x}) &= \sum_j \left(\sum_k \frac{Q^{112}_{ijk}L^{12}_{jk}\sigma_k^2}{2\gamma_k^2} + \sum_{l,k} \frac{Q^{112}_{ijl}Q^{112}_{jkl}x_k\sigma_l^2}{2\gamma_l^2} \right) \\
\boldsymbol{G}_0(\boldsymbol{x}) &= \sum_j \left(\frac{L^{12}_{ij}}{\gamma_j} \left(\sum_k L^{21}_{jk}x_k + \frac{1}{2}\sum_{k,l} Q^{211}_{jkl}x_k x_l \right) + \sum_{l,k} \frac{Q^{112}_{ilj}L^{21}_{jk}x_l x_k}{\gamma_j} \right. \\
&\quad \left. + \sum_{l,k,m} \frac{Q^{112}_{ijl}}{\gamma_j+\gamma_l} Q^{211}_{jkm}x_l x_k x_m + \sum_l \frac{Q^{122}_{ijl}}{\gamma_j+\gamma_l} \left(\frac{L^{22}_{jl}\sigma_l^2}{2\gamma_l} + \sum_m Q^{221}_{jlm}x_m \frac{\sigma_l^2}{2\gamma_l} \right) \right) \\
\boldsymbol{A}_0(\boldsymbol{x}) &= \sum_k L^{12}_{ik} \left(L^{12}_{jk} + \sum_n Q^{112}_{jnk}x_n \right) \frac{\sigma_k^2}{2\gamma_k^2} + \sum_{k,l} Q^{112}_{ikl}x_k \left(L^{12}_{jl} + \sum_n Q^{112}_{jnm}x_n \right) \frac{\sigma_l^2}{2\gamma_l^2} \\
&\quad + \frac{1}{4}\sum_{k,l} \frac{Q^{122}_{jkl}Q^{122}_{jkl}}{\gamma_k+\gamma_l} \frac{\sigma_k^2\sigma_l^2}{\gamma_k\gamma_l} \,.
\end{aligned}
$$

The Fokker-Planck equation (3.18) then follows, which is equivalent to the following system of SDEs

$$
\begin{aligned}
dx_j &= F_j(t)dt - \sum_{k\in\sigma_1} D_{jk}x_k dt - \frac{1}{2}\sum_{k,l\in\sigma_1} Q^{111}_{jkl}x_k x_l dt \\
&\quad + a_j dt - \sum_{k\in\sigma_1} \gamma_{jk}x_k dt + \sum_{k,j\in\sigma_2} \sigma_{jkl} dB_{kl}(t) \\
&\quad + \frac{1}{2}\sum_{k\in\sigma_1}\sum_{m\in\sigma_2} \frac{\sigma_m^2}{\gamma_m^2} Q^{112}_{jkm} \left(L^{12}_{km} + \sum_{l\in\sigma_1} Q^{112}_{klm}x_l \right) dt \\
&\quad + \sum_{l\in\sigma_1}\sum_{n\in\sigma_2} \frac{1}{\gamma_n} \left(L^{12}_{jn} + \sum_{k\in\sigma_2} Q^{112}_{jkn}x_k \right) \left(L^{21}_{nl}x_l + \frac{1}{2}\sum_{m\in\sigma_1} Q^{211}_{nlm}x_l x_m \right) dt \\
&\quad + \sum_{l\in\sigma_2} \frac{\sigma_l}{\gamma_l} \left(L^{12}_{jl} + \sum_{k\in\sigma_1} Q^{112}_{jkl}x_k \right) dB_l(t) \,,
\end{aligned} \tag{3.26}
$$

where B_j, B_{jk} are independent Brownian motions satisfying $EB_j(t)B_k(s) = \delta_{jk}\min(t,s)$, $EB_{jk}(t)B_{mn}(s) = \delta_{jm}\delta_{kn}\min(t,s)$ and where we have defined

$$
a_j = \frac{1}{2}\sum_{k,l\in\sigma_2} \frac{\sigma_l^2 Q^{122}_{jkl}L^{22}_{kl}}{\gamma_l(\gamma_k+\gamma_l)}, \quad \gamma_{jk} = -\frac{1}{2}\sum_{l,m\in\sigma_2} \frac{\sigma_l^2 Q^{122}_{jlm}Q^{221}_{mlk}}{\gamma_l(\gamma_l+\gamma_m)} \,, \tag{3.27}
$$

$$\sigma_{jkl} = \frac{Q^{122}_{jkl}\sigma_k\sigma_l}{2\sqrt{(\gamma_k+\gamma_l)\gamma_k\gamma_l}}\,. \tag{3.28}$$

This result, obtained by a different method, is in agreement with that of Majda et al. [2001].

3.4 Empirical Methods to Model Reduction

The MTV strategy described above relies upon the approximation of nonlinear terms by a stochastic process. This inevitably introduces some unknown parameters. In practice these are estimated from observations of very long runs of the full model. Generally there are one or two unknown parameters for each degree of freedom in the system. This would not be possible if trying to produce a model from observations of the real atmosphere. In this case only observations of the variables of interest may be available: an empirically derived model may work as well. In this section we discuss some data driven methods to producing low dimensional models of the atmosphere and ocean.

Penland [1996] used the centred Ornstein-Uhlenbeck (OU) process

$$d\mathbf{x} = \boldsymbol{C}\mathbf{x}dt + \boldsymbol{\Sigma}d\boldsymbol{B}$$

to model sea surface temperature anomalies and test their potential for predicting the El-Nino Southern Oscillation (ENSO). ENSO is a basin wide warming phenomenon in the South Pacific Ocean which occurs quasi-periodically with approximate period 18 months. Parameters of the OU process are estimated by taking moments of the Fokker-Planck equation. One computes an estimate of the so called Green's function matrix as

$$\boldsymbol{G}(\tau_0) = \exp(\boldsymbol{B}\tau_0) = < \boldsymbol{x}(t+\tau_0)\boldsymbol{x}(t)^T >< \boldsymbol{x}(t)\boldsymbol{x}^T(t) >^{-1}\ .$$

The eigenvalues of the matrix $\boldsymbol{G}^T\boldsymbol{G}$ are known as POPs Principal Oscillation Patterns (POPs) or Empirical Normal Modes and are discussed earlier in the chapter. The Linear Inverse Model (LIM) employed here is a closely related technique to POP analysis.

The author uses the Green function matrix to determine the optimal initial structure for the system to evolve to the most probable prediction. This is determined to be the leading eigenfunction of $\boldsymbol{G}^T\boldsymbol{G}(\tau_0)$.

The OU process is fitted to monthly mean data taken from a $4^\circ \times 10^\circ$ grid between 1950-1991. The first 15 EOFs account for 65% of the variance. $\boldsymbol{G}(\tau_0)$ is

estimated with $\tau_0 =$ 4 months. Subsequent estimates of $\boldsymbol{B}$ are found insensitive to τ_0 which supports the choice of a linear model. Penland [1996] finds that the decay time scale of the estimated POPs are less than that of the ENSO oscillation. This may suggest that there is some interaction between modes or that this linear model is inappropriate. They look at predictability for lead times of 3, 6 and 9 months using Root Mean Square error. The 3 month lead time has some predictive skill but performs poorly during the strong warm phase of ENSO. It is found that this forecasting method can capture the ENSO pattern and persistence but not the magnitude. This could again indicate a problem with using a linear model.

Further work has been done on fitting LIMs to ENSO. Johnson et al. [2000] find an improvement in their model of sea surface temperature EOFs by including the first two EOFs of subsurface heat content anomalies as measured by RMS error. Penland and Matrosova [1998] apply LIM to Atlantic sea surface temperature anomalies to determine if there is any predictive skill gained by using global Sea Surface Temperatures as predictors. They confirm that this is the case. In terms of applications to atmospheric data most have focussed upon the related problem of determining POPs from data. POPs are derived from the assumption of a linear model and represent the normal modes of the dynamics. They are different to EOFs in that they are not optimised for explained variance and they do not form a set of orthogonal patterns. They are dynamical modes of the system, not standing patterns like EOFs. For example, Xu and von Storch [1990] determine POPs for sea level pressure between 15°S and 40°S in order to describe the development of the Southern Oscillation. They discovered that the 30-60 day oscillation may be predicted by the POP forecast scheme for several days, better than persistence and an Auto-Regressive Moving Average (ARMA) model. von Storch and Baumhefner [1991] extend this work to predictions of the equatorial velocity field and examine the accuracy using the anomaly correlation skill score.

A LIM is applied to Northern Hemisphere wintertime low frequency variability by Winkler et al. [2001]. 30 EOFs, capturing 90% of the variability, were computed for combined 250 hPa and 750 hPa Northern Hemisphere (NH) streamfunction anomalies together with 7 EOFs, capturing 70% of the variability, for tropical diabatic heating 30°S to 30°N. The LIM was then formed from this 37 component vector using the same methods as Penland [1996], discussed above. The measure of predictive skill used is the local anomaly correlation in 250 hPa streamfunction at a lead time of 14 days. By this measure the LIM outperforms forecasts based on climatology, persistence, a barotropic numerical model and a baroclinic model. The LIM competes with the skill of the then medium range prediction model of NCEP

with $O(10^6)$ variables. The authors attribute the skill of the LIM to its ability to approximate some of the nonlinear effects that do not appear in models constructed by linearising the full system. They also note the importance of including the tropical diabatic heating as a dynamic variable rather than an external forcing which is gives a marked improvement on the work of Penland and Ghil [1993]. They also apply a LIM to extratropical variability with tropical heating as a dynamical variable and report predictive skill only modestly better than the persistence prediction. Although their poor result could be more to do with their attempt to build a model for all seasons. Winkler et al. [2001] conclude that the dynamics of extratropical variability are essentially linear and stable if sufficient variables are included in the model although LIM still fails to capture the full amount of wintertime variability.

In many geophysical systems linear dynamics with white noise forcing are not sufficient. Kravtsov et al. [2005] suggest a data driven approach to constructing a nonlinear stochastic model. In particular they consider quadratic models such that the inference is still linear in the parameters. They estimate the parameters from model data using the least squares procedure where the dependent variables are the time derivatives. Their novel suggestion is to account for the autocorrelation in the residuals by adding extra unobserved levels. Each extra level is a linear equation for the residual. In this way more levels are added until the residuals on the final level are uncorrelated in time. Their model equations are

$$
\begin{aligned}
dx_i &= (\boldsymbol{x}^T A_i \boldsymbol{x} + b_i^{(0)} \boldsymbol{x} + c_i^{(0)})dt + r_i^{(0)} dt \\
dr_i^{(0)} &= b_i^{(1)}[\boldsymbol{x}, \boldsymbol{r}^{(0)}]dt + r_i^{(1)} dt \\
dr_i^{(1)} &= b_i^{(2)}[\boldsymbol{x}, \boldsymbol{r}^{(0)}, \boldsymbol{r}^{(1)}]dt + r_i^{(2)} dt \\
&\dots \\
dr_i^{(L)} &= b_i^{(2)}[\boldsymbol{x}, \boldsymbol{r}^{(0)}, \boldsymbol{r}^{(1)}, \dots, \boldsymbol{r}^{(L)}]dt + dr_i^{(L+1)} \,.
\end{aligned}
$$

Only the first level has nonlinear terms for climate variable $\boldsymbol{x}$, the others are linear equations for the residuals. They iteratively add more levels until the lag 1 autocorrelation is zero. The structure is similar to a multivariate autoregressive moving average model except nonlinear terms are included.

They demonstrate this method by estimating parameters for the three dimensional Lorenz model, which is a deterministic chaotic system, and also for stochastic cubic models. They report that their method is able to reproduce the parameters for the nonlinear terms but that there is dependence upon the data sampling strategy. In particular the estimated errors are large for infrequent observations. One of the main themes of this thesis is to develop an inference method that works with infre-

quent observations. Kravtsov et al. [2005] also discuss the problem of having such a large number of parameters that the problem is ill-conditioned. In their application to a semi-realistic atmospheric model they have over 3,000 parameters to estimate for a 15 dimensional system. They use a regularisation procedure which makes the inference well posed: specifically the methods of principal component regression and partial least squares. Motivated by this problem, in this thesis, we use a Bayesian approach where one places priors on the parameters. This leads to a well posed inverse problem.

Kravtsov et al. [2005] apply their method to the barotropic quasi-geostrophic three level model of Marshall and Molteni [1993]. They use a cross validation method to determine the number of variables for the reduced model. By splitting the data into two sets they train the model on one and test its predictive performance on the other. They then determine the number of levels needed to account for the autocorrelation in the noise. They settle for using 15 variables and three levels and report that the reduced model has a similar climatology to the full. They analyse the PDFs for the full and reduced by fitting Gaussian mixture models to the data. In both cases four mixture components was the optimal and they had similar clusters to each other. They looked at the ability of the reduced model to attribute the correct probability mass to regions associated with persistent flow regimes for the Northern Hemisphere. They confirm that the reduced model can capture the statistics of the positive and negative phases of the Arctic Oscillation and North Atlantic Oscillation. They use Singular Spectrum Analysis to determine the skill of the reduced model in capturing the low frequency variability. They compare their results to those from a single level model and conclude that this model is indistinguishable from a red spectrum whereas the multilevel model can capture the correct spectrum of the principal components of the full model. Moreover, they state that the single level model is sensitive to the particular realisation of the noise used and can have trajectories which diverge away from the stable patterns of the full model.

As argued by Majda and Yuan [2012] multilevel, quadratic regression can produce nonphysical behaviour such as finite time blow up and non-existence of an invariant measure. They also note the effects of error due to the sampling interval of the training data. They argue that a physics based model (motivated by the homogenisation procedure outlined above) with cubic non-linearity has more predictive skill.

In this section we have argued that linear models are insufficient to model well low frequency variability of the atmosphere; sparse observations can lead to errors and inconsistency in estimates of parameters for diffusion models and that it

is desirable to use physically motivated non-linear models as in those resulting from the rigorous homogenisation (MTV) procedure. However, the MTV procedure relies upon the estimation of hundreds of parameters from observations of the full system and may be inappropriate when there is lack of time scale separation. Therefore, we argue for a data driven approach where the parametric model is motivated by the MTV procedure: in quadratic models of the atmosphere this is usually a cubic model with noise that is linear in the state. We also argue in favour of theoretically well understood likelihood based inference to estimate the parameters. This leads to estimates with quantifiable errors. In particular the Bayesian approach will allow us to overcome any possible ill-posedness of the inference and will also prove useful in restricting the parameter space to give stable models. To overcome the problem of errors associated with infrequent observations we develop data imputation methods proposed in the SDE inference literature. In the next section we introduce the models with which we will work.

3.5 Model Problems

We apply our methods to a range of toy models becoming more sophisticated approximations of the real atmosphere. We work with these particular models to highlight some of the difficulties in the estimation: we can control the time scale separation between "climate" and "weather" variables to quantify how this affects the parameter estimates; with the triad model we can investigate the practical difficulties associated with multivariate problems (the number of parameters grows large) and using the Burgers heat bath we can assess the suitability of approximating chaotic dynamics with a linear stochastic process. Finally we demonstrate a geophysical model, discuss its derivation and some of its properties.

3.5.1 Chaotic Lorenz Model

The homogenisation procedure can be applied in the case where our original system is fully deterministic. As long as it is ergodic and mixing a reduced SDE can be derived. For example, Mitchell and Gottwald [2012] assess the potential of using a reduced model for data assimilation. They consider the following deterministic

system

$$\begin{aligned}
\frac{dx}{dt} &= x - x^3 + \frac{4}{90\epsilon} y_2 \\
\frac{dy_1}{dt} &= \frac{10}{\epsilon^2}(y_2 - y_1) \\
\frac{dy_2}{dt} &= \frac{1}{\epsilon^2}(28y_1 - y_2 - y_1 y_3) \\
\frac{dy_3}{dt} &= \frac{1}{\epsilon^2}(y_1 y_2 - \frac{8}{3} y_3)\,.
\end{aligned} \tag{3.29}$$

Here, the slow variable x moves inside a double well kicked by the chaotic Lorenz system, which is of order ϵ faster. Mitchell and Gottwald [2012] show that the reduced system is

$$dX = X(1 - X^2)dt + \sigma dB_t\,, \tag{3.30}$$

where

$$\frac{\sigma^2}{2} = \left(\frac{4}{90}\right) \int_0^\infty \left(\lim_{T\to\infty} \frac{1}{T} \int_0^T y_2(s) y_2(t+s) ds \right) dt$$

due to Birkhoff's ergodic theorem . Mitchell and Gottwald [2012] estimate σ using a quadratic variation method. They find that the reduced model Eq. (3.30) is a good approximation of the full model Eq. (3.29) for time scale $\epsilon = 0.01$ by analysing the PDFs and autocorrelation time scale though they note that the quality is sensitive to the estimate of σ. They find that the reduced model outperforms the full model when used for data assimilation: it is better at tracking the truth, given noisy observations, as measured by RMS error. They argue that this is due to the larger variance in the ensemble Kalman filter when using the reduced model. In Chapter 8 we estimate σ, for the reduced model in Eq. (3.30), using the likelihood based inference techniques discussed in Chapter 4. We study the affect of time scale separation on the results. We also fit a model with a latent, unobserved noise process and find that this is a good model when there is lack of time scale separation. We analyse the effect of low frequency observations upon the estimation of parameters and the implications for the skill of the resulting model.

3.5.2 Multiplicative Triad System

We consider now an example with two dimensions and multiplicative noise terms. This system was studied by Majda et al. [1999] and Majda et al. [2002]. Consider

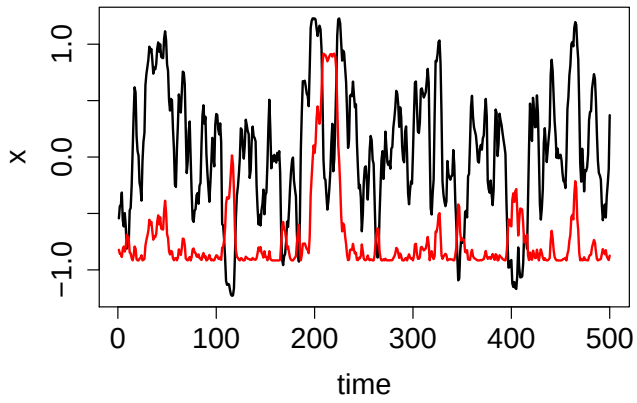


Figure 3.1: Example of solution for triad model in Eq. (3.31) for $\epsilon = 0.1$ with x_1 shown in black and x_2 in red.

the triad equations below

$$\begin{aligned} dx_1 &= \frac{b_1}{\epsilon} x_2 x_3 dt \\ dx_2 &= \frac{b_2}{\epsilon} x_1 x_3 dt \\ dx_3 &= \frac{b_3}{\epsilon} x_1 x_2 dt - \frac{\gamma}{\epsilon^2} x_3 dt + \frac{\sigma}{\epsilon} dB_t \,. \end{aligned} \tag{3.31}$$

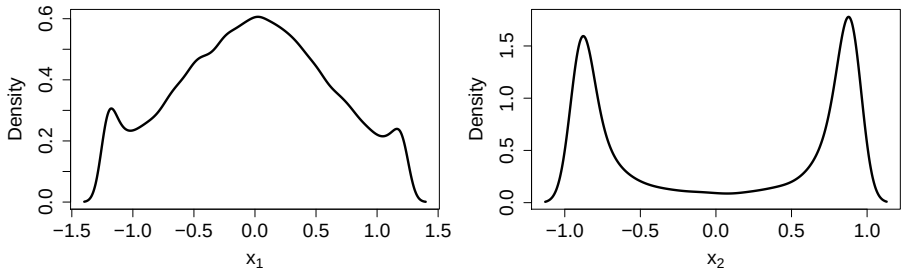
This system is stable provided that $b_1 + b_2 + b_3 = 0$. The Manley-Rowe relation $M = b_1 x_2^2 - b_2 x_1^2$ is conserved [Majda et al., 2002]. An example data set for x_1 and x_2 from this model is shown in Figure 3.1. The system moves on an ellipse if $b_1 b_2 < 0$ or a hyperbola if $b_1 b_2 > 0$. We use the values $b = (0.9, -0.5, -0.4)$, $\sigma = 0.5$ and $\gamma = 0.9$ so the system is confined to an ellipse with invariant density (see Majda et al. [2002])

$$p(x_1, x_2, x_3) = \frac{1}{Z} \exp\left(-\beta(b_1 x_3^2 - b_3 x_1^2)\right) \delta(b_1 x_2^2 - b_2 x_1^2 - M) \,,$$

where $\beta = \gamma / b_1 \sigma^2$ and $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ otherwise. Numerical approximations to the invariant distributions are shown in Figure 3.2.

We are interested in eliminating x_3 leaving equations for just x_1 and x_2. The small parameter ϵ represents the time scales within the system. The variable x_3 has fastest time scale of order $O(1/\epsilon^2)$ compared to $O(1/\epsilon)$ for x_1 and x_2. As $\epsilon \to 0$ we can use the method of homogenisation for SDEs to eliminate the fast variable x_3.

For this simple system the result can be derived by direct calculation. The



(a) Invariant distribution of x_1.

(b) Invariant distribution of x_2.

Figure 3.2: Invariant distributions for variables in Eq. (3.31) for $\epsilon = 0.1$.

solution of the third equation is

$$x_3(t) = e^{-\gamma t/\epsilon^2} x_3(0) + \frac{b_3}{\epsilon} \int_0^t e^{-\gamma(t-s)/\epsilon^2} x_1(s) x_2(s) ds + \frac{\sigma}{\epsilon} \int_0^t e^{-\gamma(t-s)/\epsilon^2} dB_s \,.$$

After substituting this expression into the equations for x_1 and x_2 we multiply by a factor of $1/\epsilon$. As we take the limit $\epsilon \to 0$ we have

$$\frac{1}{\epsilon} e^{-\gamma(t-s)/\epsilon^2} \to 0$$

and for the second term

$$\frac{b_3}{\epsilon^2} \int_0^t e^{-\gamma(t-s)/\epsilon^2} x_1(s) x_2(s) ds \to \frac{b_3}{\gamma} x_1(t) x_2(t) \,.$$

For the noise term, if we define

$$g(t) = \frac{\sigma}{\epsilon} \int_0^t e^{-\gamma(t-s)/\epsilon^2} dB_s$$

then this has zero mean and covariance

$$\langle g(t) g(t') \rangle = \frac{\sigma^2 \epsilon^2}{2\gamma} (e^{-\gamma|t-t'|/\epsilon^2} - e^{-\gamma(t+t')/\epsilon^2}) \,.$$

For arbitrary test function $\psi(t, t')$ we have

$$\frac{1}{\epsilon^2} \int_0^T \int_0^T \psi(t, t') \langle g(t) g(t') \rangle dt dt' \to \frac{\sigma^2}{\gamma^2} \int_0^T \psi(t, t) dt$$

so that as $\epsilon \to 0$, $g(t)$ is a white noise process, i.e

$$\frac{1}{\epsilon} g(t)dt \to \frac{\sigma}{\gamma} dB_t \,.$$

Note that as a process with finite correlation time dB_t must be interpreted in the Stratonovich sense (see Section 2.2). In the limit $\epsilon \to 0$ the system in (3.31) becomes

$$\begin{aligned} dx_1(t) &= \frac{b_1 b_2}{\gamma} x_2^2(t) x_1(t) dt + \frac{\sigma}{\gamma} b_1 x_2(t) \circ dB_t \\ dx_2(t) &= \frac{b_2 b_3}{\gamma} x_1^2(t) x_2(t) dt + \frac{\sigma}{\gamma} b_2 x_1(t) \circ dB_t \,. \end{aligned}$$

Written in Ito form this is

$$\begin{aligned} dx_1(t) &= \frac{b_1}{\gamma}(b_3 x_2^2(t) + \frac{\sigma^2}{2\gamma} b_2) x_1(t) dt + \frac{\sigma}{\gamma} b_1 x_2(t) dB_t \\ dx_2(t) &= \frac{b_2}{\gamma}(b_3 x_1^2(t) + \frac{\sigma^2}{2\gamma} b_1) x_2(t) dt + \frac{\sigma}{\gamma} b_2 x_1(t) dB_t \,. \end{aligned} \tag{3.32}$$

3.5.3 Burgers Equation

Here we use the same triad model as before but now, instead of the stochastic term, the equations are coupled to a high dimensional non-linear deterministic system. This will serve as a test of the first step of the MTV procedure: the approximation of chaotic dynamics by a Ornstein-Uhlenbeck process. For this purpose we use the inviscid Burgers equation

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = 0 \tag{3.33}$$

with complex Fourier amplitudes $\hat{u}_k = y_k + i z_k$ and reality condition $\hat{u}_{-k} = \hat{u}_k^*$. We simulate the Galerkin truncated system for modes $1 \le k \le \Lambda$

$$\begin{aligned} \frac{dy_k}{dt} &= -\mathrm{Re} \frac{ik}{2} \sum_{p+q+k=0} \hat{u}_p^* \hat{u}_q^* \\ \frac{dz_k}{dt} &= -\mathrm{Im} \frac{ik}{2} \sum_{p+q+k=0} \hat{u}_p^* \hat{u}_q^* \end{aligned}$$

We simulate the system using a pseudo-spectral method [Peyret, 2002]. Figure 3.3 shows the solution u at different times for truncation $\Lambda = 50$. The exact solution of 3.33 develops a shock discontinuity, whereas the Galerkin truncation dissipates the energy to the other modes.

The equations satisfy the Liouville property: they are measure preserving and conserve energy. This implies that the canonical Gibbs measure is a stationary

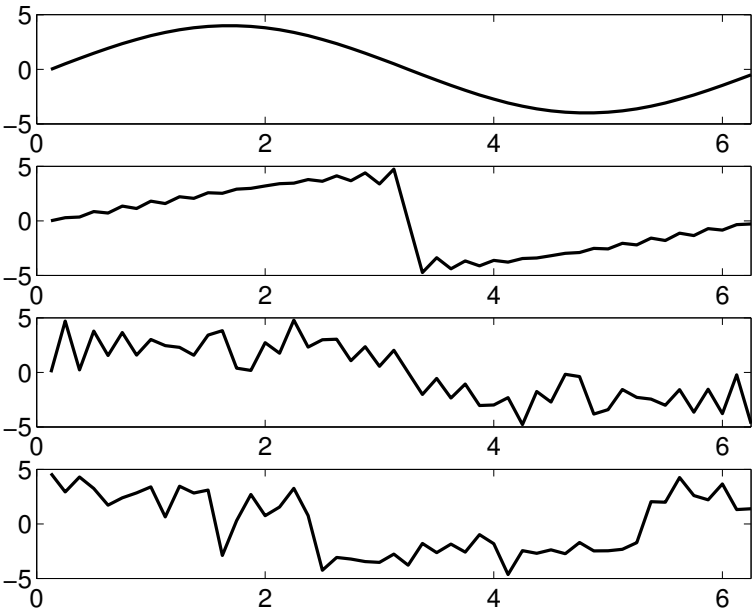


Figure 3.3: Solution of the Galerkin truncation of the Burgers equation for times $t = 0, 0.4, 1.5, 20$

distribution for the system

$$p_\beta = \frac{1}{Z} \exp\left(-\beta \sum_k^\Lambda |u_k|^2 \right)$$

for fixed β. The Gibbs measure predicts equipartition of energy between modes and so the system is a good approximation to a thermal heat bath and a suitable toy model of the atmosphere (see e.g Majda and Wang [2006]). As shown in Figure 3.4 this system is chaotic with different time scales. It represents the type of dynamics observed in atmospheric variables and is an ideal toy model to test stochastic mode reduction methods.

The Burgers system is coupled to the triad model through the mode $k = 1$ as follows

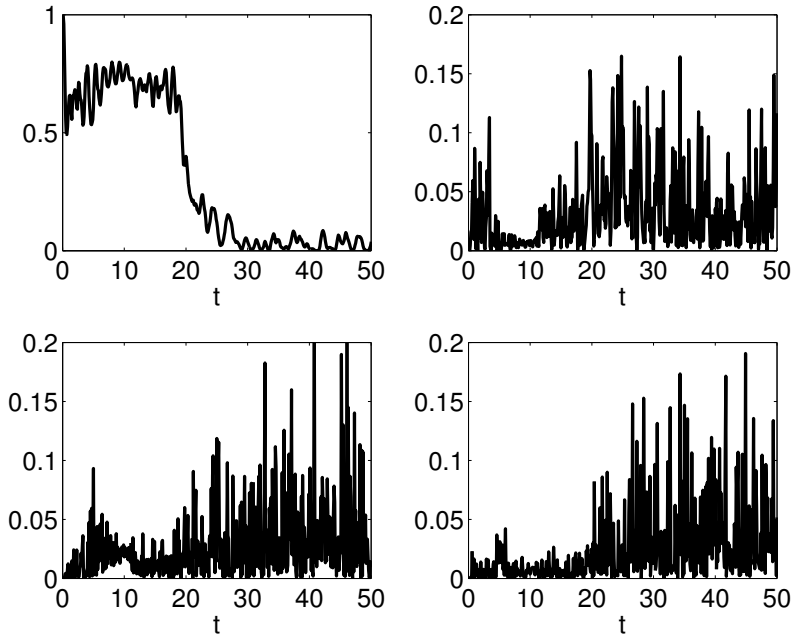


Figure 3.4: Evolution of Fourier amplitudes for $k = 1, 5, 10, 20$

$$\begin{aligned}
\frac{dx_1}{dt} &= b_1 x_2 y_1 \\
\frac{dx_2}{dt} &= b_2 x_1 y_1 \\
\frac{dy_k}{dt} &= b_3 x_1 x_2 \delta_{1,k} - \mathrm{Re}\frac{ik}{2} \sum_{p+q+k=0} \hat{u}_p^* \hat{u}_q^* \\
\frac{dz_k}{dt} &= -\mathrm{Im}\frac{ik}{2} \sum_{p+q+k=0} \hat{u}_p^* \hat{u}_q^*
\end{aligned} \tag{3.34}$$

The first step of the Stochastic Mode Reduction procedure is to make the approximation

$$dy_1 = b_3 x_1 x_2 dt - \gamma y_1 dt + \sigma dB_t \,,$$

where the parameters γ and σ are unknown. Then the homogenisation procedure is implemented as for the model in Eq. (3.31).

3.5.4 Quasi-Geostrophic Model on the β-plane with Mean Flow

Here we introduce a toy model with more of the physics of the real atmosphere. It is based on the much studied barotropic model in a beta channel of Charney and

De Vore [1979]. We first give a rough outline as to the origin of this equation. For a rigorous derivation see Pedlosky [1987] and for a further discussion of its properties see Majda and Wang [2006].

Consider first the shallow water equations in a rotating frame

$$\begin{aligned}\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}&=-\frac{1}{\rho}\frac{\partial p}{\partial x}+2\Omega v\sin\phi\\ \frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}&=-\frac{1}{\rho}\frac{\partial p}{\partial y}-2\Omega u\sin\phi\,,\end{aligned}\tag{3.35}$$

where u and v are **longitudinal and latitudinal velocity**, p is **pressure**, ρ is the density, Ω is the **rotation** of the Earth and ϕ is the polar angle. The left hand side is the advection of velocity and is simply the acceleration in a moving reference frame. The right hand side consists of forces due to the gradient in pressure and the rotation of the Earth.

The velocity is assumed to be close to **geostrophic**. Geostrophic flow is when the velocity is parallel to the isobars (lines of constant pressure). The pressure field completely determines the flow. For example, in a geostrophic atmosphere in the northern hemisphere the wind blows anti-clockwise around regions of low pressure. The geostrophic flow is perpendicular to the pressure gradient

$$\begin{aligned}-2\Omega v_g\sin\phi&=-\frac{1}{\rho}\frac{\partial p}{\partial x}\\ 2\Omega u_g\sin\phi&=-\frac{1}{\rho}\frac{\partial p}{\partial y}\,.\end{aligned}$$

It is assumed that the size of the system is small and so the curved surface of the Earth can be represented as a tangent plane, such that

$$2\Omega\sin\phi\approx f+\beta y\,.$$

Then in the shallow water equations Eq. (3.35) all velocities, except the leading Coriolis terms are approximated by the geostrophic flow

$$\begin{aligned}\frac{du_g}{dt}-fv-\beta y v_g&=-\frac{1}{\rho}\frac{\partial p}{\partial x}\\ \frac{dv_g}{dt}+fu+\beta y u_g&=-\frac{1}{\rho}\frac{\partial p}{\partial y}\,.\end{aligned}$$

This gives the system of equations

$$-\frac{1}{\rho f}\frac{\partial^2 p}{\partial y \partial t} - \frac{1}{\rho^2 f^2} J\left(p, \frac{\partial p}{\partial y}\right) - fv - \frac{\beta}{\rho f} y \frac{\partial p}{\partial x} = -\frac{1}{\rho}\frac{\partial p}{\partial x}$$
$$\frac{1}{\rho f}\frac{\partial^2 p}{\partial x \partial t} + \frac{1}{\rho^2 f^2} J\left(p, \frac{\partial p}{\partial x}\right) + fu - \frac{\beta}{\rho f} y \frac{\partial p}{\partial y} = -\frac{1}{\rho}\frac{\partial p}{\partial y}\,,$$

where $J(a,b) = (\partial a/\partial x)(\partial b/\partial y) - (\partial b/\partial x)(\partial a/\partial y)$. Differentiating the top equation by y and the bottom by x, then substituting into the continuity equation

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0$$

yields

$$\frac{\partial}{\partial t}\nabla^2 p + \frac{1}{\rho f} J\left(p, \nabla^2 p\right) + \beta \frac{\partial p}{\partial x} = 0\,.$$

Recasting in terms of the **stream function** $\psi = \frac{p}{\rho f}$ and **potential vorticity** $q = \Delta\psi$, we have

$$\frac{\partial q}{\partial t} + J(\psi, q) + \beta \frac{\partial \psi}{\partial x} = 0\,. \tag{3.36}$$

We equip the system with periodic boundary conditions in a channel domain

$$\psi(x + 2\pi, y, t) = \psi(x, y + 2\pi, t) = \psi(x, y, t)\,,$$

$$\int \psi(x, y, t) dx dy = 0\,.$$

The most general stream function satisfying the boundary conditions has the form $\psi = \psi' - Uy$. Here we have introduced a background mean flow U in the longitudinal direction. The wind velocity is then given by

$$(u, v) = \nabla^\perp \psi' = \begin{pmatrix} -\frac{\partial \psi}{\partial y} + U \\ \frac{\partial \psi}{\partial x} \end{pmatrix}.$$

Introducing the bottom **topography** $h(x, y)$, the potential vorticity is

$$q = \Delta\psi + h(x, y)\,.$$

The total energy of the system is

$$\begin{aligned} E_{\text{total}} &= E_{\text{mean flow}}(t) + E_{\text{small scale}}(t) \\ &= A_R \frac{1}{2} U^2(t) + \frac{1}{2} \int |\nabla \psi'|^2 \\ &= \text{constant} \,. \end{aligned} \tag{3.37}$$

The second term is called **topographic stress** [Majda and Wang, 2006] and it can be shown that

$$\frac{dE_{\text{small scale}}(t)}{dt} = U(t) \int \frac{\partial h}{\partial x} \psi' \,.$$

Differentiating Eq. (3.37) gives an equation for the evolution of the mean flow

$$\frac{dU}{dt} = \frac{1}{4\pi^2} \int h \frac{\partial \psi}{\partial x} \, dx \, dy \,, \tag{3.38}$$

which completes the dynamical description of the model. Initially we consider the simple topography in the longitudinal direction

$$h(x) = H(\cos(x) + \sin(x)) \,. \tag{3.39}$$

The value of H is important in determining the strength of the topographic stress. Combining the equation for mean flow with Eq. (3.36) gives the **Quasi-Geostrophic Equations on the beta-plane with Mean Flow and bottom topography**:

$$\begin{aligned} &\frac{\partial q}{\partial t} + \nabla^{\perp} \psi \cdot \nabla q + U \frac{\partial q}{\partial x} + \beta \frac{\partial \psi}{\partial x} = 0 \\ &q = \Delta \psi + h, \;\; \frac{dU}{dt} = \frac{1}{4\pi^2} \int h \frac{\partial \psi}{\partial x} dx dy \,. \end{aligned} \tag{3.40}$$

Figure 3.5 demonstrates the non-linear dynamics of the mean flow U. The predominant flow is the negative zonal direction with regime transitions to positive flow.

We derive a diffusion model for the mean flow U. Firstly we state some results about the stationary measure of the truncated system from Majda and Wang [2006].

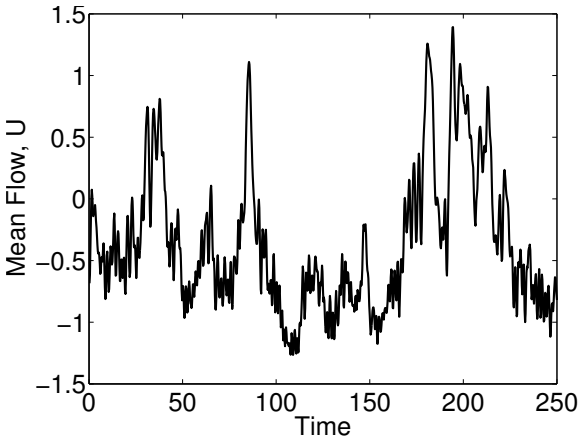


Figure 3.5: Example of dynamics of Mean flow U_t from Eq. (3.40)

Taking the Fourier transform of the system Eq. (3.40) gives

$$\frac{dU}{dt} = i \sum_{1 \leq |k|^2 \leq \Lambda} k_x h_{-k} \psi_k \tag{3.41}$$

$$\frac{d\psi_k}{dt} = \frac{ik_x\beta}{|k|^2}\psi_k - ik_x U\psi_k + \frac{ik_x U h_k}{|k|^2} - \sum_{\substack{l+m=k \\ |l|^2 \leq \Lambda \\ |m|^2 \leq \Lambda}} \frac{l^\perp \cdot m}{|k|^2}\psi_l(-|m|^2\psi_m + h_m)\,, \tag{3.42}$$

where $k = (k_x, k_y)$ is the wave number of the mode in the x and y directions. We restrict the model with spherical cut-off $1 \leq |k|^2 \leq \Lambda$. The truncated energy E_Λ is conserved for this system

$$E_\Lambda = \frac{1}{2}U^2 + \frac{1}{2}\int |\nabla^\perp \psi_\Lambda|^2 dx = \frac{1}{2}U^2 + \frac{1}{2}\sum_{1 \leq |k|^2 \leq \Lambda} |k|^2 |\psi_k|^2\,. \tag{3.43}$$

The enstrophy is also conserved. This is given by the integral of the square of the vorticity and is often used as a measure of dissipation in a system of fluids. In this model the truncated enstrophy

$$\epsilon_\Lambda = \beta U + \frac{1}{2}\int q^2 dx = \beta U + \frac{1}{2}\sum_{1 \leq |k|^2 \leq \Lambda} |-|k|^2\psi_k + h_k|^2 \tag{3.44}$$

is conserved To prove energy conservation differentiate Eq. (3.43) with respect to

time

$$\begin{aligned}\frac{dE_\Lambda}{dt} &= U\frac{dU}{dt} + \sum_k |k|^2\psi_{-k}\frac{d\psi_k}{dt} \\ &= Ui\sum_k k_x h_{-k}\psi_k + \sum_k |k|^2\psi_{-k}\left(\frac{ik_x\beta}{|k|^2}\psi_k - ik_xU\psi_k + \frac{ik_xUh_k}{|k|^2}\right. \\ &\left. - \sum_{l+m=k}\frac{l^\perp\cdot m}{|k|^2}\psi_l(-|m|^2\psi_m + h_m)\right). \end{aligned} \tag{3.45}$$

The second and third terms on the right equal zero due to being anti-symmetric in k_x. For the last term

$$\begin{aligned}&\sum_{l+m=k} l^\perp\cdot m\,\psi_{-k}\psi_l(-|m|^2\psi_m + h_m)) \\ &= \sum_{-k+m=-l} l^\perp\cdot m\,\psi_{-k}\psi_l(-|m|^2\psi_m + h_m)) \\ &= \sum_{l+m=k} -k^\perp\cdot m\,\psi_l\psi_{-k}(-|m|^2\psi_m + h_m)) \\ &= \sum_{l+m=k} -(l+m)^\perp\cdot m\,\psi_l\psi_{-k}(-|m|^2\psi_m + h_m)) \\ &= -\sum_{l+m=k} l^\perp\cdot m\,\psi_{-k}\psi_l(-|m|^2\psi_m + h_m))\end{aligned}$$

and therefore vanishes leaving

$$\frac{dE_\Lambda}{dt} = iU\sum_k k_x h_{-k}\psi_k + iU\sum_k k_x h_k\psi_{-k} = 0\,.$$

Conservation of enstrophy can be proved similarly.

Since the equations satisfy the Liouville property we can use these conservation laws and equilibrium statistical mechanics to write an invariant Gibbs measure for the ensemble

$$\begin{aligned}\rho_{\alpha,\theta} = C\exp\left(-\alpha\left(\beta U + \frac{1}{2}\sum_{1\le|k|^2\le\Lambda} |-|k|^2\psi_k + h_k|^2\right)\right. \\ \left. -\,\theta\left(\frac{1}{2}U^2 + \frac{1}{2}\sum_{1\le|k|^2\le\Lambda}|k|^2|\psi_k|^2\right)\right).\end{aligned} \tag{3.46}$$

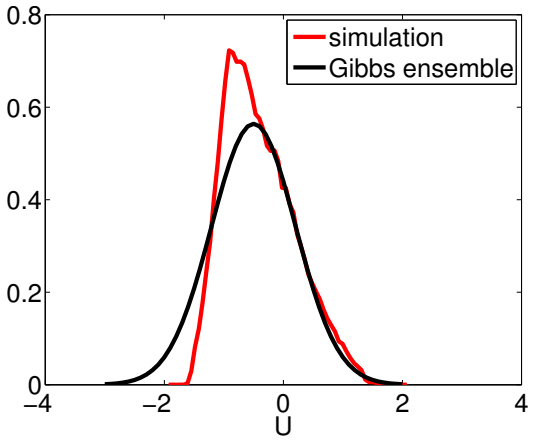


Figure 3.6: Comparison of predicted density from equilibrium statistical mechanics and the empirical density for the mean flow.

Normalising gives a Gaussian density with mean

$$(\bar{U}, \bar{\psi}_k) = (-\frac{-\beta}{\mu}, \frac{h_k}{\mu + |k|^2})$$

and variance

$$\begin{aligned} \mathrm{Var}(U) &= \frac{1}{\alpha\mu} \\ \mathrm{Var}(\psi_k) &= \frac{1}{\alpha|K|^2(\mu + |k|^2)} \,. \end{aligned}$$

Figure 3.6 compares the Gibbs ensemble predicted by statistical mechanics with the numerically computed invariant density. The statistical mechanics prediction is a good approximation, only lacking the skewness of the true invariant distribution.

We now write the equations where the non-linear terms for the stream functions have been parametrised by a complex linear stochastic process and time-scale separation has been identified with small parameter ϵ

$$\begin{aligned} \frac{d\psi_k}{dt} &= \frac{i}{\epsilon} k_x H_k U - \frac{\gamma_k(U)}{\epsilon^2}\psi_k + \frac{\sigma_k}{\epsilon}\dot{B}_k \\ \frac{dU}{dt} &= \frac{2}{\epsilon}\mathrm{Im}\sum_k k_x H_k \psi_k \,, \end{aligned}$$

where $\gamma_k(U) = \gamma_k + i\Omega_k + ik_x/\sqrt{\alpha\mu}U$ and $\gamma_k, \sigma_k \in \mathbb{R}$. The corresponding Fokker-

Plank equation, with test function v, is

$$\frac{\partial v}{\partial t} = \frac{1}{\epsilon^2}\mathcal{L}_0 v + \frac{1}{\epsilon}\mathcal{L}_1 v\,, \tag{3.47}$$

where $\mathcal{L}_0$ and $\mathcal{L}_1$ are the operators

$$\mathcal{L}_0 = -\sum_k \gamma_k(U)\psi_k \frac{\partial}{\partial \psi_k} + \frac{1}{2}\sum_k \sigma_k^2 \frac{\partial^2}{\partial \psi_k \partial \psi_k^*} = g\cdot\nabla_\psi + \frac{1}{2}B : \nabla_\psi\nabla_\psi$$
$$\mathcal{L}_1 = 2\mathrm{Im}\sum_k k_x H_k \psi_k \frac{\partial}{\partial U} + \sum_k i k_x H_k U \frac{\partial}{\partial \psi_k} = f_0\cdot\nabla_U + f_1\cdot\nabla_\psi\,.$$

Now, expanding v as $v = v_0 + \epsilon v_1 + \epsilon^2 v_2 + O(\epsilon^3)$, substituting into Eq. (3.47) and equating coefficients of ϵ gives

$$O(1/\epsilon^2):\ -\mathcal{L}_0 v_0 = 0 \tag{3.48}$$
$$O(1/\epsilon):\ -\mathcal{L}_0 v_1 = \mathcal{L}_1 v_0 \tag{3.49}$$
$$O(1):\ -\mathcal{L}_0 v_2 = -\frac{\partial v_0}{\partial t} + \mathcal{L}_1 v_1\,. \tag{3.50}$$

Eq. (3.48) implies that $v_0 = v_0(U,t)$. Eq. (3.49) becomes

$$-\mathcal{L}_0 v_1 = f_0(U,\psi)\cdot\nabla_U v_0(U,t) \tag{3.51}$$

giving solution

$$v_1(U,\psi,t) = \Phi(U,\psi)\cdot\nabla_U v_0(U,t)$$

since $\mathcal{L}_0$ is a differential operator in ψ alone. Substituting this into Eq. (3.51) gives the so called "cell problem"

$$\sum_k \gamma_k(U)\psi_k \frac{\partial \Phi}{\partial \psi_k}(\psi,U) - \frac{1}{2}\sum_k \sigma_k^2 \frac{\partial^2}{\partial \psi_k \partial \psi_k^*}\Phi(\psi,U) = 2\mathrm{Im}\sum_k k_x H_k \psi_k\,, \tag{3.52}$$

which has solution

$$\Phi(\psi,U) = 2i\sum_k \frac{k_x H_k \gamma_k}{|\gamma_k(U)|^2}\psi_k\,.$$

Note that this is centred (expectation zero) with respect to the invariant measure of the fast variables

$$\rho^\infty(\psi;U) = \prod_k \mathcal{N}(0, \frac{\sigma_k^2}{2\mathrm{Re}\gamma_k(U)})\mathcal{N}(0, \frac{\sigma_k^2}{2\mathrm{Im}\gamma_k(U)})\,.$$

The solvability condition for Eq. (3.50) gives

$$\frac{\partial v_0}{\partial t} = \int_\psi \rho^\infty(\psi; U)\mathcal{L}_1(\Phi(\psi, U) \cdot \nabla_U v_0(U, t))d\psi \,. \tag{3.53}$$

Note that

$$\mathcal{L}_1(\Phi\frac{\partial v_0}{\partial U}) = f_0\Phi\frac{\partial^2 v_0}{\partial U^2} + f_0\frac{\partial\Phi}{\partial U}\frac{\partial v_0}{\partial U} + f_1 \cdot \nabla_\psi\Phi\frac{\partial v_0}{\partial U} \,.$$

Consider the first term in Eq. (3.53)

$$\begin{aligned}
&\int \rho^\infty f_0\Phi\frac{\partial^2 v_0}{\partial U^2}d\psi \\
&= 4\int \rho^\infty \sum_k \frac{k_x H_k\gamma_k\psi_k}{|\gamma_k(U)|^2}\sum_k k_x H_k\psi_k\frac{\partial^2 v_0}{\partial U^2}d\psi \\
&= 4\int \sum_k \frac{k_x^2 H_k^2\gamma_k}{|\gamma_k(U)|^2}(\mathrm{Re}\psi_k)^2\mathcal{N}(0, \frac{\sigma_k^2}{2\gamma_k})\frac{\partial^2 v_0}{\partial U^2}d\psi \\
&= 2\sum_k \frac{k_x^2 H_k^2\sigma_k^2}{|\gamma_k(U)|^2}\frac{\partial^2 v_0}{\partial U^2} \,.
\end{aligned}$$

The second term gives

$$\begin{aligned}
&\int \rho^\infty f_0\frac{\partial\Phi}{\partial U}\frac{\partial v_0}{\partial U}d\psi \\
&= -8\int \sum_k \frac{k_x^2 H_k^2\gamma_k}{|\gamma_k(U)|^4}\frac{k_x}{\sqrt{\alpha\mu}}(\Omega_k + \frac{k_x U}{\sqrt{\alpha\mu}})(\mathrm{Re}\psi_k)^2\mathcal{N}(0, \frac{\sigma_k^2}{2\gamma_k})\frac{\partial v_0}{\partial U}d\psi_k \\
&= -4\sum_k \frac{k_x^2 H_k^2\sigma_k^2}{|\gamma_k(U)|^4}\frac{k_x}{\sqrt{\alpha\mu}}(\Omega_k + \frac{k_x U}{\sqrt{\alpha\mu}})\frac{\partial v_0}{\partial U}d\psi_k
\end{aligned}$$

and the third

$$\int \rho^\infty g\nabla_\psi\Phi\frac{\partial v_0}{\partial U}d\psi = -2\sum_k \frac{k_x^2 H_k^2\gamma_k}{|\gamma_k(U)|^2}\frac{\partial v_0}{\partial U} \,.$$

Bringing these together gives

$$\frac{\partial v_0}{\partial t} = -\tilde{\gamma}(U)U\frac{\partial v_0}{\partial U} + \gamma'(U)\frac{\partial v_0}{\partial U} + \gamma(U)\frac{\partial^2 v_0}{\partial U^2} \,, \tag{3.54}$$

where

$$\gamma(U) = 2\sum_k \frac{k_x^2 H_k^2\sigma_k^2}{|\gamma_k(U)|^2} \quad \text{and} \quad \tilde{\gamma}(U) = 2\sum_k \frac{k_x^2 H_k^2\gamma_k}{|\gamma_k(U)|^2} \,.$$

This gives the reduced SDE for the mean flow U

$$dU = (-\tilde{\gamma}(U)U + \gamma'(U))dt + \sqrt{2\gamma(U)}dB \,. \tag{3.55}$$

Majda et al. [2003] completed this model with estimates for γ_k and σ_k calculated from observations of the full original system. In Chapter 8 we estimate these parameters using likelihood based techniques and observations of U. The problem with this approach is that it is only possible if there are not many terms in the sums in Eq. (3.54). This will be the case using the simple topography in Eq. (3.39) as $H_k = 0$ for $k \neq (1, 0)$. In other cases the parameters γ_k and σ_k are not easily identifiable from observations of U.

In this chapter we have reviewed various approaches to stochastic climate modelling. We discussed the choice of basis and the method of reducing the number of variables in a system. We provided some background theory to the Stochastic Model Reduction method of Majda et al. [1999], Majda et al. [2001], Majda et al. [2009], Majda et al. [2002], Majda et al. [2003] and reproduced their result by an independent method. Using a set of minimal assumptions about the components of the original climate model and allowing for time scale separation between the resolved and unresolved variables one arrives at a reduced climate model. This model is a Stochastic Differential Equation and includes cubic terms in the drift function and linear additive and multiplicative noise in the diffusion function. We also derived a reduced model for some specific cases. We will use these as test cases in Chapter 8.

In Chapter 5 we develop methodology to infer the parameters of general cubic drift models with linear diffusion function. This is motivated from the general form of stochastic climate model derived from the full system in Eqns.. (3.20) and (3.21) and resulting in the model in Eq. (3.26). This resulting software is then applicable to most stochastic climate models. In Chapter 8 we compare the approach based on homogenisation and estimation of the few remaining parameters to the purely data driven approach of Chapter 5.

Chapter 4

Estimating Parameters in Stochastic Differential Equation Models

In this chapter we review inference methods used to estimate parameters of stochastic differential equations. Previous reviews include Sorensen [2004] and Hurn et al. [2007]. We focus our review on methods that could be applicable to the models derived in Chapter 3. We start by introducing the likelihood function and discuss the difficulties with this inference problem, particularly for the estimation of parameters in the diffusion function from discrete observations of the system. We also state some of the asymptotic properties of estimators of the drift parameters and how they relate to the frequency of observations. In Section 4.2 we demonstrate these issues for an example through a numerical study of the maximum likelihood estimator of the Ornstein-Uhlenbeck Process, which was introduced in Section 2.8. For general non-linear problems, such as those in Chapter 3, the likelihood function can not be calculated in closed form, although there are various ways in which it can be approximated. We review these in Section 4.3. In particular the theory in Section 4.3.3, which is based on the pivotal work of Pedersen [1995], provides the foundation for the Markov Chain Monte Carlo (MCMC) Methods of Section 4.3.4, which are the main focus of the methodology of this thesis. In that section, MCMC methods are discussed in the context of inference for diffusions. Problems of convergence, highlighted by Roberts and Stramer [2001], are central to these algorithms and are discussed here. Note that issues relating to the algorithmic efficiency and optimisation of the MCMC algorithms are left to Chapter 5, where they are introduced in the context of a particular SDE model. Finally, in Section 4.3.4, we give

details of a flexible MCMC algorithm named the **Innovation Scheme**. We write these in full pseudo-code in Algorithms 4.1 and 4.2 for reference later in the thesis as these form the foundation of the methods used in Chapters 5-8.

In Section 4.3 we review other ways of approximating the likelihood function. In particular, we discuss methods of local linearisation, which will form the foundation for an improved MCMC algorithm that we introduce in Chapter 5, Section 5.2.1. In Section 4.4 we discuss an alternative approach, where the exact likelihood function can be used. At present, this can only be done for a restricted class of models, but it is being generalised through a series of papers (Beskos et al. [2006], Beskos et al. [2008]) and might eventually be applicable to models of Chapter 3. Due to the difficulty of likelihood based estimation several researches have pursued non-likelihood based methods. For completeness, these are reviewed in Section 4.5. In the final section we discuss our reasoning for choosing to base our work on the Innovation Scheme of Section 4.3.4 with regards to the inference problem we motivated in Chapter 3.

4.1 Background

We assume that we have the model

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, \boldsymbol{\theta})dt + \boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\theta})d\boldsymbol{B}_t\,,\ \boldsymbol{X}_0 = \boldsymbol{x}_0\,,\ t \in [0, T] \tag{4.1}$$

with filtered probability space $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\})$. We consider the inference problem for unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^p$, state space $\boldsymbol{X} \in \mathbb{R}^d$ and m-dimensional Brownian motion $\boldsymbol{B}$. We restrict our attention to time homogeneous processes and require regularity conditions on the functions $\boldsymbol{\mu} : \mathbb{R}^d \to \mathbb{R}^d$ and $\boldsymbol{a} : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ for all $\boldsymbol{\theta}$ as discussed in Chapter 2 to guarantee a unique weak solution.

If $\boldsymbol{X}$ were observed in continuous time then the parameters $\boldsymbol{\sigma}$ entering into the diffusion coefficient would be completely determined by the **Quadratic Variation**

$$\int_0^t \boldsymbol{\Sigma}(\boldsymbol{X}_s, \boldsymbol{\sigma})ds = \lim_{\|P\|\to 0} \sum_{k=1}^{n} (\boldsymbol{X}_{t_k} - \boldsymbol{X}_{t_{k-1}}) \cdot (\boldsymbol{X}_{t_k} - \boldsymbol{X}_{t_{k-1}})^T\,, \tag{4.2}$$

where $\boldsymbol{\Sigma} = \boldsymbol{a}\boldsymbol{a}^T$ and the limit, in the supremum of the partition P, is valid for any time interval t. So in continuous time the parameters in the diffusion function can be considered known. This is a path property of diffusions as discussed in Chapter 2.

We assume that the parameters $\boldsymbol{\sigma}$ entering $\boldsymbol{a}$ are identifiable from observa-

tions of the quadratic variation. If the parameters partition as $\boldsymbol{\theta} = \{\boldsymbol{\gamma}, \boldsymbol{\sigma}\}$ such that $\boldsymbol{\gamma}$ are the remaining parameters not entering the diffusion function then they can be determined from the continuous time likelihood function $l_c(\boldsymbol{\gamma})$. If $\mathbb{Q}$ is the law of the driftless version of Eq. (4.1) then the continuous time likelihood is just the change of measure theorem discussed in Section 2.5. See Øksendal [2007, Chapter 8.6] for a proof. This is given by Girsanov's formula

$$l_c(\boldsymbol{\gamma}) = \frac{d\mathbb{P}}{d\mathbb{Q}}(\boldsymbol{X}|\boldsymbol{\gamma}) = \exp\left(\int_0^T \boldsymbol{\mu}(\boldsymbol{X}_s, \boldsymbol{\gamma})\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_s, \boldsymbol{\sigma})d\boldsymbol{X}_s \right. \\ \left. -\frac{1}{2}\int_0^T \boldsymbol{\mu}(\boldsymbol{X}_s, \boldsymbol{\gamma})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_s, \boldsymbol{\sigma})\boldsymbol{\mu}(\boldsymbol{X}_s, \boldsymbol{\gamma})ds\right) \tag{4.3}$$

However, continuous time observation is not realistic in applications so we assume observations $\boldsymbol{x}_t$ at discrete times $t_0 < t_1 \ldots < t_k < t_{k+1} < \ldots < t_N$, with interobservation interval $\Delta t_k = t_{k+1} - t_k$.

Due to the Markov property, Eq. (2.3), the likelihood function for $\boldsymbol{\theta}$ can be written as the product of transition densities between observations. These densities are defined from the transition probabilities as follows. At time s we observe $\boldsymbol{X}_s = \boldsymbol{x}_s$ then the probability of the process being in measurable set A at time t is

$$\mathbb{P}_\theta(\boldsymbol{X}_t \in A|\boldsymbol{X}_s = \boldsymbol{x}_s) = \int_{\boldsymbol{y}\in A} p(t, \boldsymbol{y}|s, \boldsymbol{x}; \boldsymbol{\theta})d\boldsymbol{y} \tag{4.4}$$

then $p(t, \boldsymbol{y}|s, \boldsymbol{x}; \boldsymbol{\theta})$ is the probability density for the process. For clarity we introduce the index notation for observations $\boldsymbol{x}_k = \boldsymbol{x}_{t_k}$. The log likelihood function is then written

$$l_N(\boldsymbol{\theta}) = \sum_{k=1}^{N-1} \log p(t_{k+1}, \boldsymbol{x}_{k+1}|t_k, \boldsymbol{x}_k; \boldsymbol{\theta})\,. \tag{4.5}$$

For ergodic diffusions, the maximum likelihood estimator has the usual good properties of consistency, asymptotic normality and efficiency as $N \to \infty$ [Dacunha-Castelle and Florens-Zmirou, 1986]. The problem is that the transition density is rarely available in closed form so many inference methods are based upon suitable approximations of the transition density.

One approach is to consider the discretised system as a Markov Chain based upon approximations to Eq. (4.1). As discussed in Section 2.9, the simplest discretised version of Eq. (4.1) is

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \boldsymbol{\mu}(\boldsymbol{X}_k, \boldsymbol{\theta})\Delta t_k + \boldsymbol{a}(\boldsymbol{X}_k, \boldsymbol{\theta})(\boldsymbol{B}_{k+1} - \boldsymbol{B}_k) \tag{4.6}$$

known as the Euler-Maruyama scheme. In this case the transition density can be approximated as a Gaussian density with mean and variance

$$\begin{aligned}\mathbb{E}_\theta(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k = \boldsymbol{x}_k) &= \boldsymbol{x}_k + \boldsymbol{\mu}(\boldsymbol{x}_k, \boldsymbol{\theta})\Delta t_k \\ \mathrm{Var}_\theta(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k = \boldsymbol{x}_k) &= \boldsymbol{\Sigma}(\boldsymbol{x}_k, \boldsymbol{\theta})\Delta t_k\,, \end{aligned} \tag{4.7}$$

where $\boldsymbol{\Sigma} = \boldsymbol{a}^T\boldsymbol{a}$.

Approximations of the likelihood based on Eq. (4.6) were studied by Florens-Zmirou [1989]. They assumed that the parameters of the model divided between those in the drift $\boldsymbol{\gamma}$ and those in the diffusion $\boldsymbol{\sigma}$. To estimate the diffusion parameter they assumed that the diffusion had the form $\boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\sigma}) = \boldsymbol{\sigma}\boldsymbol{a}(\boldsymbol{X}_t)$, for which they could then use the quadratic variation as an estimator. For the drift parameters they used the approximate likelihood function based on Eq. (4.7)

$$l_N^{\text{Euler}}(\boldsymbol{\theta}) = -\frac{1}{2\Delta t_k}\sum_{k=0}^{N-1}(\boldsymbol{x}_{k+1}-\boldsymbol{x}_k-\Delta t_k\boldsymbol{\mu}(\boldsymbol{x}_k,\boldsymbol{\theta}))^T\Sigma^{-1}(\boldsymbol{x}_k)(\boldsymbol{x}_{k+1}-\boldsymbol{x}_k-\Delta t_k\boldsymbol{\mu}(\boldsymbol{x}_k,\boldsymbol{\theta})) \tag{4.8}$$

The maximum likelihood estimator from this approximate likelihood function has good asymptotic properties with some restrictions on the maximum observation interval $h = \max_k \Delta t_k$. In the case where $Nh^2 \to 0$ the estimator is asymptotically efficient [Florens-Zmirou, 1989]. This scenario is the **increasing experimental design** assumption of Prakasa-Rao [1983]. For applications, including atmospheric science, it is more likely that the observation interval is fixed and the number of observations will increase. Florens-Zmirou [1989] showed that estimates based on the Euler-Maruyama scheme are biased to the order of the observational interval due to the misspecification of the mean and variance in the Gaussian approximation.

An improvement over the standard Euler-Maruyama approximation would be to use the Milstein scheme (see Section 2.9). This discretised system is given by

$$\boldsymbol{X}_{k+1} = \boldsymbol{X}_k + \boldsymbol{\mu}(\boldsymbol{X}_k,\boldsymbol{\theta})\Delta t_k + \boldsymbol{a}(\boldsymbol{X}_k,\boldsymbol{\theta})\sqrt{\Delta t_k}\boldsymbol{\xi}_k + \frac{\boldsymbol{a}(\boldsymbol{X}_k,\boldsymbol{\theta})}{2}\frac{d\boldsymbol{a}}{dx}(\boldsymbol{X}_k,\boldsymbol{\theta})\Delta t_k(\boldsymbol{\xi}_k^2-1)\,,$$

where $\boldsymbol{\xi} \sim \mathcal{N}_m(0,1)$. Since this includes the square of $\boldsymbol{\xi}_k$, the approximating transition density is not Gaussian. Elerian [98] showed that the density can be computed in closed form and involves a hyperbolic cosine function. This method is more accurate than the Euler method but one must take care that the approximating density gives numerically stable results: it is possible that the argument of the hyperbolic cosine can have non-zero imaginary part [Hurn et al., 2007].

4.2 Maximum Likelihood for the Ornstein-Uhlenbeck Process

As a demonstration of the importance of the size of the sampling interval in SDE inference we consider here the **Maximum Likelihood Estimator** (MLE) for the Ornstein-Uhlenbeck (O-U) process. In this case the exact transition density is tractable and the continuous time MLE is available in closed form. We compare this to the MLE derived from the approximation in (4.6).

The O-U model we consider is

$$dX_t = \phi X_t dt + \sigma dB_t,\ t \in [0, T]\,, \tag{4.9}$$

with $\phi \in (-\infty, 0)$ and $\sigma \in (0, \infty)$. We consider observations at fixed, constant interval Δ. Given X_t, the exact solution for $X_{t+\Delta}$ (see Section 2.8) is

$$X_{t+\Delta} = e^{\phi\Delta} X_t + \sqrt{\frac{-\sigma^2}{2\phi}(1 - e^{2\phi\Delta})}\xi_t,\ \xi_t \sim \mathcal{N}(0, 1)$$

whereas the Euler model gives

$$X_{t+\Delta} = X_t + \phi X_t \Delta + \sigma\sqrt{\Delta}\xi_t,\ \xi_t \sim \mathcal{N}(0, 1)\,.$$

The continuous time MLE is

$$\hat{\phi}_N = \frac{1}{N}\log\left(\frac{\sum_{i=0}^{N-1} X_{(i+1)\Delta} X_{i\Delta}}{\sum_{i=0}^{N-1} X_{i\Delta}^2}\right) \to \phi_0$$

$$\hat{\sigma}_N^2 = \frac{-2\hat{\phi}_N}{N(1 - \exp(2\Delta\hat{\phi}_N))} \sum_{i=0}^{N-1} (X_{(i+1)\Delta} - X_{i\Delta}\exp(\Delta\hat{\phi}_N))^2 \to \sigma_0^2$$

where ϕ_0 and σ_0^2 are the true values and the limit is $N \to \infty$ [Pedersen, 1995]. The MLE is asymptotically Normal with variance

$$V(\phi_0, \sigma_0^2, \Delta) = \frac{1}{N}\begin{pmatrix} \frac{1-e^{2\phi_0\Delta}}{\Delta^2 e^{2\phi_0\Delta}} & \frac{2\sigma_0^2}{\Delta} + \frac{\sigma_0^2}{\phi_0\Delta^2}\frac{(1-e^{2\phi_0\Delta})}{e^{2\phi_0\Delta}} \\ \frac{2\sigma_0^2}{\Delta} + \frac{\sigma_0^2}{\phi_0\Delta^2}\frac{(1-e^{2\phi_0\Delta})}{e^{2\phi_0\Delta}} & \frac{\sigma_0^4}{\phi_0^2\Delta^2}\frac{1-e^{2\phi_0\Delta}}{e^{2\phi_0\Delta}} + \frac{4\sigma_0^4}{\phi_0\Delta} + \frac{2\sigma_0^4(1+e^{2\phi_0\Delta})}{1-e^{2\phi_0\Delta}} \end{pmatrix}$$

The discrete time MLE is

$$\tilde{\phi}_n = \frac{1}{\Delta}\left(\frac{\sum_{i=0}^{N-1} X_{(i+1)\Delta}X_{i\Delta}}{\sum_{i=0}^{N-1} X_{(i=0}^{N-1}X_{i\Delta}^2} - 1\right) \to \frac{\exp(\Delta\phi_0)-1}{\Delta} > \phi_0$$

$$\tilde{\sigma}_N^2 = \frac{1}{N\Delta}\sum_{i=0}^{N-1}(X_{(i+1)\Delta} - X_{i\Delta})^2 \to \sigma_0^2\frac{1-\exp(\Delta\phi_0)}{-\Delta\phi_0} < \sigma_0^2\,.$$

Therefore, these estimators are not consistent. The sampling variance of this estimator can also be shown to be asymptotically Normal (see e.g Broze et al. [1998]) with variance

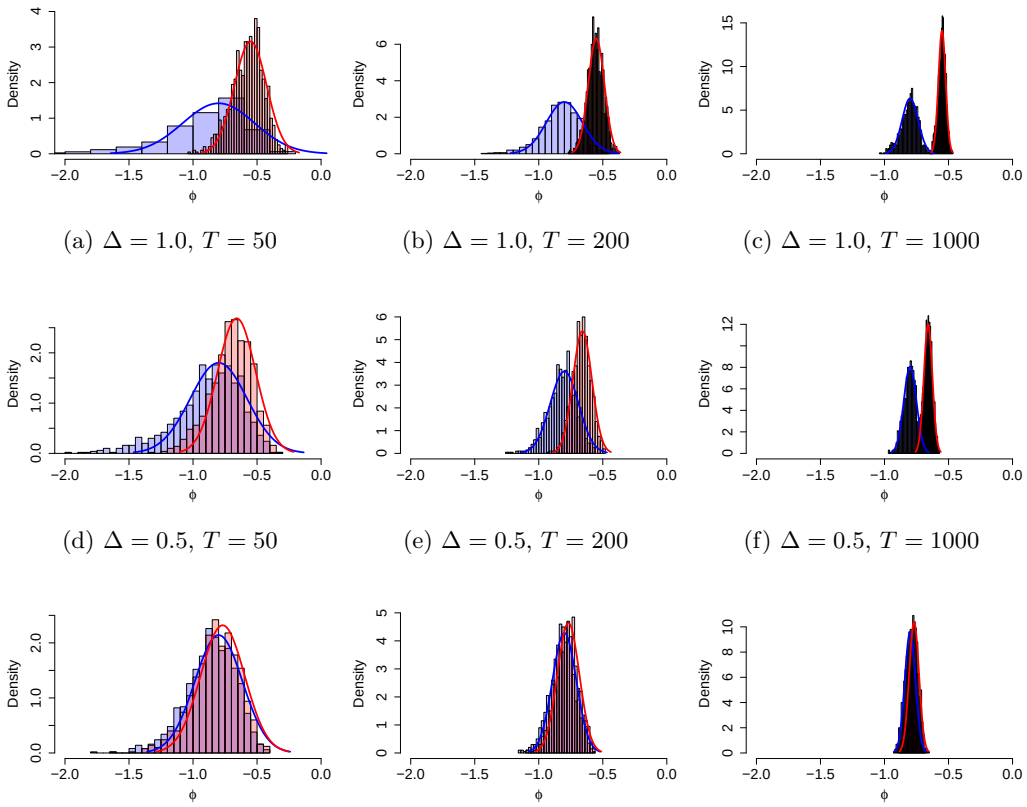
$$V_\Delta(\phi_0,\sigma_0^2,\Delta) = \frac{1}{N}\frac{1-\exp(2\phi_0\Delta)}{2\Delta^2\phi_0^2}\begin{pmatrix} 2\phi_0^2 & 0 \\ 0 & \sigma_0^4(1-\exp(2\phi_0\Delta)^2 \end{pmatrix}.$$

Figures 4.1 and 4.2 show the results of a simulation study comparing the continuous and discrete time MLEs for the parameters in Eq. (4.9). For each value of $\Delta \in \{0.1, 0.5, 1.0\}$ and $T \in \{50, 200, 1000\}$ we simulated Eq. (4.9) 1000 times and calculated both MLEs. We plotted the distribution of the continuous estimates in blue and the discrete in red. The theoretical asymptotic distributions are also displayed. For all cases the true values were $(\phi_0, \sigma_0^2) = (-0.8, 0.5)$.

For $\hat{\phi}$ the empirical distributions converge quickly to the asymptotic distributions, whereas there is some discrepancy between the empirical and asymptotic distributions for $\hat{\sigma}^2$. The distributions of the discrete MLE $\tilde{\phi}$ are biased until $\Delta \leq 0.1$. The problem of estimation based on the discrete model gets worse as T increases: there is no overlap between the distributions of $\hat{\phi}$ and $\tilde{\phi}$. Also, the distributions of $\tilde{\sigma}^2$ show that until $\Delta \leq 0.1$ the estimates are wrong.

In this thesis we assume that we have a fixed constant observation interval Δ. All the methods studied can be easily extended to a variable observation interval. We assume that Δ is large enough such that the naive maximum likelihood discussed above introduces significant errors.

Only for simple models like the Ornstein-Uhlenbeck process, Geometric Brownian motion and the Cox-Ingersol-Ross model are the continuous time transition densities available. In other cases the likelihood must be approximated. Figures 4.1 and 4.2 show that the simple Euler approximation is not sufficient even in simple models like the O-U process. It is likely to be much worse in nonlinear models. It is vital to improve upon this simple approximation. There are several ways of doing this, which we discuss in the next section.

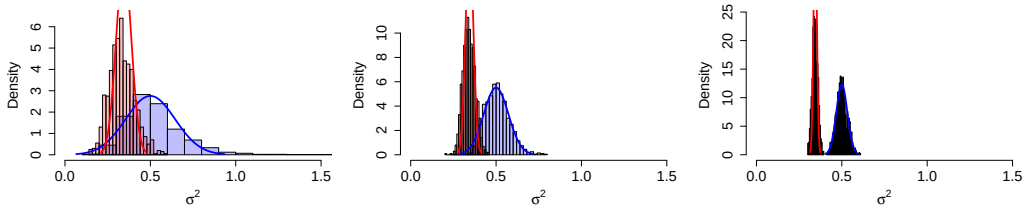


(a) $\Delta = 1.0$, $T = 50$ (b) $\Delta = 1.0$, $T = 200$ (c) $\Delta = 1.0$, $T = 1000$

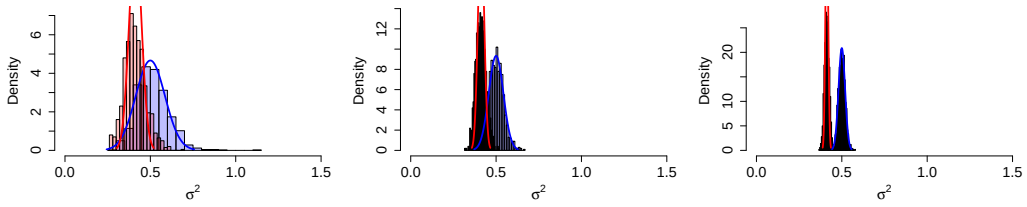
(d) $\Delta = 0.5$, $T = 50$ (e) $\Delta = 0.5$, $T = 200$ (f) $\Delta = 0.5$, $T = 1000$

(g) $\Delta = 0.1$, $T = 50$ (h) $\Delta = 0.1$, $T = 200$ (i) $\Delta = 0.1$, $T = 1000$

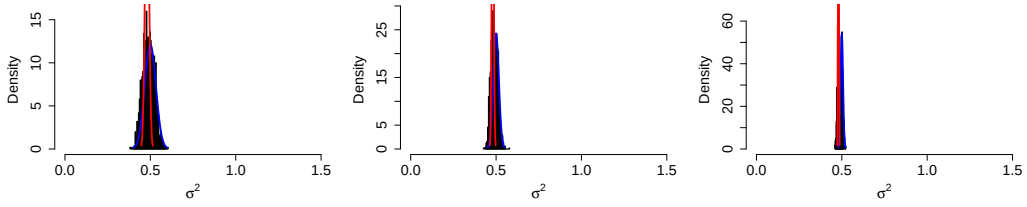
Figure 4.1: Maximum Likelihood Estimates for ϕ in the O-U process model Eq. (4.9). The blue (red) histogram are the estimates from the continuous (discrete) time model. The blue (red) curve is the asymptotic distribution of estimates of the continuous (discrete) time model. The true value is $\phi = -0.8$.



(a) $\Delta = 1.0$, $T = 50$ (b) $\Delta = 1.0$, $T = 200$ (c) $\Delta = 1.0$, $T = 1000$



(d) $\Delta = 0.5$, $T = 50$ (e) $\Delta = 0.5$, $T = 200$ (f) $\Delta = 0.5$, $T = 1000$



(g) $\Delta = 0.1$, $T = 50$ (h) $\Delta = 0.1$, $T = 200$ (i) $\Delta = 0.1$, $T = 1000$

Figure 4.2: Maximum Likelihood Estimates for σ^2 in the O-U process model Eq. (4.9). The blue (red) histogram are the estimates from the continuous (discrete) time model. The blue (red) curve is the asymptotic distribution of estimates of the continuous (discrete) time model. The true value is $\sigma^2 = 0.5$.

4.3 Approximations of the Likelihood Function

4.3.1 Numerical Solutions of the Fokker-Planck Equation

One method, to approximate the transition densities, is through the solution of the Fokker-Planck (FP) equation (see Section 2.4). LO [1988] applied this approach to geometric Brownian motion, jump processes and a system with an absorbing barrier. In these cases the FP equation could be solved analytically. They discuss the asymptotic consistency and normality of the maximum likelihood estimator $\hat{\boldsymbol{\theta}}$ so that in the limit of increasing observation period

$$\lim_{N\to\infty} \hat{\boldsymbol{\theta}} = \boldsymbol{\theta}, \;\; \sqrt{N}(\vec{\hat{\theta}} - \boldsymbol{\theta}) \sim \mathcal{N}(0, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))\,,$$

where the Fisher information matrix

$$\boldsymbol{I}(\boldsymbol{\theta}) = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left[-\frac{\partial p(\boldsymbol{x}_k|\boldsymbol{x}_{k-1};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} \right].$$

Maximum likelihood through numerical integration of the Fokker-Planck equation was investigated by Jensen and Poulsen [2002]. Details of the method are given in Hurn et al. [2007]: standard central difference formula lead to a system of tri-diagonal equations which must be solved at each value of θ. The problem with this approach is the need for a very fine spatial grid in order to approximate the initial condition, which is a delta function, as a smooth Gaussian. Care needs to be taken in including boundary conditions and large amounts of computation time are needed to repeatedly solve the PDE accurately. Hurn et al. [2007] suggest that a representation in terms of the Cumulative Density Function (CDF) can give better results for lower grid resolutions as the initial condition is now a step function. The obvious drawback is that this introduces higher order derivatives into the FP equation and would only be straight forward for a one dimensional problem.

Another approach is to solve the FP equation using spectral methods. Hurn and Lindsay [1999] use Chebyshev polynomials as basis functions leading to a system of ODEs that can be solved numerically. The benefit of this approach is the exponential rate of convergence in the number of Chebyshev polynomials although before one can apply this solution method a mapping to the interval $[-1, 1]$ must be found.

Hurn et al. [2003] use a simulation approach of the actual SDE to estimate the transition density. They integrate the SDE from initial condition x_{i-1} to some point y and then use Kernel density estimation to approximate the transition density.

Repeating this M times gives the estimator

$$\hat{p}^M(t_i, x_i|t_{i-1}, x_{i-1}; \theta) = \frac{1}{Mh} \sum_{j=1}^{M} K\left(\frac{x_i - y_j}{h}\right) ,$$

where h is the bandwidth and K is a kernel function. The same underlying Wiener process is used to estimate the density for each parameter.

Crommelin and Vanden-Eijnden [2006] use a non-parametric method based on the F-P equation. The drift and diffusion functions are optimised such that the eigenspectrum of the F.P operator is as close as possible to the eigenspectrum estimated from the observed data. This method is applied to the estimation of reduced climate models in Majda et al. [2009].

4.3.2 Particle Filters

If the diffusion path is updated at the same time as $\boldsymbol{\sigma}$ so that they are always consistent then the algorithm should not deteriorate for large m. However, due to updating such a large vector of variables, this is likely to lower the acceptance rate unless effective proposals are used. One method that aims to do this is the **particle filter** of Golightly and Wilkinson [2006a,b]. Their idea is to approximate the joint density of $\boldsymbol{X}_j$ and $\boldsymbol{\theta}$, given the observations up to that time, by a collection of particles $\{\boldsymbol{X}_j^{(i)}, \boldsymbol{\theta}^{(i)}\}_{i=1\ldots P}$ for some large P. This discrete distribution is then smoothed and used as an approximate prior when sampling the next path segment $\boldsymbol{X}_{j+1} \ldots \boldsymbol{X}_{j+m-1}$ and parameter $\boldsymbol{\theta}$, conditional on the next observation $\boldsymbol{X}_{j+m}$. In this way the diffusion and parameter are consistent and only one block of data need be sampled at once. P new paths are sampled using the Metropolis-Hastings algorithm, the end points of which are retained as the particles for the next path segment: $\{\boldsymbol{X}_{j+m}^{(i)}, \boldsymbol{\theta}^{(i)}\}_{i=1\ldots P}$. The algorithm can be applied to any multivariate diffusion and also has the benefit that it can be applied to **online estimation**: inference can include new data without reusing the whole data set. A downside to this approach is the need to have a large number of particles for the approximation and the need to choose some smoothing bandwidth that could propagate errors from one approximation to the next. There are guidelines for choosing the bandwidth but these may not be as reliable in high dimensional problems.

4.3.3 Importance Samplers

The inference method in this section covers some of the theoretical foundation for methods used later in this thesis and will be used to fix notation. Let $\boldsymbol{X}_{t_{im}} = \boldsymbol{x}_{t_{im}}$

be a sequence of observations of process 4.1 at times t_{im}, $i = 0, \ldots, N$ for integer m. Without loss of generality we consider equal observation times so that $t_{im} = im\Delta$ for fixed $\Delta > 0$.

The **Simulated Maximum Likelihood** of Pedersen [1995] is a method to estimate the transition density $p(t_{(i+1)m}, \boldsymbol{x}_{t_{(i+1)m}} | t_{im}, \boldsymbol{x}_{t_{im}}; \boldsymbol{\theta})$ between observations by partitioning the interval $\Delta = t_{(i+1)m} - t_{im}$ into m smaller subintervals $\delta = \Delta/m$. This introduces $m - 1$ unobserved random variables per observation interval. We write $\boldsymbol{X}_{im+k} = \boldsymbol{X}_{t_{im+k}}$, $i = 0 \ldots N$ and $k = 0 \ldots m$, where $t_{im+k} = t_{im} + k\delta$, to denote the complete data of observed and "missing" variables.

Consider a single observation interval. For notational convenience, and without loss of generality, this can be for $t_0 = 0$ to $t_m = \Delta$ so that $i = 0$. The transition probability for a fixed subinterval k can be approximated by the Euler-Maruyama discretisation

$$p^{(1)}(t_{k+1}, \boldsymbol{x}_{k+1} | t_k, \boldsymbol{x}_k; \boldsymbol{\theta}) \approx \phi(\boldsymbol{x}_{k+1} | \boldsymbol{x}_k + \delta \boldsymbol{\mu}_k, \delta \boldsymbol{\Sigma}_k) \,,$$

where $\boldsymbol{\mu}_k = \boldsymbol{\mu}(\boldsymbol{x}_k, \boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}(\boldsymbol{x}_k, \boldsymbol{\theta})$. Then the transition density for the complete interval can be approximated as

$$\begin{aligned} &p(t_m, \boldsymbol{x}_m | t_0, \boldsymbol{x}_0; \boldsymbol{\theta}) \\ &\approx p^{(m)}(t_m, \boldsymbol{x}_m | t_0, \boldsymbol{x}_0; \boldsymbol{\theta}) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \prod_{k=0}^{m-1} p^{(1)}(t_{k+1}, \boldsymbol{x}_{k+1} | t_k, \boldsymbol{x}_k; \boldsymbol{\theta}) d\boldsymbol{x}_1 \ldots d\boldsymbol{x}_{m-1} \,, \end{aligned} \tag{4.10}$$

where $\mathcal{X}$ is the domain of the process. Pedersen [1995] proves L^1 convergence of $p^{(m)}(t_m, \boldsymbol{x}_m | t_0, \boldsymbol{x}_0; \boldsymbol{\theta})$ to $p(t_m, \boldsymbol{x}_m | t_0, \boldsymbol{x}_0; \boldsymbol{\theta})$ as $m \to \infty$. Note that $\boldsymbol{x}_k$ are not observations: they are missing variables between observation times so must be integrated out.

Consider the case with $m = 2$ so that there is only one missing variable $\boldsymbol{x}_1$. The transition density in Eq. (4.10) is estimated using the **Chapman-Kolmogorov** equation

$$\int_{\mathcal{X}} p^{(1)}(t_2, \boldsymbol{x}_2 | t_1, \boldsymbol{x}_1; \boldsymbol{\theta}) p^{(1)}(t_1, \boldsymbol{x}_1 | t_0, \boldsymbol{x}_0; \boldsymbol{\theta}) d\boldsymbol{x}_1 = \mathbb{E}(p^{(1)}(t_2, \boldsymbol{x}_2 | t_1, \boldsymbol{x}_1; \boldsymbol{\theta})) \,, \tag{4.11}$$

where the expectation is with respect to density $p^{(1)}(t_1, \boldsymbol{x}_1 | t_0, \boldsymbol{x}_0; \boldsymbol{\theta})$. The proposed scheme is then based upon integrating out the unobserved paths of the diffusion by simulation. Pedersen [1995] suggests computing the expectation using the Monte

Carlo estimator of Eq. (4.11)

$$\hat{p}^{m,n}(t_2, \boldsymbol{x}_2, t_0, \boldsymbol{x}_0) = \frac{1}{n}\sum_{i=1}^{n} p^{(1)}(t_2, \boldsymbol{x}_2 | t_1, \boldsymbol{x}_1^{(i)}; \boldsymbol{\theta}) \,, \tag{4.12}$$

where, in this case, $m = 2$ and $\boldsymbol{x}_1^{(i)}$ is the end point sampled from the Euler-Maruyama scheme up to time t_1. This can be generalised to larger m, to estimate Eq. (4.10), so that the estimate converges to the true transition density. Pedersen [1995] prove convergence asymptotically but, in general, it is not known how large m should be in practice. To compute the MLE the transition density for all i is recalculated for multiple $\boldsymbol{\theta}$ and maximised.

In practice Eq. (4.12) may be a poor estimator due to the samples $\boldsymbol{x}_1^{(i)}$ having low mass under the density $p^{(1)}(t_2, \boldsymbol{x}_2 | t_1, \boldsymbol{x}_1^{(i)}; \boldsymbol{\theta})$, simply because $\boldsymbol{x}_1^{(i)}$ is not close to $\boldsymbol{x}_2$. Durham and Gallant [2002] propose techniques to reduce the variance of this sampler. They suggest generating samples from a process that is conditioned on the end point $\boldsymbol{X}_2 = \boldsymbol{x}_2$. This, bridge process, has density

$$\begin{aligned} p(\boldsymbol{x}_1 | \boldsymbol{x}_0 = \boldsymbol{x}_0, \boldsymbol{X}_2 = \boldsymbol{x}_2; \boldsymbol{\theta}) &\propto p(t_1, \boldsymbol{x}_1 | t_0, \boldsymbol{x}_0; \boldsymbol{\theta}) p(\boldsymbol{x}_2 | t_1, \boldsymbol{x}_1; \boldsymbol{\theta}) \\ &\approx \phi(\boldsymbol{x}_1; \boldsymbol{x}_0 + \tilde{\boldsymbol{\mu}}_0 \delta, \tilde{\boldsymbol{\Sigma}}_0 \delta) \,, \end{aligned}$$

where

$$\tilde{\boldsymbol{\mu}}_k = \frac{\boldsymbol{x}_m - \boldsymbol{x}_k}{t_m - t_k}, \quad \tilde{\boldsymbol{\Sigma}}_k = \left(\frac{t_m - t_{k+1}}{t_m - t_k} \right) \boldsymbol{\Sigma}_k \,. \tag{4.13}$$

Using this proposal distribution for $\boldsymbol{x}_1$ gives the **importance sampling** estimator

$$\hat{p}^{m,n}(t_2, \boldsymbol{x}_2, t_0, \boldsymbol{x}_0) = \frac{1}{n}\sum_{i=1}^{n} p^{(1)}(t_2, \boldsymbol{x}_2 | t_1, \boldsymbol{x}_1^{(i)}; \boldsymbol{\theta}) \rho(\boldsymbol{x}_1^{(i)}) \,, \tag{4.14}$$

where the weight function

$$\rho(\boldsymbol{x}_1^{(i)}) = \frac{\phi(\boldsymbol{x}_1^{(i)} | \boldsymbol{x}_0 + \boldsymbol{\mu}_0 \delta, \boldsymbol{\Sigma}_0 \delta)}{\phi(\boldsymbol{x}_1^{(i)} | \boldsymbol{x}_0 + \tilde{\boldsymbol{\mu}}_0 \delta, \tilde{\boldsymbol{\Sigma}}_0 \delta)} \tag{4.15}$$

corrects for the fact that we are not sampling from the Euler approximation.

For arbitrary m the importance weights are

$$\rho(\boldsymbol{x}_{m-1}^{(i)}) = \frac{\prod_{k=0}^{m-2} \phi(\boldsymbol{x}_{k+1}^{(i)} | \boldsymbol{x}_k + \boldsymbol{\mu}_k \delta, \boldsymbol{\Sigma}_k \delta)}{\prod_{k=0}^{m-2} \phi(\boldsymbol{x}_{k+1}^{(i)} | \boldsymbol{x}_k + \tilde{\boldsymbol{\mu}}_k \delta, \tilde{\boldsymbol{\Sigma}}_k \delta)} \tag{4.16}$$

Durham and Gallant [2002] note that, for general m, the importance sampler in Eq.

(4.14) is an estimate of

$$p^{(m)}(t_m, \boldsymbol{x}_m | t_0, \boldsymbol{x}_0; \boldsymbol{\theta}) = \int p^{(1)}(t_m, \boldsymbol{x}_m | \boldsymbol{x}_{m-1}, t_{m-1}) d\mathbb{P}^{(m)}(\boldsymbol{x}_{m-1}) \\ = \int p^{(1)}(t_m, \boldsymbol{x}_m | \boldsymbol{x}_{m-1}, t_{m-1}) \rho^{(m)}(\boldsymbol{x}_{m-1}) d\mathbb{Q}^{(m)}(\boldsymbol{x}_{m-1}) \,,$$

where $\rho^{(m)}(\boldsymbol{x}_{m-1})$ is the Radon-Nikodym derivative between $\mathbb{P}^{(m)}(\boldsymbol{x}_{m-1})$, the law of the Euler discretisation and $\mathbb{Q}^{(m)}(\boldsymbol{x}_{m-1})$, the law of the importance sampler. In the modified bridge sampler above the Radon-Nikodym derivative is given by the importance weights in Eq. (4.15). In continuous time, as $m \to \infty$, we have the expression

$$p(t_m, \boldsymbol{x}_m | t_0, \boldsymbol{x}_0; \boldsymbol{\theta}) = \int p(t_m, \boldsymbol{x}_m | \boldsymbol{x}_{m-1}, t_{m-1}) d\mathbb{P}(\boldsymbol{x}_{m-1}) \\ = \int p(t_m, \boldsymbol{x}_m | \boldsymbol{x}_{m-1}, t_{m-1}) \rho(\boldsymbol{x}_{m-1}) d\mathbb{Q}(\boldsymbol{x}_{m-1}) \,,$$

which is equivalent to Eq. (4.11). Now $\mathbb{P}(\boldsymbol{x}_{m-1})$ is the law of the continuous time process in Eq. (4.1) and $\mathbb{Q}(\boldsymbol{x}_{m-1})$ is the law of the continuous time importance sampler given by

$$d\tilde{\boldsymbol{X}}_t = \tilde{\boldsymbol{\mu}}(\tilde{\boldsymbol{X}}_t, \boldsymbol{\theta}) dt + \boldsymbol{a}(\tilde{\boldsymbol{X}}_t, \boldsymbol{\theta}) d\tilde{\boldsymbol{B}}_t \,. \tag{4.17}$$

The Radon-Nikodym derivative is given by Girsanov's theorem

$$\rho(\boldsymbol{X}_t) = \frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mathbb{Q}_{\boldsymbol{\theta}}}(\boldsymbol{X}_t) = \exp\left(\int_0^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_s, \boldsymbol{\theta}) \boldsymbol{b}(\boldsymbol{X}_s, \boldsymbol{\theta}) d\boldsymbol{X}_s \right. \\ \left. -\frac{1}{2} \int_0^t \boldsymbol{b}(\boldsymbol{X}_s, \boldsymbol{\theta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}_s, \boldsymbol{\theta}) \boldsymbol{b}(\boldsymbol{X}_s, \boldsymbol{\theta}) ds \right) \,, \tag{4.18}$$

where

$$\boldsymbol{b}(\boldsymbol{X}_s, \boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{X}_t, \boldsymbol{\theta}) - \tilde{\boldsymbol{\mu}}(\boldsymbol{X}_t, \boldsymbol{\theta}) \,. \tag{4.19}$$

In continuous time the modified bridge sampler is given by

$$d\tilde{\boldsymbol{X}} = \frac{\boldsymbol{x}_m - \tilde{\boldsymbol{X}}_t}{t_m - t} dt + \boldsymbol{a}(\tilde{\boldsymbol{X}}_t, \boldsymbol{\theta}) d\tilde{\boldsymbol{B}}_t \,, \;\; \tilde{\boldsymbol{X}}_m = \boldsymbol{x}_m \,, \tilde{\boldsymbol{X}}_0 = \boldsymbol{x}_0 \,. \tag{4.20}$$

It is vital that the importance sampler has the same diffusion function as the target process because, as mentioned in Section 2.5, probability laws for processes with different diffusion functions are mutually singular. This is due to the quadratic variation property: in continuous time the diffusion parameters can be perfectly determined by the quadratic variation of the process. The Radon-Nikodym derivative,

between processes with different diffusions, would not exist in the continuous time limit and the importance sampler weights would be meaningless. All algorithms should take account of this fact if they are to be meaningful in the continuous time limit. Proposals for missing data should be consistent with the diffusion function of the target process. Otherwise one would notice very poor convergence of the estimator as m increases. Durham and Gallant [2002] report dramatic decreases in variance when using the modified bridge over the Brownian bridge as expected.

4.3.4 Markov Chain Monte Carlo Methods

It may not always be possible to find an importance sampler that works well, especially in high dimensional problems. An alternative is to construct a Markov Chain that has invariant distribution equal to the distribution of the missing data. **Markov Chain Monte Carlo** (MCMC) is a widely used technique for sampling from high dimensional distributions. Subsequent samples are not i.i.d but effective algorithms can be designed to have low autocorrelation and fast mixing times for specific problems. A popular method is the Metropolis-Hastings algorithm (MH). Consider the problem of sampling a collection of missing data $\boldsymbol{X}$ according to a probability density $p(\boldsymbol{X})$. In the MH algorithm one first samples $\boldsymbol{X}^*$ from a simpler distribution $\boldsymbol{X^*} \sim q(\boldsymbol{X}^*|\boldsymbol{X})$, where q may depend on the current value of $\boldsymbol{X}$. This proposal is then accepted with probability

$$\alpha = \frac{p(\boldsymbol{X})q(\boldsymbol{X}|\boldsymbol{X}^*)}{p(\boldsymbol{X}^*)q(\boldsymbol{X}^*|\boldsymbol{X})}\,. \tag{4.21}$$

It can be shown that the resulting Markov Chain is reversible with respect to $p(\boldsymbol{X})$ and so leaves this distribution invariant. One need only then ensure ergodicity of the Markov Chain to guarantee convergence to the target (see e.g Robert and Casella [2005] and Gilks and Spiegelhalter [1996]).

In our problem the target density, with respect to Lebesgue measure, is the product in Eq. (4.10)

$$p(\boldsymbol{X}_1 \cdots \boldsymbol{X}_{m-1}|\boldsymbol{X}_0 = \boldsymbol{x}, \boldsymbol{X}_m = \boldsymbol{y}; \boldsymbol{\theta}) = \prod_{i=0}^{m-1} p^{(1)}(\tau_{k+1}, \boldsymbol{x}_{k+1}|\tau_k, \boldsymbol{x}_k; \boldsymbol{\theta})\,, \tag{4.22}$$

where $p^{(1)}$ is approximated by the Euler transition density given in Eq. (4.7). Equivalently we can write the target as a density with respect to a **dominating measure** $\mathbb{Q}$, which is the law of a driftless version of the process

$$d\boldsymbol{X}_t = \boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\theta})d\boldsymbol{B}_t\,,\ \boldsymbol{X}_0 = \boldsymbol{x}_0\,,\ t \in [0, T]\,. \tag{4.23}$$

Then $\mathbb{Q}$ is the law of the **local martingale** and the target density is given by Girsanov's formula Eq. (4.18).

Consider now the Bayesian problem where we specify priors on $\boldsymbol{\theta}$ and wish to perform inference based on the **posterior distribution**. We augment the observed data $\boldsymbol{X}_{\text{obs}}$ with missing data $\boldsymbol{X}_{\text{miss}}$ as discussed in Section 4.3.3 in order to facilitate accurate computation of the likelihood function. The target distribution for our MCMC algorithm is the observed data posterior for $p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}})$ but what we can evaluate is the full data posterior $p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{miss}})$. If we alternate between sampling values of missing data from $p(\boldsymbol{X}_{\text{miss}}|\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta})$ then the parameters from $p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{miss}})$ we can average over the missing data to estimate the observed data posterior. This type of data augmentation algorithm was shown to converge by Tanner and Wong [1987].

In the context of MCMC this is an example of the **Gibbs algorithm** and is central to the inference in this thesis. Sahu and Roberts [1999] showed that the convergence properties of the Gibbs algorithm are closely related to those of the EM algorithm for maximum likelihood. Meng and vanDyk [1997] discuss how the rate of convergence of the EM algorithm is closely related to the ratio of information in the observed data to missing data. Roberts and Stramer [2001] highlight how this fact is very important in the current context of inference for diffusions. The problem stems from the quadratic variation relation in Eq. (4.2) which states that there is an infinite amount of information about the diffusion parameters if it is observed continuously. Therefore, there is complete dependence between the missing data and the diffusion parameters. Updating one just confirms the value of the other. This phenomenon has been the focus of much research into SDE inference. The slowing down of the Gibbs sampler as the amount of missing data increases was noted by Elerian et al. [2001] in a simulation study and predicted from a theoretical viewpoint by Roberts and Stramer [2001]. An algorithm that alternates between sampling $\boldsymbol{X}_{\text{miss}}$ and $\boldsymbol{\sigma}$ will break down as $m \to \infty$, where $m-1$ is the amount of missing data per observation interval as in the importance samplers discussed in Section 4.3.3. This is because the dominating measure $\mathbb{Q}$ in Girsanov's theorem Eq. (4.18) depends on the value of $\boldsymbol{\sigma}$ and the measures $\mathbb{Q}_{\boldsymbol{\sigma}}$ and $\mathbb{Q}_{\boldsymbol{\sigma}^*}$ are mutually singular for $\boldsymbol{\sigma} \neq \boldsymbol{\sigma}^*$. Using Eq. (4.18) as a likelihood for $\boldsymbol{\sigma}$ would result in a posterior which is just a point mass on the value given by the quadratic variation. In the following sections we review some approaches that attempt to overcome this problem.

Reparametrisation of the Dominating Measure

In the practical implementation of a discrete approximation to Eq. (4.18), Roberts and Stramer [2001] argue that the mixing time of the algorithm would be $O(m)$. They suggest a reparametrisation which is equivalent to refactorising the dominating measure so that it is independent of $\boldsymbol{\sigma}$. Firstly the data is transformed via the Lamperti transformation $\dot{\boldsymbol{X}} = \boldsymbol{h}(\boldsymbol{X})$, where the function $\boldsymbol{h}(\boldsymbol{X})$ satisfies

$$\frac{d\boldsymbol{h}}{d\boldsymbol{x}} = \boldsymbol{a}(\boldsymbol{X}, \boldsymbol{\sigma})^{-1} \tag{4.24}$$

to the unit diffusion process

$$d\dot{\boldsymbol{X}}_t = \boldsymbol{b}(\dot{\boldsymbol{X}}, \boldsymbol{\theta}) + d\boldsymbol{B}_t\,, \tag{4.25}$$

where

$$\boldsymbol{b} = \boldsymbol{a}(h_{\boldsymbol{\theta}}^{-1}(\dot{\boldsymbol{X}}), \boldsymbol{\sigma})^{-1}\boldsymbol{\mu}(h_{\boldsymbol{\theta}}^{-1}(\dot{\boldsymbol{X}}), \boldsymbol{\theta}) - \frac{1}{2}\frac{d\boldsymbol{a}(h_{\boldsymbol{\theta}}^{-1}(\dot{\boldsymbol{X}}), \boldsymbol{\theta})}{d\boldsymbol{x}}\,. \tag{4.26}$$

Eq. (4.25) does not suffer due to the dependence on diffusion parameters $\boldsymbol{\sigma}$. The second step is to transform the data such that $\boldsymbol{X}_0 = \boldsymbol{X}_m = 0$. The unique linear transformation that does this is

$$\hat{\boldsymbol{X}}_j = \boldsymbol{\eta}(\dot{\boldsymbol{X}}) = \dot{\boldsymbol{X}}_j + \frac{(m-j)\dot{\boldsymbol{X}}_0 + (j-m)\dot{\boldsymbol{X}}_m}{m} \tag{4.27}$$

According to Girsanov's change of measure theorem, under the dominating measure, each interval of missing data is a Brownian bridge process. We write this law as $\mathbb{B} = \otimes_{i=1}^{N-1}\mathbb{B}(t_{i+1}, 0|t_i, 0)$, where $\mathbb{B}(t, 0|s, 0)$ is the law of the standard **Brownian Bridge**. The likelihood of the missing data in Eq. (4.18) is then

$$\frac{d\mathbb{P}_{\boldsymbol{\theta}}}{d\mathbb{B}}(\hat{\boldsymbol{X}}_{\text{miss}}|\dot{\boldsymbol{X}}_{\text{obs}}) \propto G(\boldsymbol{\eta}^{-1}(\hat{\boldsymbol{X}}), \boldsymbol{b}, 1) \tag{4.28}$$

so that the dominating measure is independent of $\boldsymbol{\sigma}$.

In terms of the transformed data the conditional posterior for $\boldsymbol{\sigma}$ can be written

$$p(\boldsymbol{\sigma}|\hat{\boldsymbol{X}}, \boldsymbol{X}_{\text{obs}}, \boldsymbol{\gamma}) \propto G(\boldsymbol{\eta}^{-1}(\hat{\boldsymbol{X}}), \boldsymbol{b}, 1) f(t, \boldsymbol{X}_{\text{obs}}, \boldsymbol{\sigma}) p(\boldsymbol{\sigma})\,, \tag{4.29}$$

where $f(t, \boldsymbol{X}_{\text{obs}}, \boldsymbol{\sigma})$ is the density of the observations with respect to Lebesgue measure under the dominating measure. For example, for a one dimensional SDE

with constant σ this is the Gaussian density

$$f(t, X_{\text{obs}}, \sigma) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{N-1} \frac{(X_{i+1} - X_i)}{2(t_{i+1} - t_i)} \right) .$$

Updating of $\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$ can be achieved using Metropolis-Hastings steps within the Gibbs sampler. For updating the missing data, Roberts and Stramer [2001] suggest using an independence sampler which is absolutely continuous with respect to the target

$$d\dot{\boldsymbol{Z}} = \boldsymbol{\xi}(\dot{\boldsymbol{Z}})dt + d\boldsymbol{B}_t \,, \tag{4.30}$$

where there are several possibilities for the drift function $\boldsymbol{\xi}(\cdot)$, then transforming to $\hat{\boldsymbol{Z}} = \boldsymbol{\eta}(\dot{\boldsymbol{Z}})$ to agree with the end point conditions $\hat{\boldsymbol{Z}}_0 = \hat{\boldsymbol{Z}}_m = 0$. A new segment of data is accepted with the MH acceptance probability

$$\alpha = \min\left(1, \frac{G(\boldsymbol{\eta}^{-1}(\hat{\boldsymbol{Z}}), \boldsymbol{b}, 1) G(\boldsymbol{\eta}^{-1}(\hat{\boldsymbol{X}}), \boldsymbol{\xi}, 1)}{G(\boldsymbol{\eta}^{-1}(\hat{\boldsymbol{X}}), \boldsymbol{b}, 1) G(\boldsymbol{\eta}^{-1}(\hat{\boldsymbol{Z}}), \boldsymbol{\xi}, 1)} \right) . \tag{4.31}$$

The simplest method is to choose $\boldsymbol{\xi} = 0$ then the proposal is just a Brownian bridge. Roberts and Stramer [2001] also suggest a linear process.

We need to be able to simulate from $\dot{\boldsymbol{X}}$ easily, which in practice restricts us to using linear SDEs only. In Chapter 5 we investigate the efficiency of different choices for $\boldsymbol{\xi}$ in the independence sampler albeit in the context of a different algorithm. In particular we demonstrate the efficiency of the linear approximation for multivariate processes.

Roberts and Stramer [2001] implement the algorithm on a one dimensional cubic model and the Cox-Ingersol-Ross model for interest rates [Cox et al., 1985]. They show that the posterior distributions converge for increasing m and that the algorithm does not deteriorate. The algorithm is applied to a multivariate problem by Kalogeropoulos et al. [2011]. The algorithm is well motivated theoretically but unfortunately relies upon the existence of a transformation to unit volatility: a solution to Eq. (4.24). Ait-Sahalia [2008] showed that a necessary and sufficient condition is given by Eq. (2.23). This greatly restricts the range of models and in fact the form of model, for which we argued in Chapter 3, is outside of this class: the diffusion matrix should be linear in the state variables. To satisfy Eq. (2.23) we would have to set all off diagonal terms of the diffusion matrix to zero and the diagonal terms $a_{ii}(\boldsymbol{X}, \boldsymbol{\sigma}) = a_{ii}(X_i, \boldsymbol{\sigma})$. This would seem overly restrictive.

Time Change Transformation

An alternative reparametrisation that aims to overcome the dependence between the missing data and quadratic variation is the time change transformations of Kalogeropoulos et al. [2010]. Again the aim is to write the likelihood with a parameter free dominating measure so that the diffusion parameters do not enter Girsanov's formula. For constant diffusion σ this is relatively straightforward. A new time scale is defined as

$$s = \eta_1(t, \sigma) = \int_0^t \sigma^2 d\omega = t\sigma^2 \,.$$

Then transform the SDE via

$$U_s = \begin{cases} X_{\eta_1^{-1}(s,\sigma)}, & 0 \leq s \leq \sigma^2 \\ M_{\eta_1^{-1}(s,\sigma)}, & s > \sigma^2 \end{cases} ,$$

where M is the corresponding driftless process. The new SDE is

$$dU_s = \begin{cases} \frac{\mu(U_s,\theta)}{\sigma^2} dt + dB_s^U & 0 \leq s \leq \sigma^2 \\ dW_s^U & s > \sigma^2 \end{cases} ,$$

This has transformed the SDE to one of unit diffusion although there is still dependence on σ^2 through the end point condition $U_{\sigma^2} = y_1$, where y_1 is the endpoint observation. Like in the two step transformation of Roberts and Stramer [2001] a second transformation is needed. Kalogeropoulos et al. [2010] suggest changing to a third time scale via

$$u = \eta_2(s, \sigma) = \frac{s}{\sigma^2(\sigma^2 - s)}$$

and defining a new process Z via

$$U_s = (\sigma^2 - s) Z_{\eta_2(s,\sigma)} + \left(1 - \frac{s}{\sigma^2}\right) y_0 + \frac{s}{\sigma^2} y_1, \; 0 \leq t < \sigma^2 \,,$$

where y_0 is the initial observation. Then Z is given by the SDE

$$dZ_u = \left(\frac{\mu(U_s, \theta)}{1 + \sigma^2} + U_2 \right) dt + dB_u^Z$$

where B^Z is a Brownian motion on the new timescale, which runs from 0 to ∞. The two transformations imply that σ does not enter into Girsanov's formula.

In the practical implementation of this algorithm each time σ is updated a new time scale is defined. This means that new values of the process Z must

be imputed between current values. This is likely to lower the acceptance rates although Kalogeropoulos et al. [2010] report that the algorithm is effective for large m. The algorithm can also be applied to cases where the noise depends upon an unobserved process as in the case of stochastic volatility models although the time scale is also state dependent. It is not clear whether the algorithm could be adapted to apply to more general processes. For example, if the diffusion depends upon the state of the observed variable then the Lamperti transform should be applied first. As in Roberts and Stramer [2001] this greatly limits the scope of the algorithm.

Innovation Scheme

Another algorithm that ensures consistency between the parameters and path was suggested initially in Chib et al. [2004] and further developed by Golightly and Wilkinson [2008]. The idea is to overcome the dependency between the diffusion parameters and the missing data by changing variables to the underlying Brownian motion $\boldsymbol{B} \in \mathbb{R}^d$ and conditioning on this, rather than $\boldsymbol{X}$, when performing the parameter update.

Assuming we have the SDE in Eq. (4.1) and $\boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\sigma}) \in \mathbb{R}^{d\times d}$ is invertible in its first argument then

$$d\boldsymbol{B}_t = \boldsymbol{a}^{-1}(\boldsymbol{X}_t, \boldsymbol{\sigma})(d\boldsymbol{X}_t - \boldsymbol{\mu}(\boldsymbol{X}_t, \boldsymbol{\theta})dt) \tag{4.32}$$

is a $d-$dimensional Brownian motion. Let this map be $\boldsymbol{X}_t = \boldsymbol{h}(\boldsymbol{B}_t, \boldsymbol{\sigma})$, then the conditional density of the parameters given the data is transformed via the Jacobian determinant as

$$p(\boldsymbol{\sigma}|\boldsymbol{X}, \boldsymbol{\theta}) = p(\boldsymbol{\sigma}|\boldsymbol{h}(\boldsymbol{B}, \boldsymbol{\sigma}), \boldsymbol{\theta}) \left| \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{B}}(\boldsymbol{B}, \boldsymbol{\sigma}) \right| , \tag{4.33}$$

where $|\cdot|$ denotes the determinant. When updating $\boldsymbol{\sigma}$ one first calculates the Brownian motion using $\boldsymbol{B} = \boldsymbol{h}^{-1}(\boldsymbol{X}, \boldsymbol{\sigma})$ then proposes a new $\boldsymbol{\sigma}^* \sim q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma})$. The conditional density is now

$$p(\boldsymbol{\sigma}^*|\boldsymbol{X}^*, \boldsymbol{\theta}) = p(\boldsymbol{\sigma}^*|\boldsymbol{h}(\boldsymbol{B}, \boldsymbol{\sigma}^*), \boldsymbol{\theta}) \left| \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{B}}(\boldsymbol{B}, \boldsymbol{\sigma}^*) \right| . \tag{4.34}$$

and so the data is updated as $\boldsymbol{X}^* = \boldsymbol{h}(\boldsymbol{B}, \boldsymbol{\sigma}^*)$.

In practice a discrete approximation

$$\boldsymbol{B}_{j+1} = \boldsymbol{B}_j + \boldsymbol{a}^{-1}(\boldsymbol{X}_j, \boldsymbol{\sigma})(\boldsymbol{X}_{j+1} - \boldsymbol{X}_j - \boldsymbol{\mu}(\boldsymbol{X}_j, \boldsymbol{\theta})\Delta t_j) \tag{4.35}$$

will be used. As discussed by Dargatz [2010] this mapping does not take account

of the end point conditions so that the reverse mapping may not be consistent with the observation. If $\boldsymbol{x}$ is observed then $\boldsymbol{h}^{-1}(\boldsymbol{h}(\boldsymbol{x},\boldsymbol{\sigma}),\boldsymbol{\sigma}^*) \neq \boldsymbol{x}$ for $\boldsymbol{\sigma} \neq \boldsymbol{\sigma}^*$.

A better idea is to define the process $\boldsymbol{Z}$, which conditions on the endpoint $\boldsymbol{x}_T$, by

$$d\boldsymbol{Z}_t = \boldsymbol{a}^{-1}(\boldsymbol{X}_t,\boldsymbol{\sigma})\left(d\boldsymbol{X}_t - \frac{\boldsymbol{x}_T - \boldsymbol{X}_t}{T-t}dt\right),\ \boldsymbol{Z}_0 = 0 \tag{4.36}$$

where T is the next observation time. If $\boldsymbol{Z}$ was Brownian motion then

$$d\boldsymbol{X}_t = \boldsymbol{\sigma}(\boldsymbol{X}_t,\boldsymbol{\sigma})d\boldsymbol{Z}_t + \frac{\boldsymbol{x}_T - \boldsymbol{X}_t}{T-t}dt,\ \boldsymbol{X}_0 = \boldsymbol{x}_0 \tag{4.37}$$

would be the continuous time version of the modified bridge as in Eq. (4.20). $\boldsymbol{Z}$ is not Brownian motion, however, but it has unit diffusion and so is absolutely continuous with respect to Brownian motion as shown by Dargatz [2010]. Therefore, it has the desired property that densities with respect to $\boldsymbol{Z}$ will have parameter free dominating measure.

In discrete time the transformation is

$$\boldsymbol{X}_{i+1} = \boldsymbol{X}_i + \boldsymbol{a}(\boldsymbol{X}_i,\boldsymbol{\sigma})(\boldsymbol{Z}_{i+1} - \boldsymbol{Z}_i) + \frac{\boldsymbol{x}_T - \boldsymbol{X}_i}{m-i}\,. \tag{4.38}$$

Let $\boldsymbol{X}_t = \boldsymbol{g}(\boldsymbol{Z}_t,\boldsymbol{\sigma})$ denote the transformation of all of the data. The Jacobian determinant is equal to the product

$$\left|\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{Z}}\right| = \prod_{i=1}^{N-1} |\boldsymbol{a}(\boldsymbol{X}_i,\boldsymbol{\sigma})|\,. \tag{4.39}$$

We use Eqns. (4.38) and (4.39) to write down an algorithm that does not degenerate as m increases. It is based on the reparametrisation of the conditional density of the data

$$p(\boldsymbol{X}|\boldsymbol{\theta},\boldsymbol{\sigma}) = p(\boldsymbol{g}(\boldsymbol{Z},\boldsymbol{\sigma})|\boldsymbol{\theta},\boldsymbol{\sigma})\left|\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{Z}}\right|\,.$$

The density on the right hand side can be written as Radon-Nikodym derivative of the law of $\boldsymbol{Z}$ with respect to a Brownian motion and so the dominating measure is parameter free [Dargatz, 2010]. This means that the Metropolis-Hastings acceptance probability will have non zero numerator and denominator.

We sample $\boldsymbol{\sigma}$ according to Algorithm 4.1. As this algorithm is central to this thesis we give details useful for implementation in computer software. We use zero-based numbering and consider there to be N observations and therefore, $N-1$ observation intervals indexed $0 \ldots N-2$. We assume interobservation times Δ are all equal and that there are $m-1$ imputed points per interval, giving a time interval

$\delta = \Delta/m$. We use the notation $\boldsymbol{X}_i = \boldsymbol{X}_{t_i}$ and $\boldsymbol{\mu}_i = \boldsymbol{\mu}(\boldsymbol{X}_{t_i}, \boldsymbol{\theta})$. The extension to variable interobservation times is straight forward. For simplicity, we write the algorithm for perfect observation of the system so that $\boldsymbol{X}_{im}, i = 0, \ldots, N-1$ are fixed.

Algorithm 4.1 Sample parameters entering the diffusion function.

Draw $\boldsymbol{\sigma}^* \sim q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma})$
Initialise $\alpha = \log(q(\boldsymbol{\sigma}|\boldsymbol{\sigma}^*)) - \log(q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma})) + \log(p(\boldsymbol{\sigma}^*)) - \log(p(\boldsymbol{\sigma}))$
for $i = 0$ to $N-2$ **do**
  for $j = 0$ to $m-2$ **do**

$$\boldsymbol{Z}_{im+j+1} = \boldsymbol{Z}_{im+j} + \boldsymbol{a}^{-1}(\boldsymbol{X}_{im+j}, \boldsymbol{\sigma}) \left(\boldsymbol{X}_{im+j+1} - \boldsymbol{X}_{im+j} - \frac{\boldsymbol{X}_{im+m} - \boldsymbol{X}_{im+j}}{m-j} \right)$$

$$\boldsymbol{X}^*_{im+j+1} = \boldsymbol{X}^*_{im+j} + \frac{\boldsymbol{X}_{im+m} - \boldsymbol{X}^*_{im+j}}{m-j} + \boldsymbol{a}(\boldsymbol{X}^*_{im+j}, \boldsymbol{\sigma}^*)(\boldsymbol{Z}_{im+j+1} - \boldsymbol{Z}_{im+j})$$

$$\begin{aligned} \alpha = \alpha &+ \log(\phi(\boldsymbol{X}^*_{im+j+1}; \boldsymbol{X}^*_{im+j} + \boldsymbol{\mu}^*_{im+j}\delta, \delta\boldsymbol{\Sigma}^*_{im+j})) + \log|a(\boldsymbol{X}^*_{im+j}, \boldsymbol{\sigma}^*)| \\ &- \log(\phi(\boldsymbol{X}_{im+j+1}; \boldsymbol{X}_{im+j} + \boldsymbol{\mu}_{im+j}\delta, \delta\boldsymbol{\Sigma}_{im+j}) - \log|a(\boldsymbol{X}_{im+j}, \boldsymbol{\sigma})| \end{aligned}$$

  end for
end for
Set $\{\boldsymbol{\sigma}, \boldsymbol{X}\} = \{\boldsymbol{\sigma}^*, \boldsymbol{X}^*\}$ with probability $\min(1, \exp(\alpha))$ else retain $\{\boldsymbol{\sigma}, \boldsymbol{X}\}$

In Algorithm 4.1 we have written the likelihood in terms of the Gaussian approximation to the transition density using the Markov property. As mentioned at the start of this section this is proved to converge to the true likelihood by Pedersen [1995].

To update missing data between observations we use an independence sampler as in Roberts and Stramer [2001] using the proposal process

$$d\boldsymbol{X}^* = \boldsymbol{\xi}(\boldsymbol{X}^*, \boldsymbol{X}_T)dt + \boldsymbol{a}(\boldsymbol{X}^*, \boldsymbol{\sigma})d\boldsymbol{B}^*_t\,, \tag{4.40}$$

where $\boldsymbol{X}_T$ is the next observation. This process will have measure that is absolutely continuous with respect to the target process in Eq. (4.1) because of their common diffusion function. However, we have a choice of drift function $\boldsymbol{\xi}$. The major restrictions being that it should be efficient to simulate from, its probability densities should be available explicitly and it should form a bridge between start and end observations. The simplest choice is just the modified bridge of Durham and Gallant [2002]. We investigate the benefits of using more sophisticated drift functions that approximate the target process in Chapter 5.

To update all of the missing data one proposes a block at a time from Eq.

(4.40) and then accepts this according to the MH ratio. If the inter observation interval is large then the acceptance rate may become very low and so one may subsample smaller blocks. However, Shephard and Pitt [1997] and Elerian [1999] showed that smaller block sizes generally lead to higher autocorrelation in the Markov Chain. This can be proved for Gaussian models but must be investigated case by case generally. In Chapter 5 we find that the efficiency of algorithms becomes very low when subsampling blocks of data for the case of cubic models. One also has the option of updating more than one block of missing data at a time: this could lead to increases in efficiency but for simplicity we give implementation details for the case of single block updates.

For some interval i we set $\boldsymbol{X}_0^* = \boldsymbol{X}_{im}$ and $\boldsymbol{X}_m^* = \boldsymbol{X}_{(i+1)m}$ then we propose $\boldsymbol{X}_1^* : \boldsymbol{X}_{m-1}^*$ and accept or reject using the Metropolis-Hastings acceptance probability

$$\alpha = \frac{p_\delta(\boldsymbol{X}_m^*|\boldsymbol{X}_{m-1}^*,\boldsymbol{\theta})\prod_{j=0}^{m-2} p_\delta(\boldsymbol{X}_{j+1}^*|\boldsymbol{X}_j^*,\boldsymbol{\theta})q_\delta(\boldsymbol{X}_{im+j+1}|\boldsymbol{X}_{im+j},\boldsymbol{\xi},\boldsymbol{\sigma})}{p_\delta(\boldsymbol{X}_{(i+1)m}|\boldsymbol{X}_{im+m-1},\boldsymbol{\theta})\prod_{j=0}^{m-2} p_\delta(\boldsymbol{X}_{im+j+1}|\boldsymbol{X}_{im+j},\boldsymbol{\theta})q_\delta(\boldsymbol{X}_{j+1}^*|\boldsymbol{X}_j^*,\boldsymbol{\xi},\boldsymbol{\sigma})}, \tag{4.41}$$

where p_δ is the transition density of the target Eq. (4.1) over time interval δ and q_δ is the transition density of the proposal. We choose proposal processes so that given $\boldsymbol{X}_j^*$, $\boldsymbol{X}_{j+1}^*$ is approximately Gaussian distributed. Eq. (4.40) is not a true Gaussian process because of the state dependent noise term. Details for updating the missing data are given in Algorithm 4.2.

In Chapter 5 we investigate the efficiency of different choices of $\boldsymbol{\xi}(\boldsymbol{X}^*, \boldsymbol{X}_T)$. In particular we develop the time dependent linear bridge and demonstrate that it is more effective than the standard modified bridge of Durham and Gallant [2002] at least for highly nonlinear processes.

Algorithms 4.1 and 4.2 are combined with standard Metropolis-Hastings updates for the parameters $\boldsymbol{\theta}$ entering into the drift function. One could use Random-Walk proposals but in our case of polynomial models it is more efficient to implement another Gibbs sampling step. This is described in Chapter 5. Repeatedly alternating between these three steps will produce MCMC samples that can be used to estimate the parameters. In practice we increase the amount of missing data m until we see convergence in the marginal distributions of the parameters.

4.3.5 Analytical Approximations of the Likelihood Function

Ait-Sahalia [2002] devised a method of approximating the transition densities $p_X(\boldsymbol{X}_{t+\delta}|\boldsymbol{X}_t,\boldsymbol{\theta})$ in the likelihood Eq. (4.5) analytically. The idea is to use a con-

Algorithm 4.2 Sample missing data between observations.

for $i = 0$ to $N - 2$ **do**
  Set $\boldsymbol{X}_0^* = \boldsymbol{X}_{im}$
  Set $\alpha = 0$
  for $j = 0$ to $m - 2$ **do**
    $\boldsymbol{X}_{j+1}^* \sim q_\delta(\boldsymbol{X}_{j+1}^* | \boldsymbol{\xi}(\boldsymbol{X}_j^*, \boldsymbol{X}_{im+m}), \boldsymbol{\sigma})$

$$\begin{aligned}\alpha =& \alpha + \log(\phi(\boldsymbol{X}_{j+1}^*; \boldsymbol{X}_j^* + \delta\boldsymbol{\mu}_j^*, \delta\boldsymbol{\Sigma}_j^*) \\ &+ \log(q_\delta(\boldsymbol{X}_{im+j+1} | \boldsymbol{\xi}(\boldsymbol{X}_{im+j}, \boldsymbol{X}_{im+m}), \boldsymbol{\sigma})) \\ &- \log(\phi(\boldsymbol{X}_{im+j+1}; \boldsymbol{X}_{im+j} + \delta\boldsymbol{\mu}_{im+j}, \delta\boldsymbol{\Sigma}_{im+j})) \\ &- \log(q_\delta(\boldsymbol{X}_{j+1}^* | \boldsymbol{X}_j^*, \boldsymbol{\xi}(\boldsymbol{X}_j^*, \boldsymbol{X}_{im+m}), \boldsymbol{\sigma}))\end{aligned}$$

  end for
  $\alpha = \alpha + \log(\phi(\boldsymbol{X}_{im+m}; \boldsymbol{X}_{m-1}^* + \delta\boldsymbol{\mu}_{m-1}^*, \delta\boldsymbol{\Sigma}_{m-1}^*))$
  $- \log(\phi(\boldsymbol{X}_{im+m}; \boldsymbol{X}_{im+m-1} + \delta\boldsymbol{\mu}_{im+m-1}, \delta\boldsymbol{\Sigma}_{im+m-1}))$
  if $\exp(\alpha) > \mathcal{U}(0, 1)$ **then**
    for $j = 0$ to $m - 2$ **do**
      $\boldsymbol{X}_{im+j+1} = \boldsymbol{X}_{j+1}^*$
    end for
  end if
end for

vergent series that adds corrections to the initial Gaussian approximation of the density. The first step is again to Lamperti transform the equation for $\boldsymbol{X}$ to one for $\boldsymbol{Y}$. Then the tails of the transition density $p_Y(\boldsymbol{Y}_{t+\delta} | \boldsymbol{Y}_t, \boldsymbol{\theta})$ are "light" enough for the series to converge. A second transformation to $\boldsymbol{Z}$ centres and normalises so that $\boldsymbol{Z} = (\boldsymbol{Y} - \boldsymbol{Y}_0)/\sqrt{\Delta}$, where $\boldsymbol{Y}_0$ is the peak of $\boldsymbol{Y}$, so that the density is close to a $\mathcal{N}(0, 1)$ variable. Then a Hermite series expansion is used to approximate $p_Z(\boldsymbol{Z}_{t+\Delta} | \boldsymbol{Z}_t, \boldsymbol{\theta})$.

The Hermite polynomials are given by

$$H_j(z) = e^{z^2/2} \frac{d^j}{dz^j} \left(e^{-z^2/2} \right) ,$$

The density is approximated as

$$p_Z^{(J)}(\boldsymbol{Z}_{t+\Delta} | \boldsymbol{Z}_t, \boldsymbol{\theta}) = \phi(\boldsymbol{Z}_{t+\Delta}) \sum_{j=0}^{J} \eta_Z^{(j)} H_j(\boldsymbol{Z}_{t+\Delta}) , \tag{4.42}$$

where ϕ is the standard Gaussian density and the coefficients are

$$\eta_Z^{(j)} = 1/j! \int_{-\infty}^{\infty} H_j(Z_{t+\Delta}) p_Z(Z_{t+\Delta}|Z_t;\boldsymbol{\theta}) dZ_{t+\Delta}\,. \tag{4.43}$$

Eq. (4.42) indicates that the first term in the expansion is just the Gaussian density.

Ait-Sahalia [2002] show that the back transformed density converges uniformly

$$p_X^{(J)}(\boldsymbol{X}_{t+\Delta}|\boldsymbol{X}_t;\boldsymbol{\theta}) \underset{J\to\infty}{\longrightarrow} p_X(\boldsymbol{X}_{t+\Delta}|\boldsymbol{X}_t;\boldsymbol{\theta})$$

and they demonstrate that retaining as few as 2 or 3 terms in the expansion can lead to maximum likelihood estimates with much lower errors than other methods such as that of Pedersen [1995]. However, the approximation is complicated by the need to also expand the coefficient functions in a convergent series and, although, the method has been adapted for multivariate diffusions [Ait-Sahalia, 2008] it still relies upon the Lamperti transform as do so many other estimation techniques.

4.3.6 Local Linearisation

Ozaki [1992] and Shoji and Ozaki [1998] use a **local linearisation** method that approximates the non-linear diffusion over some small time window. Then since the linear SDE is tractable they perform the estimation on this. It is an extension of the Euler method: whereas the Euler approximation is piecewise constant the local linearisation is piecewise linear. Shoji and Ozaki [1998] extend this to multivariate diffusions and Shoji [1998] analyse the rate of convergence of the linear approximation. Roberts and Stramer [2001] use a linear approximation, calculate the bridge distribution and use this in the context of MCMC inference for proposing missing data for a one dimensional model. They approximate the drift function, at time t using values at $s < t$ by

$$\boldsymbol{\mu}(\boldsymbol{X}_t,\boldsymbol{\theta}) = \boldsymbol{\mu}(\boldsymbol{X}_s,\boldsymbol{\theta}) + (\boldsymbol{X}_t - \boldsymbol{X}_s)\boldsymbol{\mu}'(\boldsymbol{X}_s,\boldsymbol{\theta})\,.$$

This is not the full approximation given by Ito's Formula. Applying Ito's formula to the drift function of multivariate SDE

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t)dt + \boldsymbol{a}(\boldsymbol{X}_t)d\boldsymbol{B}_t\,, \tag{4.44}$$

where $\boldsymbol{X} \in \mathbb{R}^d$, $\boldsymbol{\mu} : \mathbb{R}^d \to \mathbb{R}^d$ and $\boldsymbol{a} : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ we have

$$d\mu_i(\boldsymbol{X}_t) = \sum_j \frac{\partial \mu_i(\boldsymbol{X}_t)}{\partial x_j} dX_j + \frac{1}{2} \sum_{j,k,l} a_{jl}(\boldsymbol{X}_t) a_{kl}(\boldsymbol{X}_t) \frac{\partial^2 \mu_i}{\partial x_j \partial x_k}(\boldsymbol{X}_t) dt \,.$$

Therefore, we can approximate $\boldsymbol{\mu}(\boldsymbol{X}_t)$ by

$$\begin{aligned} \mu_i(\boldsymbol{X}_t) \approx \mu_i(\boldsymbol{X}_s) &+ \sum_j \frac{\partial \mu_i(\boldsymbol{X}_s)}{\partial x_j} (X_j(t) - X_j(s)) \\ &+ \frac{1}{2} \sum_{j,k,l} a_{jl}(\boldsymbol{X}_s) a_{kl}(\boldsymbol{X}_s) \frac{\partial^2 \mu_i}{\partial x_j \partial x_k}(\boldsymbol{X}_s)(t-s) \,, \end{aligned}$$

where $0 \leq s < t$. Substituting this into (4.44) gives the approximating process

$$d\boldsymbol{Z}_t = (\boldsymbol{Q}\boldsymbol{Z}_t + \boldsymbol{r}(t))dt + \boldsymbol{\Sigma} d\boldsymbol{B}_t \,,$$

with

$$\begin{aligned} Q_{ij} &= \frac{\partial \mu_i(\boldsymbol{X}_s)}{\partial x_j} \\ r_i(t) &= \mu_i(\boldsymbol{X}_s) - \sum_j \frac{\partial \mu_i}{\partial x_j} \boldsymbol{X}_j(s) + \frac{1}{2} \sum_{j,k,l} a_{jl}(\boldsymbol{X}_t) a_{kl}(\boldsymbol{X}_t) \frac{\partial^2 \mu_i}{\partial x_j \partial x_k}(\boldsymbol{X}_t)(t-s) \\ \boldsymbol{\Sigma} &= \boldsymbol{a}(\boldsymbol{X}_s) \,. \end{aligned}$$

In Chapter 5 we study the efficiency of using the linear bridge for multivariate diffusions as a proposal distribution for missing data in nonlinear diffusions. Note that the diffusion coefficient, as described above, is constant. With regard to the remarks at the end of Section 4.3.3 this will not work well as a proposal in the continuous time limit. Therefore, in Chapter 5 we also investigate the linear bridge where the diffusion depends upon the state: this is akin to the modified bridge as discussed in Section 4.3.3.

4.4 Exact Algorithms

Beskos and Roberts [2005] developed an algorithm for simulating diffusion paths for a certain class of models without discretisation error. This **Exact Algorithm 1** (EA1) was followed by EA2 [Beskos et al., 2006] and EA3 [Beskos et al., 2008], which enlarged the class of eligible diffusions. These algorithms have also been used to estimate the transition densities between observations in order to estimate

parameters through maximum likelihood or Bayesian methods [Beskos et al., 2006]. EA1 is applicable to one dimensional diffusions of the form

$$dX_t = \mu(X_t, \boldsymbol{\theta})dt + dB_t\,.$$

This is not overly restrictive as one dimensional diffusions can be transformed to this form via the Lamperti transform.

The algorithm is a rejection sampler: one proposes from a simple density $x \sim g(x)$ then accepts the proposal with probability $Mf(x)/g(x) < 1$, where $1/M$ is the upper bound of the ratio $f(x)/g(x)$. In the case of diffusions one proposes $Z \sim \mathbb{Z}$ from an absolutely continuous diffusion then calculates the Radon-Nikodym derivative with respect to the target $X \sim \mathbb{P}$. Therefore, we must be able to calculate Girsanov's formula and furthermore, it must be bounded.

This places some restrictions on μ. Firstly it must be differentiable so that Ito's formula can be applied to the stochastic integral in Girsanov's formula. In order to bound Girsanov's formula Beskos and Roberts [2005] suggest that $\mathbb{Z}$ should be a biased Brownian motion that has an endpoint distributed as $h(B_T) \propto \exp(A(B_T) - B_T^2/(2T))$, where $A(B_T) = \int_0^{B_T} \mu(u)du$. Further requirements are

- The integral $\int \exp(A(u) - \frac{u^2}{2T})du < \infty$ so that $h(u)$ is a probability density
- There exist constants $k_1, k_2 \in \mathbb{R}$ such that $k_1 \leq 0.5(\mu^2(u) + \mu'(u)) \leq k_2$ for all $u \in \mathbb{R}$
- The time interval T is small enough so that $0 \leq \phi(u) \leq T^{-1}$, where $\phi(u) = 0.5(\mu^2(u) + \mu'(u)) - k_1$

With these constraints Beskos and Roberts [2005] show that the Radon-Nikodym derivative is

$$\frac{d\mathbb{P}}{d\mathbb{Z}}(X_{[0,T]}) = \exp(-H(X_{[0,T]}))\,, \tag{4.45}$$

where $H(X_{[0,T]}) = \int_0^T \phi(X_t)dt$. This is indeed bounded by 1.

The next step is to consider how to calculate the probability in Eq. (4.45). It is clear that one can not calculate this exactly because that would require knowledge of the entire path $X_{[0,T]}$. However, it is possible to simulate a **skeleton** of X which is accepted or rejected according to Eq. (4.45) then to fill in any gaps. This is the concept of **retrospective sampling** and was applied in a different context by Beskos and Roberts.

Beskos et al. [2006] provided a simpler acceptance criteria than the original by Beskos and Roberts [2005]. The acceptance criteria is calculated as fol-

lows. First simulate points of a Poisson Process with unit intensity on the interval $[0,T]\times[0,M]$ where M is an upper bound for $\phi(u)$. This gives a collection of points $(t_1,y_1),\ldots,(t_k,y_k)$. Now given the endpoint simulated from $\mathbb{Z}$, simulate the proposal at times $t_1,\ldots,t_k$. The probability that $\phi(X_{t_k})<y_k$ for all k is equal to $\exp(-H(X_{[0,T]}))$. Therefore, if this criteria is met the proposal should be accepted, otherwise it is rejected.

Beskos et al. [2006] extend the method to EA2. This no longer requires $\phi(u)$ to be bounded above as in EA1. It is now only required that $\limsup_{u\to\infty}\phi(u)<\infty$ or $\limsup_{u\to-\infty}\phi(u)<\infty$. This is accomplished by decomposing the Brownian bridge to be conditioned upon its lowest or highest point. This bound is first simulated, then Bessel processes (see Øksendal [2007]) are used to connect this to the endpoints. This implies that the bound used to simulate the Poisson process is stochastic but proves no extra restriction in the implementation of the algorithm, although one problem is that the length of time (number of Poisson points) needed to determine if a proposed path is accepted varies strongly according to each proposal.

The one sided boundedness of $\phi(u)$ in EA2 is quite restrictive: it is not satisfied by the Ornstein-Uhlenbeck process, for example. Beskos et al. [2008] introduce EA3, which does not require $\phi(u)$ to be bounded. The algorithm works by simulating a partition of the sample space. Within each partition $\phi(u)$ can be bounded and the path simulated exactly.

The exact algorithms are advanced methods that are broadly applicable to the simulation of one dimensional diffusions and can be used in construction efficient estimators of the likelihood function [Beskos et al., 2006]. However, they are not as applicable to multivariate diffusions. Firstly, if there is a non-unit diffusion function then the Lamperti transform must first be used: this is not always possible. Secondly, the drift function must be the gradient of some potential so that it can be integrated. In models relevant for low frequency variability of the atmosphere discussed in Chapter 3 it was argued that, due to the homogenisation procedure, a linear noise term was important. Even if this could be transformed to unit diffusion it is not likely that this would produce a drift that was the gradient of a potential. Even before a transformation we would not expect this to be the case as travelling waves in the atmosphere would preclude the existence of a potential. Therefore, we focus on more generally applicable methods that are not exact for finite observation interval but their error can be reduced systematically by imputing missing data between observations.

4.5 Alternatives to Likelihood Estimation

4.5.1 Estimating Functions

In order to estimate a $p-$dimensional parameter $\boldsymbol{\theta}$ then an **estimating function** $F_N \in \mathbb{R}^p$ is a function of the observed data and the parameter such that $F_N(\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta}) = 0$. If the estimating function is the score function (derivative of the likelihood with respect to $\boldsymbol{\theta}$) then the solution is the MLE, however, as already discussed, the likelihood is not available for inference in SDEs.

Estimating functions should be unbiased and the parameters should be identifiable

$$\mathbb{E}_{\boldsymbol{\theta}_0}\left[F_N(\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta})\right] = 0, \text{ if and only if } \boldsymbol{\theta} = \boldsymbol{\theta}_0\,, \tag{4.46}$$

where the $\boldsymbol{\theta}_0$ are the true parameters. If the estimating function is the score function then the solution of

$$S_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\theta}} p(\boldsymbol{X}_{i+1}|\boldsymbol{X}_i; \boldsymbol{\theta}) \tag{4.47}$$

gives the maximum likelihood estimator. Since this is not available one considers estimating functions that have the same structure

$$F_N(\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta}) = \sum_{i=1}^{N} f(\boldsymbol{X}_{i+1}, \boldsymbol{X}_i, \boldsymbol{\theta})\,. \tag{4.48}$$

The score function is a **martingale** so one choice is to look for estimating functions that also obey the martingale property

$$\mathbb{E}_{\boldsymbol{\theta}}(F_N(\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta})|\mathcal{F}_{N-1}) = F_{N-1}(\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta}) \tag{4.49}$$

where $\mathcal{F}_{N-1}$ is the σ-algebra generated by the process up to time t_{N-1}. An advantage to using **martingale estimating functions** is that all of the known asymptotic results for $N \to \infty$ associated with martingales are applicable [Sorensen, 2004].

The functions in Eq. (4.48) have the structure

$$f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = \sum_{j=1}^{J} \alpha_j(\boldsymbol{x}, \boldsymbol{\theta}) h_j(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta})\,, \tag{4.50}$$

where J is some integer and the α_j are weight functions. The martingale property then implies

$$\mathbb{E}_{\boldsymbol{\theta}}\left[h_j(\boldsymbol{x}, \boldsymbol{X}_1, \boldsymbol{\theta})|\boldsymbol{X}_0 = \boldsymbol{x}\right] = 0 \tag{4.51}$$

for all j.

Sorensen [1999] proved that, given suitable convergence of the martingale, then a solution $\tilde{\boldsymbol{\theta}}$ of $F_N(\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta}) = 0$ exists, $\tilde{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ in probability and $\tilde{\boldsymbol{\theta}}$ will have an asymptotically Gaussian distribution. Sorensen [1997] derived optimal weights α^* to minimise the asymptotic variance but discusses the need to use approximations due to numerical problems.

A usual choice for the functions $h_1 \dots h_N$ is

$$h_j(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\theta}) = g_j(\boldsymbol{y}, \boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}}\left[g_j(\boldsymbol{X}_1, \boldsymbol{\theta}) | \boldsymbol{X}_0 = \boldsymbol{x}\right] , \tag{4.52}$$

for some simple functions g_j with finite expectation. Often g_j are polynomials: $g_j(\boldsymbol{y}, \boldsymbol{\theta}) = y^{k_j}$ [Bibby and Sorensen, 1995, 1996]. The expectation in Eq. (4.52) can be estimated easily by simulation. An alternative is to choose eigenfunctions of the generator. Then, under some regularity conditions, the expectation in Eq. (4.52) can be calculated as

$$\mathbb{E}_{\boldsymbol{\theta}}\left[g_j(\boldsymbol{X}_1, \boldsymbol{\theta}) | \boldsymbol{X}_0 = \boldsymbol{x}\right] = \exp(-\lambda_j(\boldsymbol{\theta})\Delta) g_j(\boldsymbol{x}, \boldsymbol{\theta}) , \tag{4.53}$$

where $\lambda_j(\boldsymbol{\theta})$ are the eigenvalues [Sorensen, 2004].

Alternatives to martingale estimating functions are **simple estimating functions** with the form

$$F_N(\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta}) = \sum_{i=1}^{N} f(\boldsymbol{X}_i, \boldsymbol{\theta}) \tag{4.54}$$

so that f takes, as argument, a state variable at only one time t_i. Then Eq. (4.46) implies $\mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{X}_0, \boldsymbol{\theta}) = 0$ if and only if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. This involves only the marginal distributions and completely ignores any dependence structure. It is suitable for estimating parameters that appear in the invariant distribution. Possibilities for f include the score of the invariant distribution [Sorensen, 2001] or low order polynomials [Kessler, 2000]. Martingale estimating functions are not as easily constructed in non-Markovian models or when some components of the diffusion are unobserved such as Stochastic Volatility models. In such cases a useful technique could be the **prediction based estimating function** of Sorensen [2000] (see also Sorensen [2011]).

4.5.2 Generalised Method of Moments

Related to the technique of estimating functions is the **Generalised Method of Moments** [Hansen, 1982]. To estimate parameter $\boldsymbol{\theta} \in \mathbb{R}^p$ one chooses $p' > p$

moment conditions

$$w(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[f(\boldsymbol{\theta}, \boldsymbol{X})]\,, \tag{4.55}$$

where the f_j are functions of one or more of the state variables. In practice this is estimated as

$$\hat{w}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N} f(\boldsymbol{\theta}, \boldsymbol{X}_i) = 0 \tag{4.56}$$

and the estimate is obtained by minimising the quadratic form

$$\hat{w}(\boldsymbol{\theta})^T \Omega \hat{w}(\boldsymbol{\theta}) \tag{4.57}$$

for some weight matrix Ω. Hansen [1982] shows that the resulting estimator is consistent, asymptotically normal and that it is efficient with the optimal weight matrix

$$\Omega = \left(\mathbb{E}\left[f(\boldsymbol{\theta}_0, \boldsymbol{X}) f(\boldsymbol{\theta}_0, \boldsymbol{X})^T\right]\right)^{-1}\,. \tag{4.58}$$

Since this matrix depends upon the unknown true parameter, in practice, it is calculated iteratively by plugging in preliminary estimates.

An example of moment conditions for SDEs are

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{A}_{\boldsymbol{\theta}}(g(\boldsymbol{X}_t))\right] = 0 \\ &\mathbb{E}_{\boldsymbol{\theta}}\left[\mathcal{A}_{\boldsymbol{\theta}}(g(\boldsymbol{X}_t))h(\boldsymbol{X}_s) - \mathcal{A}_{\boldsymbol{\theta}}(h(\boldsymbol{X}_s))f(\boldsymbol{X}_t)\right] = 0\,, \end{aligned} \tag{4.59}$$

where $t > s \geq 0$, h and f are any suitable functions and $\mathcal{A}_{\boldsymbol{\theta}}$ is the infinitesimal generator of the process [Hansen and Scheinkman, 1995].

4.5.3 Estimation Via an Auxiliary Model

Consider data $\boldsymbol{X}$ with density $p_N(\boldsymbol{X}|\boldsymbol{\theta})$. If we can simulate data from p and evaluate a suitable **auxiliary density** $q_N(\boldsymbol{X}|\boldsymbol{\rho})$ then we can find a link $\boldsymbol{\rho} = f(\boldsymbol{\theta})$ that we can use to calculate the MLE of $\boldsymbol{\theta}$ [Gourieroux et al., 1993]. The idea is that the auxiliary model reflects important aspects of the full model and emphasises these in the estimation.

Firstly one finds the MLE from the auxiliary model

$$\hat{\boldsymbol{\rho}} = \underset{\boldsymbol{\rho}}{\operatorname{argmax}} \log q_N(\boldsymbol{X}|\boldsymbol{\rho}) \tag{4.60}$$

based on the actual observed data. Then one simulates data from the true model for a range of different parameters $\boldsymbol{\theta}$ and calculates $\hat{\boldsymbol{\rho}}$ for each. The estimator $\hat{\boldsymbol{\theta}}$

minimises the difference between the estimate of $\boldsymbol{\rho}$ from the real and simulated data.

$$(\hat{\boldsymbol{\rho}} - f(\boldsymbol{\theta}))^T \Omega (\hat{\boldsymbol{\rho}} - f(\boldsymbol{\theta}))\,, \tag{4.61}$$

where Ω is the weight matrix. If $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{\theta}}$ are of the same dimension then f can be taken as the identity. An alternative scheme called **efficient method of moments** [Gallant and Tauchen, 1996] minimises

$$\left(\frac{\partial}{\partial \boldsymbol{\rho}} \log q_N(\boldsymbol{X}^{\boldsymbol{\theta}} | \hat{\boldsymbol{\rho}})\right)^T \Omega \left(\frac{\partial}{\partial \boldsymbol{\rho}} \log q_N(\boldsymbol{X}^{\boldsymbol{\theta}} | \hat{\boldsymbol{\rho}})\right)\,, \tag{4.62}$$

where $\boldsymbol{X}^{\boldsymbol{\theta}}$ is the data simulated using parameter $\boldsymbol{\theta}$.

The quality of the estimators depend upon the chosen auxiliary model. One suggestion is just to use the Euler approximation model but another idea would be to use a linearised version. The method would be useful when performing inference for complicated models with missing data. For example, models with unobserved components or unusual structure such as some components that are not directly driven by noise. However, choosing the auxiliary model is close to guessing and the technique can be computationally intensive.

4.6 Conclusion

In this chapter we have given a broad overview of methods to estimate parameters from SDE models. We have summarised the literature explaining why this is a challenging inference problem and described some approaches to its solution. For our application we choose to use the Algorithms 4.1 and 4.2 as they are the most flexible, being applicable to general multivariate diffusions, and can be made arbitrarily accurate by imputing more missing data. This does introduce a computational challenge: sampling missing data can become very inefficient in higher dimensions. We explore the limitations of the algorithm and attempt to alleviate some of the problems in the next chapter specifically focussing on the model class discussed in Chapter 3.

Chapter 5

Inference for Models with Cubic Drift and Linear Diffusion

In this Chapter we study the following D-dimensional model

$$dX_i = \left(\alpha_i + \sum_{j=1}^{D} \beta_{i,j} X_j + \sum_{j=1}^{D} \sum_{k=1}^{j} \gamma_{i,j,k} X_j X_k + \sum_{j=1}^{D} \sum_{k=1}^{j} \sum_{l=1}^{k} \lambda_{i,j,k,l} X_j X_k X_l \right) dt + \sum_{j=1}^{i} a_{i,j} dB_j + \sum_{j=1}^{D} \sum_{k=1}^{i} b_{j,i,k} X_j dB_k \,, \tag{5.1}$$

where the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}$, $\boldsymbol{a}$ and $\boldsymbol{b}$ are unknown. This is the most general form of cubic model with linear noise, which we motivated in Chapter 3. We denote all of the parameters as $\boldsymbol{\theta}$. We place a prior $p(\boldsymbol{\theta})$ on the parameters then, given observations $\boldsymbol{X}_{\text{obs}}$, we aim to estimate the Bayesian posterior distribution

$$p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}}) \propto p(\boldsymbol{X}_{\text{obs}}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \,, \tag{5.2}$$

where $p(\boldsymbol{X}_{\text{obs}}|\boldsymbol{\theta})$ is the likelihood function that must be approximated according to the methods in Chapter 4. The inference is obtained using Markov Chain Monte Carlo Methods, specifically the **Innovation Scheme** of Section 4.3.4.

In this Chapter we study the computational aspects of the inferring parameters in a model of the form Eq. (5.1). We first consider the growth in parameter numbers with the dimension of the system. As the dimension D of the system increases the number of parameters in the drift increases rapidly. For each component, $\boldsymbol{\alpha}$ consists of 1 parameter, there are D in $\boldsymbol{\beta}$, $D(D+1)/2$ in $\boldsymbol{\gamma}$ and $D(D+1)(D+2)/6$ in $\boldsymbol{\lambda}$. The last two are given by the triangle and tetrahedral numbers respectively.

In total, for each component, there are

$$P = 1 + \frac{11D}{6} + D^2 + \frac{D^3}{6} \tag{5.3}$$

parameters and DP parameters in altogether. Therefore, the number of parameters in the drift grows as the fourth power of D. The sequence increases as $\{4, 20, 60, 140, \ldots\}$, so even for two or three dimensional models there are a lot of parameters.

The theoretical results of Florens-Zmirou [1989], discussed in Chapter 4, show that these parameters can be estimated as $Nh^2 \to \infty$ using the Euler approximation of the transition density, where h is the maximum interval and N is the number of observations. However, we aim to achieve an understanding of how much data is needed in practice and in particular to determine whether it is feasible with a realistic amount of data that would be available for a study in the atmospheric sciences. We will do this in Section 5.4.

Consider now the diffusion parameters. There is a coefficient matrix $\boldsymbol{a}$ for constant terms and one for each linear term $\boldsymbol{b}_j, j = 1 \ldots D$. We have constructed the matrices so that they only have lower diagonal components. This means that the parameters will be identifiable. In total there will be $D(D+1)^2/2$ diffusion parameters so this again grows rapidly in the dimension of the system. From Chapter 4 we know that if we observe the system in continuous time then the diffusion parameters are known exactly. However, realistically we only observe the process at discrete intervals of time and this observation frequency may not even be sufficient to approximate continuous observation satisfactorily. In Section 5.3 we investigate the data requirements, in order to accurately infer the diffusion parameters, by varying the observation interval and length. This is also a check on parameter identifiability.

If it happens that our data set is not observed at a high frequency all is not lost as we can impute data between observations to improve our approximation of the likelihood function. As discussed in Section 4.3.4 one then implements a Gibbs sampler to alternate between sampling parameters from $p(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{miss}})$ and missing data from $p(\boldsymbol{X}_{\text{miss}}|\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta})$. In Section 5.2 we discuss Metropolis-Hastings Independence Samplers for missing data with different proposal processes. All algorithms are based on Algorithm 4.2 in Section 4.3.4, but differ in the proposal distribution used to sample missing segments of data. The variants of this algorithm are summarised in Table 5.1. The efficiency of each proposal is studied for models of the form Eq. (5.1) and it is demonstrated numerically that the most efficient algorithm is based on the linear proposal process. This proposal is based

upon a linearisation of the drift function, as discussed in Section 4.3.6 in a different context. This is novel applied in a MCMC algorithm and is one of the major contributions of this thesis. The results of an extensive numerical study of the efficiency of different proposal distributions are shown to be stable as the dimension of the system increases.

The update of parameters is split between those in the drift and those in the diffusion. The diffusion parameters are updated according using Algorithm 4.1 of Section 4.3.4. Numerical verification of the algorithm applied to the model in Eq. (5.1) is presented in Section 5.3.

The drift parameters enter quadratically into the likelihood function. This implies that their conditional posterior will be multivariate normal. In Section 5.4 we determine the mean and covariance for this normal distribution and demonstrate that this indeed regains the true parameter values with sufficiently many observations.

In this chapter we are concerned with computational issues as well as constraints of data. The main algorithms are all implemented in C/C++. However, in Section 5.5 we investigate the potential benefits of a new computing paradigm that benefits from the massive parallelisation in **Graphics Processing Units** (GPU). We design our standard algorithm to exploit a parallel architecture. We implement it in the CUDA language, run it on a basic laptop GPU and find massive reductions in computation time.

Also important for reducing the required computation time is to design efficient MCMC algorithms. Ideally the Markov Chain should move around the space rapidly to get as close to i.i.d sampling as possible. This is a major theme running through this Chapter and so firstly, in the next Section, we discuss the issues of convergence and efficiency of MCMC applied to Bayesian inference.

5.1 Aspects of Bayesian Inference via Markov Chain Monte Carlo

The problem is to obtain samples distributed according to probability density $\pi(\boldsymbol{\theta})$ in order to estimate functionals such as the mean. In Bayesian statistics this would be a posterior distribution conditional upon observations and we want to estimate the unknown parameters. It is often not possible to obtain independent, identically distributed (i.i.d) samples from π. One possibility is to construct a **Markov Chain** that has invariant density π. That is a process with transitions $\boldsymbol{\theta}_1 \to \boldsymbol{\theta}_2 \to \ldots \to \boldsymbol{\theta}_{n-1} \to \boldsymbol{\theta}_n$ that for n large enough has $\boldsymbol{\theta}_n \sim \pi$. This technique is known as Markov

Chain Monte Carlo (MCMC). For an introduction and guide to MCMC in practice see Robert and Casella [2005] and Gilks and Spiegelhalter [1996]). For a summary of theoretical results see Roberts and Rosenthal [2004].

A simple way to construct such a Markov chain is to ensure that it is reversible with respect to π. If $p(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is the transition density for moving from state $\boldsymbol{\theta}$ to state $\boldsymbol{\theta}^*$ then $\pi(\boldsymbol{\theta})p(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \pi(\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*, \boldsymbol{\theta})$ implies that the Markov Chain is reversible and therefore stationary with respect to π [Roberts and Rosenthal, 2004]. The Metropolis-Hastings algorithm exploits this property.

One should be able to initialise the Markov Chain at any point in the space and have it converge to the unique density π. Intuitively this means that all states should be realisable from any starting point. In particular the chain should not be reducible since, even if π is a stationary distribution of the Markov Chain, it may not be unique and the chain may not converge to it. Furthermore, we require the chain to be aperiodic. If these requirements are fulfilled then we can guarantee that the chain will converge to π asymptotically, however, practically we want to know how large n has to be [Roberts and Rosenthal, 2004].

Properties of the Markov chain such as **uniform** or, the weaker, **geometric** ergodicity can be proved to obtain qualitative results about its convergence rate [Roberts and Rosenthal, 2004]. In some cases one can make quantitative estimates on the number of iterations to ensure convergence (for example Rosenthal [1995]). In general cases often convergence is determined empirically by observing the behaviour of multiple chains applied to the same target. The chains are started in over-dispersed initial states and then between chain and within chain information is compared to diagnose convergence [Gelman and Rubin, 1992].

After convergence, MCMC methods can be used to obtain samples from the distribution and these can then be used to estimate expectations of functions $h : \Theta \to \mathbb{R}$. In Bayesian inference one is often interested in the expectation of unknown parameters under the posterior. Most simply an estimate of unknown parameter θ is given by

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \theta^{(i)} \,, \tag{5.4}$$

where the $\theta^{(i)}$ are the MCMC output. If the samples were i.i.d then this estimator would be unbiased and a **Central Limit Theorem** (CLT) would guarantee convergence to normality with variance σ^2/n, where $\sigma^2 < \infty$ is the posterior variance of θ. For uniformly ergodic Markov Chains the same $\sqrt{n}$ CLT applies. For geometrically ergodic chains a $\sqrt{n}$ CLT exists if the $2 + \delta$ moment of the target is finite, where $\delta > 0$ [Roberts and Rosenthal, 2004].

Even if a CLT exists, since the MCMC samples are correlated, the effective sample size is smaller and the errors larger. We aim to design rapidly mixing chains that have low autocorrelation and smaller errors. We denote the autocorrelation at lag j as $C(j) = \sum_i (\theta^{(i)} - \hat{\theta})(\theta^{(i+j)} - \hat{\theta})/\hat{\text{Var}}(\theta)$, where $\hat{\text{Var}}(\theta)$ is an estimate of the variance. For correlated samples the variance of Eq. (5.4) is

$$\text{Var}(\hat{\theta}) = \sigma^2/n \left(1 + 2 \sum_{j=1}^{n-1} (1 - \frac{j}{n}) C(j) \right) .$$

(see Robert and Casella [2005]). The difference to the i.i.d case is the term in brackets and is called the **integrated autocorrelation time** τ for the algorithm. The less correlation within the MCMC samples the more efficient the algorithm. As a measure of **efficiency** we consider τ for $n \to \infty$ so that

$$\tau = 1 + 2 \sum_{j=1}^{\infty} C(j) . \tag{5.5}$$

In this thesis we are often interested in estimating the efficiency of MCMC algorithms.

Definition 1 (Estimated efficiency of MCMC algorithms)**.** *We report the efficiency of an algorithm as* $\eta = 100/\hat{\tau}$*, where* $\hat{\tau}$ *is an estimate of the integrated autocorrelation time of Eq. (5.5):* $\hat{\tau} = \sum_{j=1}^{j^*} C(j)$*, where* j^* *is the first* j *such that* $C(j) < 0.05$ *and* $j > 6$*.*

Various optimisation strategies arise for all the algorithms in this thesis. It is often possible to design the Markov Chain transition probabilities to account for some properties of the target. We do this in Section 5.2 by implementing a linearisation of the target process when sampling missing diffusion paths.

In Metropolis-Hastings algorithms it is sensible to monitor the acceptance rate of the proposals and adjust the jump size. We do this in Section 5.3 with the aim of keeping the acceptance rate in the range $0.2 - 0.4$. This is in accordance with the rate of 0.234 motivated by theoretical arguments in Roberts et al. [1997].

Even if the MCMC algorithm has converged and mixes well one should check that the true values of parameters are recoverable. When way of doing this is to quantify the error of the estimate as the amount of data used for the inference increases. The error of the estimates for each data set can be quantified using the

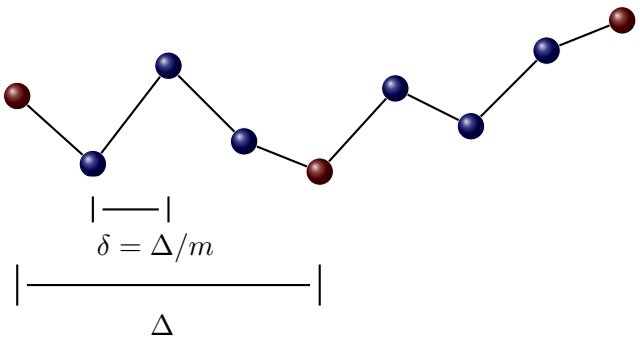


Figure 5.1: Illustration of the inference problem. Red circles represent observations and blue are missing values to impute. The inclusion of missing data reduces the time interval to $\delta = \Delta/m$. Here $m = 4$.

quadratic Posterior Expected Loss (PEL) function

$$f(\hat{p}, \boldsymbol{X}_{\text{obs}}) = \int_{\Theta} (\boldsymbol{\theta}^* - \boldsymbol{\theta})^2 \hat{p}(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}}) d\boldsymbol{\theta}\,, \tag{5.6}$$

where $\boldsymbol{\theta}$ represents all of the parameters, $\hat{p}$ is the estimated posterior distribution and $\boldsymbol{\theta}^*$ is the true value of the parameter. Note that the PEL does not distinguish between parameters with different sizes and may unequally weight those with large variances. It it important to check that the parameter estimates have variances of similar magnitude.

5.2 Inference for Missing data

In this section we study the efficiency of methods for simulating diffusion paths from Eq. (5.1) that are conditioned upon start $\boldsymbol{X}_0 = \boldsymbol{x}_0$ and end $\boldsymbol{X}_T = \boldsymbol{x}_T$ points. In the introduction this was written $p(\boldsymbol{X}_{\text{miss}}|\boldsymbol{X}_{\text{obs}}, \boldsymbol{\theta})$. For simplicity consider the probability density of a single observation interval, where $t_T - t_0 = \Delta$ is divided into m equidistant subintervals so that $t_{k+1} - t_k = \Delta/m = \delta$ and there are $m - 1$ missing data vectors to sample. As discussed in Chapter 4 this target density, with respect to Lebesgue measure, is the product

$$p(\boldsymbol{X}_1 \cdots \boldsymbol{X}_{m-1}|\boldsymbol{X}_0 = \boldsymbol{x}_0, \boldsymbol{X}_m = \boldsymbol{x}_m; \boldsymbol{\theta}) = \prod_{k=0}^{m-1} p(t_{k+1}, \boldsymbol{x}_{k+1}|t_k, \boldsymbol{x}_k; \boldsymbol{\theta})\,, \tag{5.7}$$

where now p is the transition density for the process in Eq. (5.1) and is, in practice, approximated by the Euler transition density. An illustration of the problem is given in Figure 5.1.

Algorithm Name	Proposal Distribution
Brownian Bridge (BB)	Eq. (5.11)
Modified Bridge (MB)	Eq. (5.13)
Linear Bridge (LB)	Eq. (5.21)
Modified Linear Bridge (MLB)	Eq. (5.22)
Brownian Bridge Lamperti (BL)	Brownian Bridge of Eq. (5.11) applied to transformed data of Eq. (5.11)
Linear Bridge Lamperti (LL)	Linear Bridge of Eq. (5.21) applied to transformed data of Eq. (5.11)

Table 5.1: List of proposal distributions for Algorithm 4.2 that are studied and tested in this chapter.

We focus on the case of complete (noiseless) observation of the process. This implies that blocks of missing data are independent and can be considered separately. For more general cases see, for example, Golightly and Wilkinson [2008].

It is not possible to simulate directly from the law of the conditioned process in Eq. (5.1) so we use an **independence sampler**. This is a Metropolis-Hastings algorithm with proposal density of the form $q(\boldsymbol{X}^*|\boldsymbol{X}) = q(\boldsymbol{X}^*)$ so that it does not depend upon the current state. It is still a Markov Chain since the current state enters into the acceptance probability Eq. (4.41).

Here, we consider proposal processes of the form

$$d\boldsymbol{X}^* = \boldsymbol{\xi}(\boldsymbol{X}_t^*, \boldsymbol{X}_T)dt + \boldsymbol{a}(\boldsymbol{X}^*, \boldsymbol{\sigma})d\boldsymbol{W}^*\,, \tag{5.8}$$

where $\boldsymbol{a}(\boldsymbol{X}^*, \boldsymbol{\sigma})$ is the same diffusion function as that in Eq. (5.1). This is motivated from the arguments in Chapter 4 about absolute continuity of diffusion measures. Associated with Eq. (5.8) is the proposal transition density

$$q(\boldsymbol{X}^*) = \prod_{k=0}^{m-2} q(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k, \boldsymbol{\xi}, \boldsymbol{\sigma}) \tag{5.9}$$

and inference is implemented according to Algorithm 4.2. In the following we discuss possible drift and diffusion functions for the proposal SDE in Eq. (5.8). For convenience Table 5.1 provides a reference of all variants of Algorithm 4.2 we consider in this chapter.

The simplest choice is the **Brownian Bridge** (see e.g Gardiner [2004]) process

$$d\boldsymbol{X}^* = \left(\frac{\boldsymbol{x}_T - \boldsymbol{X}_t^*}{T-t}\right)dt + \boldsymbol{a}(\boldsymbol{X}_0, \boldsymbol{\sigma})d\boldsymbol{B}^*\,. \tag{5.10}$$

This is designed so that $\boldsymbol{X}_T = \boldsymbol{x}_T$. Note that the constant diffusion function means that the Brownian Bridge can be simulated exactly so that, given a proposed value $\boldsymbol{x}_k^*$ sampled at time t_k, the proposal distribution for $\boldsymbol{X}_{k+1}^*$ is

$$q(\boldsymbol{X}_{k+1}^*|\boldsymbol{X}_k^*) = \mathcal{N}\left(\frac{(T-t_{k+1})\boldsymbol{x}_k^* + (t_{k+1}-t_k)\boldsymbol{x}_T}{T-t_k}, \frac{(t_{k+1}-t_k)(T-t_{k+1})}{T-t_k}\boldsymbol{\Sigma}(\boldsymbol{x}_0,\boldsymbol{\sigma})\right), \tag{5.11}$$

$\boldsymbol{\Sigma}(\boldsymbol{x}_0,\boldsymbol{\sigma}) = \boldsymbol{a}(\boldsymbol{x}_0,\boldsymbol{\sigma})\boldsymbol{a}(\boldsymbol{x}_0,\boldsymbol{\sigma})^T$. Given the background in Chapter 4 about absolute continuity of measures we expect the Brownian Bridge to be poor due to the constant diffusion function.

An alternative for Eq. (5.8) is the **Modified Bridge** (MB) proposal of Durham and Gallant [2002]. In this case

$$\boldsymbol{\xi}(\boldsymbol{X}_t^*, \boldsymbol{X}_T) = \frac{\boldsymbol{X}_T - \boldsymbol{X}_t^*}{T-t}\,. \tag{5.12}$$

In discrete form this implies a proposal distribution

$$q(\boldsymbol{X}_{k+1}^*|\boldsymbol{X}_k^*, \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{X}_{k+1}^*; \boldsymbol{X}_k^* + \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k)\,, \tag{5.13}$$

where

$$\tilde{\boldsymbol{\mu}}_k = \frac{\boldsymbol{X}_m - \boldsymbol{X}_k^*}{m-k}, \quad \tilde{\boldsymbol{\Sigma}}_k = \boldsymbol{a}(\boldsymbol{X}_k^*, \boldsymbol{\sigma})\boldsymbol{a}(\boldsymbol{X}_k^*, \boldsymbol{\sigma})^T\delta\,. \tag{5.14}$$

5.2.1 Linear Bridge as a Proposal Process

The aim of this section is to design an efficient independence sampler for Eq. (5.1) by focussing on the drift in Eq. (5.8). The proposal is constructed by first linearising the SDE, then forming the bridge process. The linearisation was demonstrated in Section 4.3.6 for a multivariate diffusion. Here we demonstrate the approximation using a two dimensional example

$$\begin{aligned} dX_t &= (2X_t + 3X_tY_t - X_t^3)dt + Y_t dB_1(t) + dB_2(t) \\ dY_t &= (Y_t - X_tY_t - Y_t^3)dt + dB_1(t) + X_t dB_2(t) \end{aligned} \tag{5.15}$$

These equations are linearised as explained in Section 4.3.6 to give the approximating equation

$$d\boldsymbol{Z}_t = (\boldsymbol{Q}\boldsymbol{Z}_t + \boldsymbol{r}(t))dt + \boldsymbol{\Sigma}d\boldsymbol{B}_t\,, \tag{5.16}$$

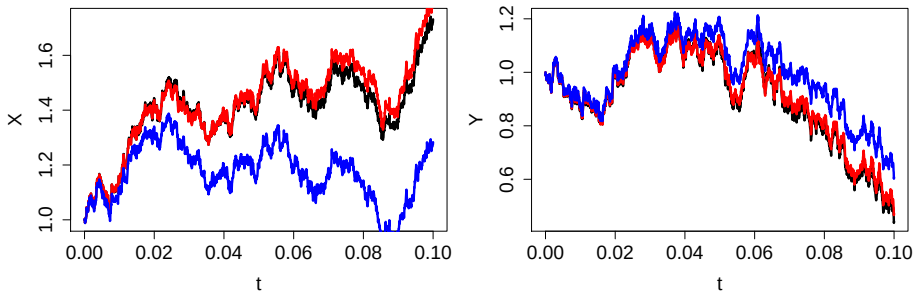


Figure 5.2: Sample paths of both components of the non-linear SDE (Eq. 5.15) (black), the linear approximation (red) and the Brownian motion (blue) using the same random variables.

with the quantities

$$\boldsymbol{Q} = \begin{pmatrix} 2 + 3Y_s - 3X_s^2 & 3X_s \\ -Y_s & 1 - X_s - 3Y_s^2 \end{pmatrix}$$
$$\boldsymbol{r}(t) = \begin{pmatrix} 2X_s^3 - 3X_sY_s + 3(X_s + Y_s - X_s(Y_s^2 + 1))(t - s) \\ X_sY_s + 2Y_s^3 - (X_s + Y_s + 3Y_s(X_s^2 + 1))(t - s) \end{pmatrix}$$
$$\boldsymbol{\Sigma} = \begin{pmatrix} Y_s^2 + 1 & X_s + Y_s \\ X_s + Y_s & X_s^2 + 1 \end{pmatrix}$$

Figure 5.2 compares the strong, pathwise solutions to the SDE Eq. (5.15) with the linear approximation and Brownian motion with constant diffusion. Note that the linear approximation remains close to the non-linear solution for significantly longer than the Brownian motion. Figure 5.3 is concerned with weak solutions. The linear approximation does much better than the Brownian motion at reproducing the strong drift.

One of the contributions of this thesis is the demonstration of the time inhomogeneous Linear Bridge distribution for efficient sampling of multivariate diffusions. We are able to do this using results for constructing bridge distributions of general multivariate linear diffusions of the form in Eq. (5.16) [Barczy and Kern, 2010]. If at time s we have $\boldsymbol{X}_s = \boldsymbol{a}$ and at time T, $\boldsymbol{X}_T = \boldsymbol{b}$ then the distribution of $\boldsymbol{X}_t$ for $0 \leq s < t \leq T$ can be shown to be Gaussian with mean

$$\boldsymbol{\nu}_{\boldsymbol{a},\boldsymbol{b}}(s,t) = \boldsymbol{\Gamma}(t,T)\boldsymbol{\Gamma}(s,T)^{-1}\boldsymbol{m}_{\boldsymbol{a}}^{+}(s,t) + \boldsymbol{\Gamma}(s,T)^{T}(\boldsymbol{\Gamma}(s,T)^{T})^{-1}\boldsymbol{m}_{\boldsymbol{b}}^{-}(t,T) \qquad (5.17)$$

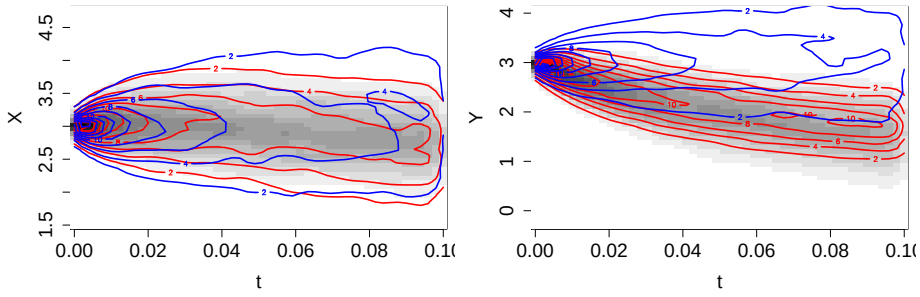


Figure 5.3: Comparison of the distributions of the original process in Eq. 5.15 (black/grey) compared with contour plots of the linear approximation in Eq. (5.16) (red) and Brownian motion (blue) evolving from a fixed initial condition for both components of Eq. (5.15).

where

$$\mathbf{\Gamma}(s,t) = \int_s^t e^{(s-u)\boldsymbol{Q}} \mathbf{\Sigma}\mathbf{\Sigma}^T e^{(t-u)\boldsymbol{Q}^T} du\,,$$

$$\boldsymbol{m}_{\boldsymbol{x}}^+(s,t) = \boldsymbol{x} + \int_s^t e^{(s-u)\boldsymbol{Q}} \boldsymbol{r}(u) du \quad \text{and} \quad \boldsymbol{m}_{\boldsymbol{x}}^-(s,t) = \boldsymbol{x} - \int_s^t e^{(t-u)\boldsymbol{Q}} \boldsymbol{r}(u) du\,.$$

The covariance matrix is given by

$$\mathbf{\Sigma}(s,t) = \mathbf{\Gamma}(t,T)\mathbf{\Gamma}(s,T)^{-1}\mathbf{\Gamma}(s,t)\,. \tag{5.18}$$

For the one dimensional case we can easily calculate the quantities involved: let $v(s) = \mu(X_s) - \mu'(X_s)\mu_s - 1/2\mu''(X_s)\Sigma^2 s$, then

$$m_a^+(s,t) = a + \frac{1}{Q}(1-e^{(s-t)Q})v(s) + \frac{1}{2Q}\left((s - e^{(s-t)Q}t) + \frac{1}{Q}(1-e^{(s-t)Q})\right)\mu''(X_s)\Sigma^2$$

$$m_b^-(s,t) = b + \frac{1}{Q}(1-e^{(t-s)Q})v(s) + \frac{1}{2Q}\left((s - e^{(t-s)Q}t) + \frac{1}{Q}(1-e^{(t-s)Q})\right)\mu''(X_s)\Sigma^2$$

$$\Gamma(s,t) = \frac{\Sigma^2}{2Q}\sinh((t-s)Q)$$

$$\Sigma(s,t) = \frac{\Sigma^2}{2Q}\frac{\sinh((T-t)Q)\sinh((t-s)Q)}{\sinh((T-s)Q)}$$

In a multivariate problem we can only compute these terms in a nice form if the matrices $\mathbf{\Sigma}$ and $\boldsymbol{Q}$ commute. In this case

$$\mathbf{\Gamma}(s,t) = \mathbf{\Sigma}(\boldsymbol{Q}+\boldsymbol{Q}^T)^{-1}\left(e^{(t-s)\boldsymbol{Q}^T} - e^{(s-t)\boldsymbol{Q}}\right)\mathbf{\Sigma}^T\,.$$

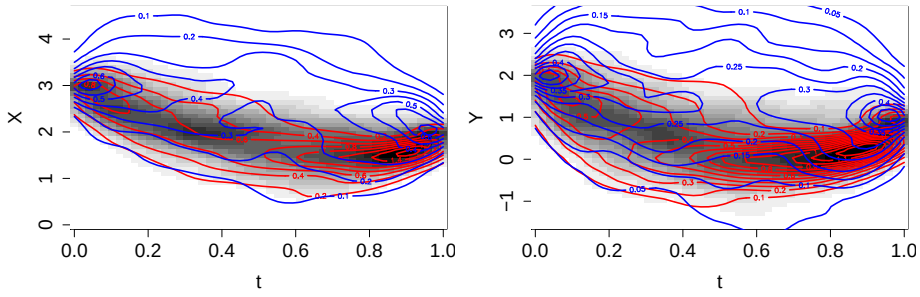


Figure 5.4: Comparison of the distributions of the non-linear bridge process derived from Eq. 5.15 (black/grey) compared with contour plots of the modified linear bridge in Eq. (5.21) (red) and Brownian bridge (blue). Here we use $\boldsymbol{a} = (3, 2)$, $\boldsymbol{b} = (2, 1)$ and $\epsilon = 0.1$.

In general this matrix can be computed as follows: if we diagonalise $\boldsymbol{Q}$ so that $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}$ then compute the matrix $\boldsymbol{A}$ with components

$$A_{ij} = \frac{(\boldsymbol{U}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^{-T})_{ij}}{\Lambda_{ii} + \Lambda_{jj}} \left(e^{(t-s)\Lambda_{jj}} - e^{(s-t)\Lambda_{ii}} \right) \tag{5.19}$$

then

$$\boldsymbol{\Gamma}(s, t) = \boldsymbol{U}\boldsymbol{A}\boldsymbol{U}^T . \tag{5.20}$$

We compute the bridge process for Eq. (5.15) using Eqns. (5.19,5.20) between fixed $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^2$. To test the performance we simulated the target SDE (Eq. 5.15) with $(X_0, Y_0) = \boldsymbol{a}$ and accepted paths if $X_T \in [b_1 - \epsilon, b_1 + \epsilon]$ and $Y_T \in [b_2 - \epsilon, b_2 + \epsilon]$ for some small ϵ, typically $\epsilon \leq 0.1$. This computationally intensive procedure forms an approximation to the nonlinear bridge. For a test example we have found that there is not much improvement over using a Brownian bridge for short bridge intervals ($T < 0.1$, not shown) so we compared the performance for $T = 1$. Figure 5.4 shows that the linear bridge performs much better than the Brownian Bridge. The variance is a lot lower and the density tracks the high probability regions of the target more accurately.

The distribution for sampling $\boldsymbol{X}_{k+1}$ given $\boldsymbol{X}_k$ and $\boldsymbol{X}_m$ using the **Linear Bridge** (LB) proposal is

$$q(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k, \boldsymbol{X}_m, \boldsymbol{\theta}) = \mathcal{N}\left(\boldsymbol{X}_{k+1}; \boldsymbol{\nu}_{\boldsymbol{x}_k, \boldsymbol{x}_m}(k\delta, (k+1)\delta), \boldsymbol{\Sigma}_{\boldsymbol{x}_0}(k\delta, (k+1)\delta)\right) , \tag{5.21}$$

where $\boldsymbol{\nu}_{\boldsymbol{x}_k, \boldsymbol{x}_m}(k\delta, (k+1)\delta)$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}_0}(k\delta, (k+1)\delta)$ are given in Esq. (5.17) and (5.18)

respectively. In this case the matrices $\boldsymbol{Q}$ and $\boldsymbol{\Gamma}$ are both computed using the value $\boldsymbol{x}_0$ and we emphasise this with the subscript on the covariance matrix in Eq. (5.21). As it is, we do not expect the linear bridge to be effective at proposing missing data for the target SDE. This is for the reasons discussed in Chapter 4. The constant diffusion matrix means that the law of linear bridge paths will not be absolutely continuous with respect to the law of the target process.

Instead we introduce the **Modified Linear Bridge** (MLB) proposal. This is the same as the linear bridge except that the diffusion matrix in the linear approximation is updated at each imputed point. This means recomputing the matrices $\boldsymbol{\Gamma}(s,t)$ at each point although $\boldsymbol{Q}$ and $\boldsymbol{U}$ are only calculated once. The proposal distribution is only a small modification of Eq. (5.21):

$$q(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k, \boldsymbol{X}_m, \boldsymbol{\theta}) = \phi\left(\boldsymbol{X}_{k+1}; \boldsymbol{\nu}_{\boldsymbol{x}_k,\boldsymbol{x}_m}(k\delta,(k+1)\delta), \boldsymbol{\Sigma}_{\boldsymbol{x}_k}(k\delta,(k+1)\delta)\right) , \quad (5.22)$$

with the $\boldsymbol{x}_k$ subscript to emphasise that $\boldsymbol{\Sigma}(\boldsymbol{x}_k)$ is used instead of $\boldsymbol{\Sigma}(\boldsymbol{x}_0)$.

We compare the Modified Bridge, Linear Bridge and Modified Linear Bridge on the following one dimensional model

$$dX_t = \alpha(1 + X_t + X_t^2 - X_t^3) + X_t dB_t, \;\; X_0 = 1\,. \quad (5.23)$$

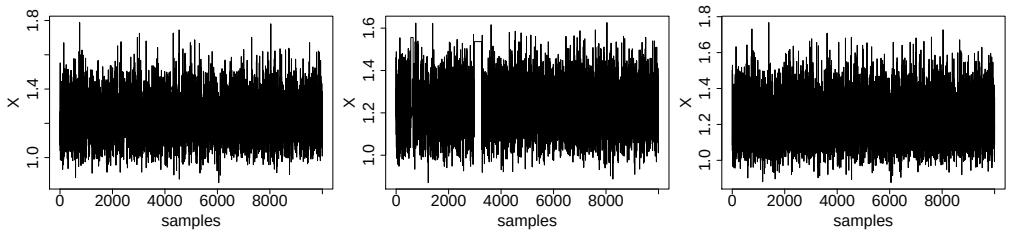
we observe the process $N = 101$ times with $\Delta = 0.1$. An example of the MCMC output of the average value for an arbitrary interval is shown in Figure 5.5 for varying amounts of missing data m. It is clear that the Linear Bridge mixes poorly compared with the other two algorithms as the amount of missing data increases. Figure 5.6 shows the autocorrelation functions of the MCMC output averaged over all data. There is significant autocorrelations in the Linear Bridge output, whereas the Modified Linear Bridge has a rapidly decaying autocorrelation function even for large m.

We consider the multivariate generalisation of Eq. (5.23)

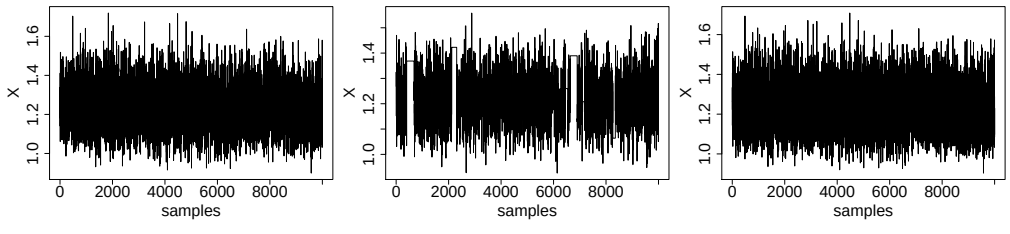
$$dX_i = \alpha(X_i(t) + X_i(t)\sum_j^d X_j(t) - X_i^3(t))dt + X_i(t)dB_i(t), \quad \boldsymbol{X} \in \mathbb{R}^d\,, \quad (5.24)$$

We check that the Modified Linear Bridge reproduces the same distribution of missing data as the much simpler Modified Bridge. We do this for $d = 2$, $N = 101$ and $\Delta = 0.1$. The results are shown in Figure 5.7 for varying m. It is clear that there is visual agreement between the densities proposed by the two different distributions.

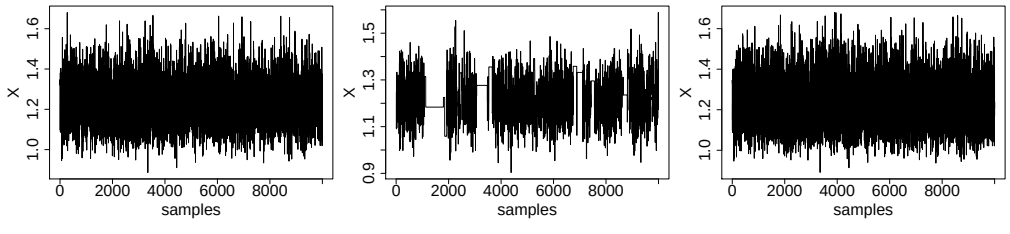
We now compare the Monte Carlo efficiency of different bridge processes



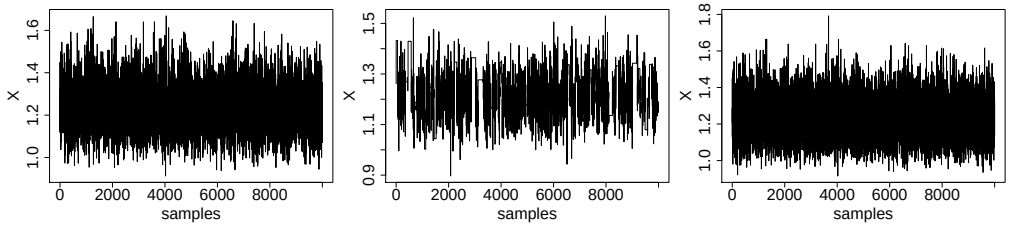
(a) $m = 10$, Modified Bridge (b) $m = 10$, Linear Bridge (c) $m = 10$, Modified Linear Bridge



(d) $m = 50$, Modified Bridge (e) $m = 50$, Linear Bridge (f) $m = 50$, Modified Linear Bridge

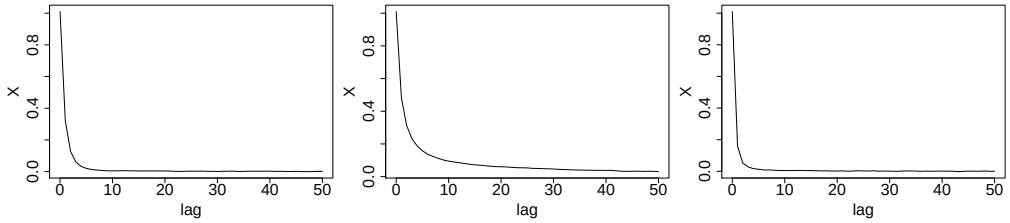


(g) $m = 70$, Modified Bridge (h) $m = 70$, Linear Bridge (i) $m = 70$, Modified Linear Bridge

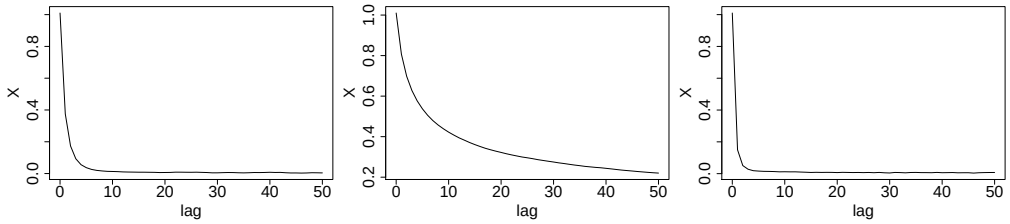


(j) $m = 100$, Modified Bridge (k) $m = 100$, Linear Bridge (l) $m = 100$, Modified Linear Bridge

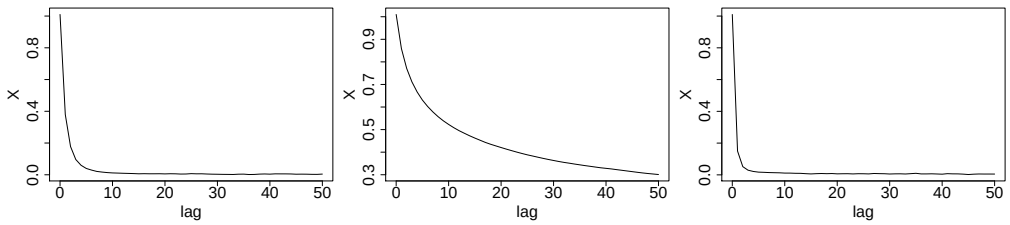
Figure 5.5: Trace plots of the MCMC output for sampling missing data from the model in Eq. (5.23). The data shown is the average value for an arbitrary observation interval with $\Delta = 0.1$. The Modified Bridge is on the left, the Linear Bridge is centre and the Modified Linear Bridge on the right.



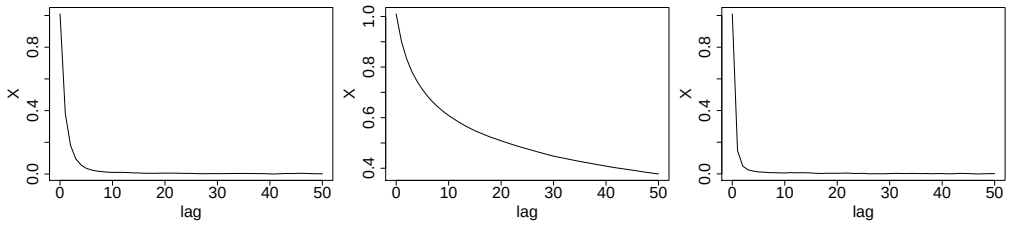
(a) $m = 10$, Modified Bridge (b) $m = 10$, Linear Bridge (c) $m = 10$, Modified Linear Bridge



(d) $m = 50$, Modified Bridge (e) $m = 50$, Linear Bridge (f) $m = 50$, Modified Linear Bridge

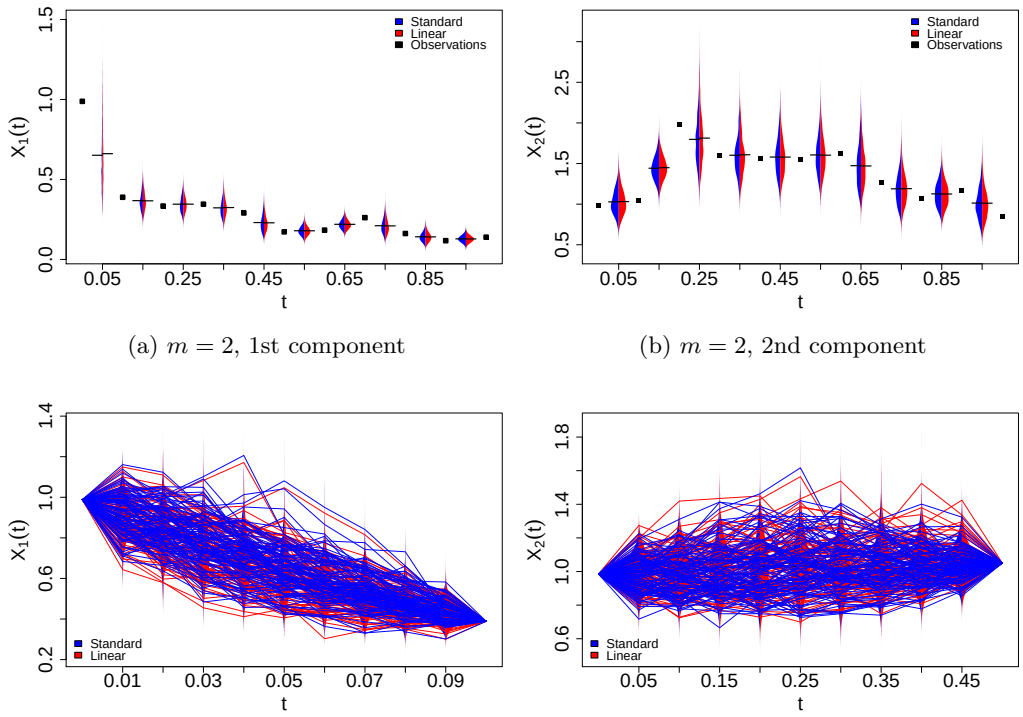


(g) $m = 70$, Modified Bridge (h) $m = 70$, Linear Bridge (i) $m = 70$, Modified Linear Bridge



(j) $m = 100$, Modified Bridge (k) $m = 100$, Linear Bridge (l) $m = 100$, Modified Linear Bridge

Figure 5.6: Average autocorrelation functions computed for MCMC output of $N = 100$ data intervals from the model in Eq. (5.23) with interobservation time $\Delta = 0.1$. The Modified Bridge is on the left, the Linear Bridge is centre and the Modified Linear Bridge on the right.



(a) $m = 2$, 1st component

(b) $m = 2$, 2nd component

(c) $m = 10$, 1st component

(d) $m = 10$, 2nd component

Figure 5.7: Output from Standard and Linear Bridge samplers applied to Eq. 5.24 in two dimensions observed at $\Delta = 0.1$. For each MCMC algorithm 10^5 samples were retained after discarding a burn in of 10^4. Plots (a) and (b) show a series of 11 observations over $T = 1.0$ with imputed data $m = 2$. At each imputed data point the density of both samplers is plotted using Kernel Density Estimation and the "beanplot" package in R. Plots (c) and (d) show the estimated densities for the imputed data with $m = 10$ for a single observation interval. Also shown are some sample paths from both MCMC algorithms.

acting as proposals for missing data applied to the model Eq. (5.24). We use the SDE in Eq. (5.24) in place of our general model Eq. (5.1) because, while retaining the non-linearity and state dependent noise, it can be transformed to one of unit diffusion via the Lamperti transform using Eq. (2.22). This transformation $Y_i = \log(X_i)$ leads to

$$dY_i = \left(\alpha(1 + \sum_{j=1}^{d} Y_j - e^{2Y_j}) - \frac{1}{2} \right) dt + dB_i \,. \tag{5.25}$$

The constant diffusion means that the Bridge proposal will be absolutely continuous

with respect to the target bridge and there will be no discretisation error due to modifying the diffusion function. We use the standard Brownian Bridge proposal and linear bridge proposal applied to Eq. (5.25) as a demonstration of efficiency in the ideal case.

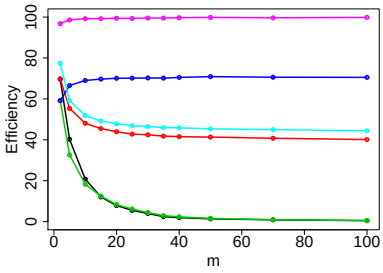
Altogether we compare six proposal methods: the **Brownian Bridge** (BB) proposal will serve as a demonstration of how not to do it, the **Modified Bridge** (MB) proposal as a benchmark, the **Linear Bridge** (LB) proposal, the **Modified Linear Bridge** (MLB), the **Brownian Bridge Lamperti transformed** (BL) and the **Linear Bridge Lamperti transformed** (LL). These algorithms are listed in Table 5.1

We estimate the efficiency of the proposal for varying dimension and amount of missing data m. We also vary α which controls the relative contributions from the drift and the diffusion terms. In each case the efficiency is averaged over all missing data points. We use a total of $N = 101$ observations with fixed interval $\Delta = 0.1$. The efficiency as given in Definition 1 is calculated from the integrated autocorrelation function estimated from 10^5 samples.

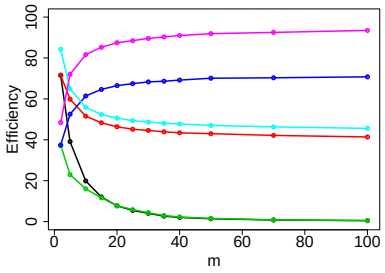
The results, shown in Figure 5.8, are for the case of updating all components simultaneously. Notice that the estimates smoothly converge for increasing m which implies that sufficient samples have been used for each estimation. As expected the most efficient proposal is the Lamperti transformed linear bridge (shown in magenta). Of course the Lamperti transformation can not be calculated generally for multi dimensional models. The Modified Linear Bridge can be applied generally and it performs better than any other general method. As in the Lamperti transformed linear bridge the efficiency actually improves with m as the linear model becomes a better approximation to the target and then reaches a plateau. It does not deteriorate at large m as the proposal has measure absolutely continuous with respect to the target. The original Modified Bridge and the standard Lamperti transformation method deteriorate for increasing m, which must be due to the increasing dimension of the problem as they quickly reach a plateau and then do not deteriorate any further. It is surprising that they do not perform very well for this model as one would hope the efficiency of an independence sampler higher. The standard Brownian Bridge and Linear Bridge perform very poorly: the efficiency rapidly goes to zero for increasing m. As mentioned previously this is due to the mutual exclusivity of their measures.

For $\alpha = 1.0$ all methods perform poorly as the dimension increases. There is a strong drift in this case, which it is harder to approximate in higher dimensions. A potential solution to this problem is to update one component at a time, keeping

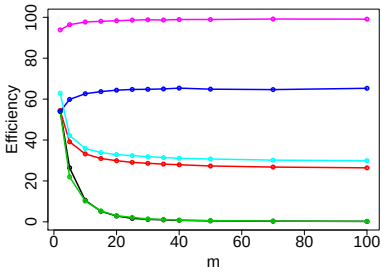
the others fixed. The results of this procedure are shown in Figure 5.9. It is clear that this approach does not deteriorate as quickly with the dimension. However, this approach is much more computationally expensive to the extent of being unpractical.



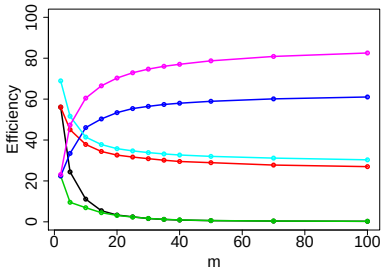
(a) $d = 1$, $\alpha = 0.1$



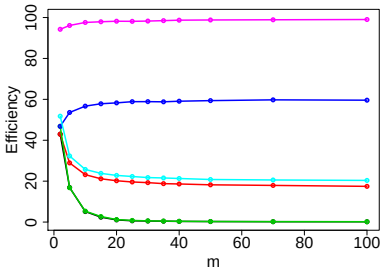
(b) $d = 1$, $\alpha = 1.0$



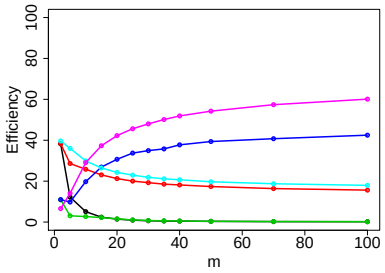
(c) $d = 2$, $\alpha = 0.1$



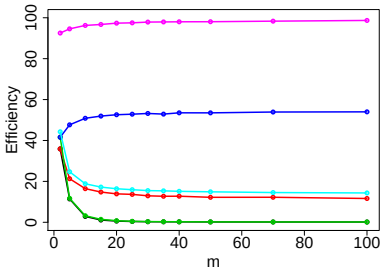
(d) $d = 2$, $\alpha = 1.0$



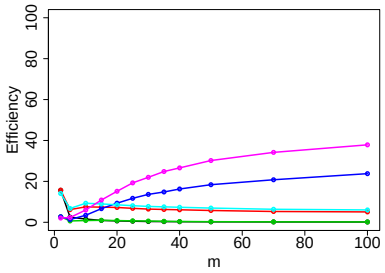
(e) $d = 3$, $\alpha = 0.1$



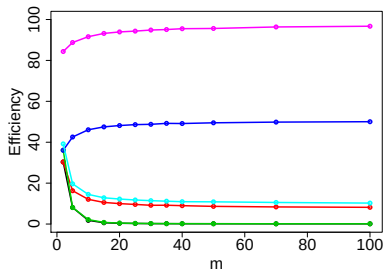
(f) $d = 3$, $\alpha = 1.0$



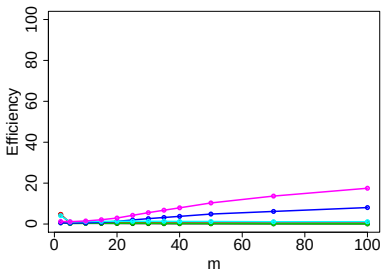
(g) $d = 4$, $\alpha = 0.1$



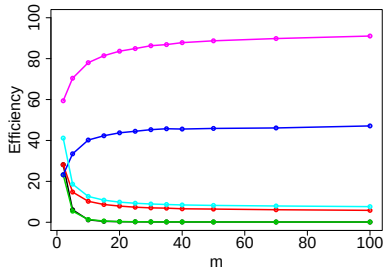
(h) $d = 4$, $\alpha = 1.0$



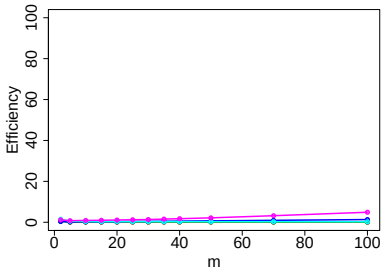
(i) $d = 5$, $\alpha = 0.1$



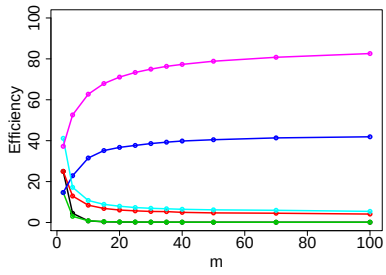
(j) $d = 5$, $\alpha = 1.0$



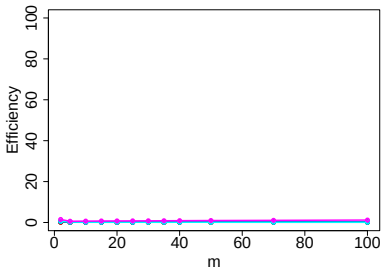
(k) $d = 6$, $\alpha = 0.1$



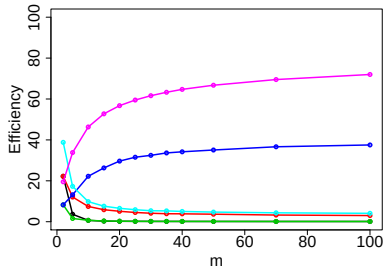
(l) $d = 6$, $\alpha = 1.0$



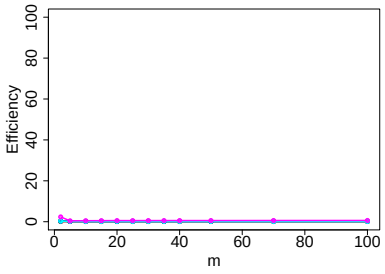
(m) $d = 7$, $\alpha = 0.1$



(n) $d = 7$, $\alpha = 1.0$

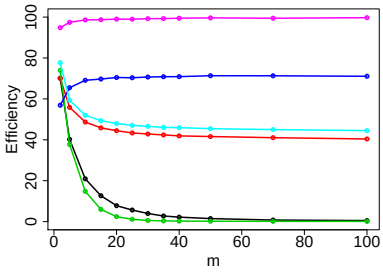


(o) $d = 8$, $\alpha = 0.1$

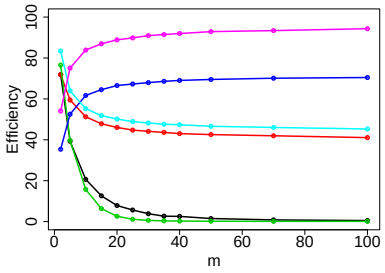


(p) $d = 8$, $\alpha = 1.0$

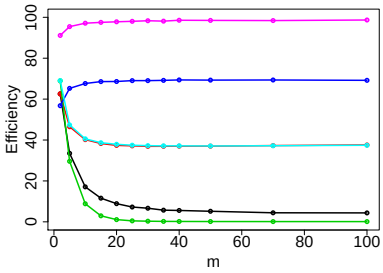
Figure 5.8: Efficiency of different data imputation proposals described in the text: BB - black, MB - red, LB - green, MLB - blue, BL - cyan, LL - magenta applied to the model in Eq. (5.24). In this case all components $\boldsymbol{X}$ were updated simultaneously. The data consisted of $N = 101$ samples at observation interval $\Delta = 0.1$. Only missing data was sampled in these algorithms. Each estimate of efficiency was calculated using 10^5 samples from three MCMC chains after a burn in of 10^4 samples.



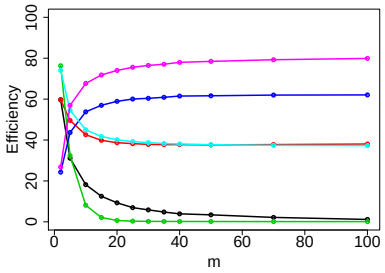
(a) $d = 1$, $\alpha = 0.1$



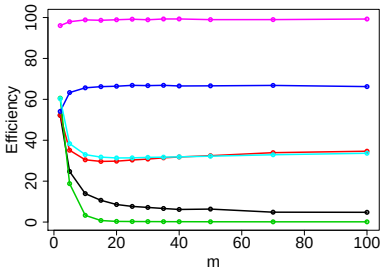
(b) $d = 1$, $\alpha = 1.0$



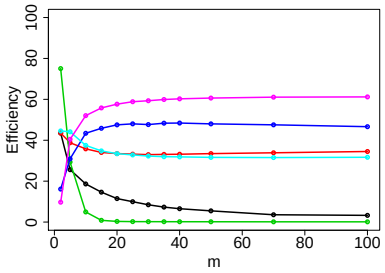
(c) $d = 2$, $\alpha = 0.1$



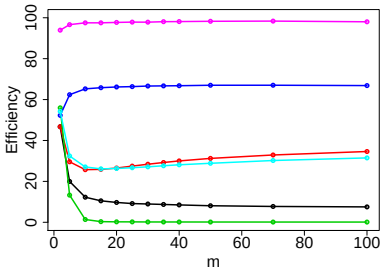
(d) $d = 2$, $\alpha = 1.0$



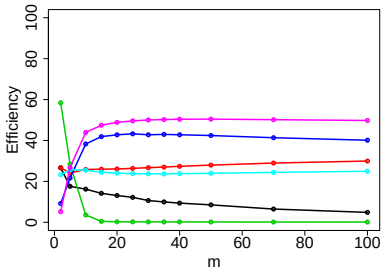
(e) $d = 3$, $\alpha = 0.1$



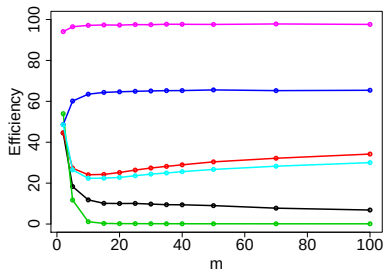
(f) $d = 3$, $\alpha = 1.0$



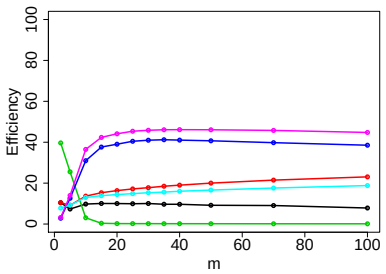
(g) $d = 4$, $\alpha = 0.1$



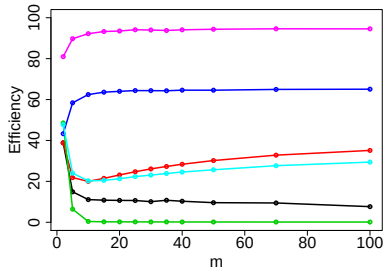
(h) $d = 4$, $\alpha = 1.0$



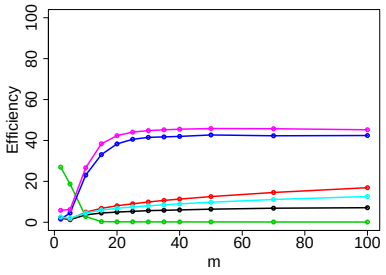
(i) $d = 5$, $\alpha = 0.1$



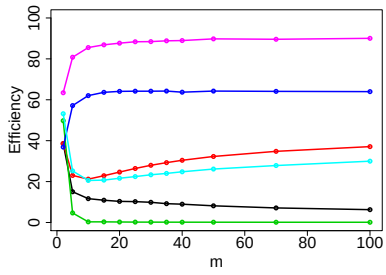
(j) $d = 5$, $\alpha = 1.0$



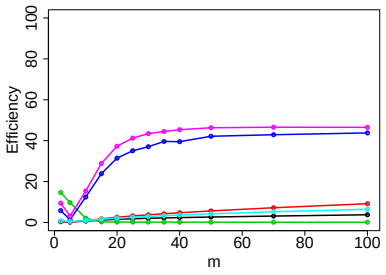
(k) $d = 6$, $\alpha = 0.1$



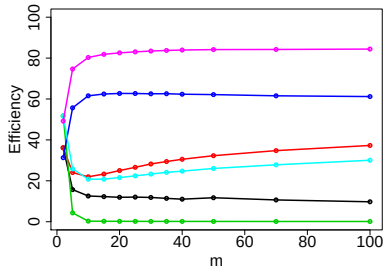
(l) $d = 6$, $\alpha = 1.0$



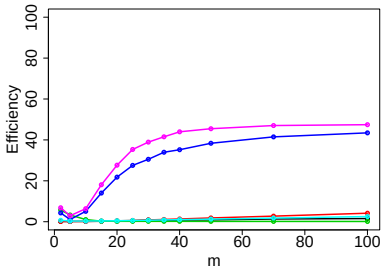
(m) $d = 7$, $\alpha = 0.1$



(n) $d = 7$, $\alpha = 1.0$



(o) $d = 8$, $\alpha = 0.1$



(p) $d = 8$, $\alpha = 1.0$

Figure 5.9: Efficiency of different proposals described in the text for the component-wise updating: BB - black, MB - red, LB - green, MLB - blue, BL - cyan, LL - magenta applied to the model in Eq. (5.24). In this case each component of $\boldsymbol{X}$ was updated separately. The data consisted of $N = 101$ samples at observation interval $\Delta = 0.1$. Only missing data was sampled in these algorithms. Each estimate of efficiency was calculated using 10^5 samples from three MCMC chains after a burn in of 10^4 samples.

5.3 Inference for Diffusion Parameters

In this section we consider the sampling of parameters $\boldsymbol{\sigma}$ that enter into the diffusion function of Eq. (5.1). Recalling the general form of SDE

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, \boldsymbol{\theta})dt + \boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\sigma})d\boldsymbol{B}_t\,,\ \boldsymbol{X} \in \mathbb{R}^d\,, \tag{5.26}$$

As a test case we use the following

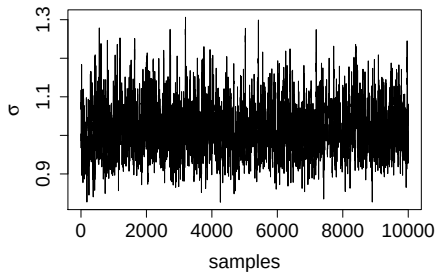
$$dX_i = \alpha(X_i(t) + X_i(t)\sum_{j=1}^{d} X_j(t) - X_i^3(t))dt + \sigma_i X_i(t)dB_i(t),\quad X \in \mathbb{R}^d\,, \tag{5.27}$$

A simple way to sample the diffusion parameters is to use the symmetric **Random Walk** proposal in the Metropolis-Hastings algorithm. Given the current value $\boldsymbol{\sigma}$ a new value is proposed from the Gaussian distribution $\boldsymbol{\sigma}^* \sim \mathcal{N}(\boldsymbol{\sigma}, \tau)$ where τ is a tuning parameter. In this case the proposal density drops out and the acceptance probability is just the ratio of posteriors

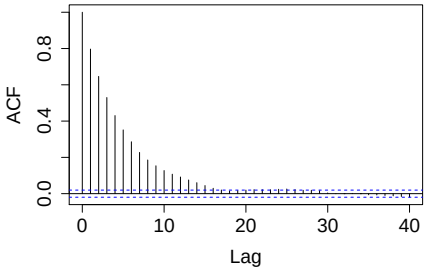
$$\alpha = \frac{p(\boldsymbol{\sigma}^*)}{p(\boldsymbol{\sigma})} \frac{\prod_{i=0}^{N-2}\prod_{j=0}^{m-1} p_\delta(\boldsymbol{X}_{im+j+1}|\boldsymbol{X}_{im+j}, \boldsymbol{\sigma}^*)}{\prod_{i=0}^{N-2}\prod_{j=0}^{m-1} p_\delta(\boldsymbol{X}_{im+j+1}|\boldsymbol{X}_{im+j}, \boldsymbol{\sigma})}\,,$$

where $p(\theta)$ is the prior distribution. This parameter update alternates with the sampling of missing data using Algorithm 4.2. The various choices for proposal distribution in Algorithm 4.2 are listed in Table 5.1. In this section we use the Modified Bridge proposal. Trace plots and autocorrelation functions of this algorithm applied to Eq. (5.27) with $d = 1$, $N = 101$ observations, interobservation time $\Delta = 0.1$, fixed $\alpha = 1.0$ and true value $\sigma = 1$ are shown in Figure 5.10. The trace plots show that the mixing of the algorithm becomes very poor as m increases; the autocorrelation becomes very high for large lags. This is due to the reasons discussed in Chapter 4: naive methods like the Random Walk actually become degenerate in the continuous time limit.

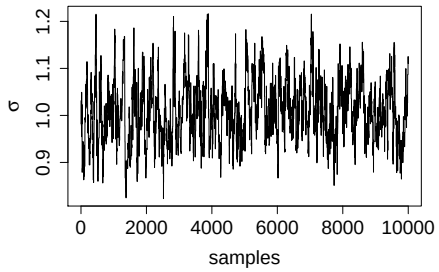
To overcome this we use the **Innovation Scheme** [Chib et al., 2004; Golightly and Wilkinson, 2008; Dargatz, 2010] (see Section 4.3.4). This applies the change of variables $\boldsymbol{Z} = \boldsymbol{g}^{-1}(\boldsymbol{X}, \boldsymbol{\sigma}) \in \mathbb{R}^d$, to give a process with unit diffusion. The general background and motivation for this algorithm was given in Chapter 4 and the detailed implementation in Algorithm 4.1. We applied the Innovation Scheme to the same problem as Figure 5.10. The results, in Figure 5.11, show that the mixing is much better than when using the Random Walk. The autocorrelation does not



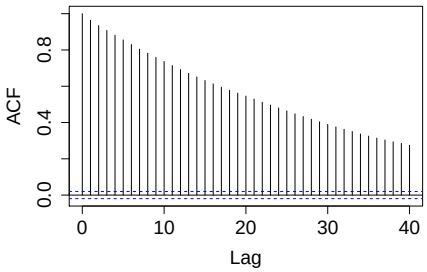
(a) Trace plot $m = 2$



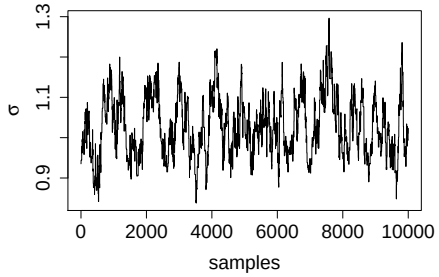
(b) Autocorrelation $m = 2$



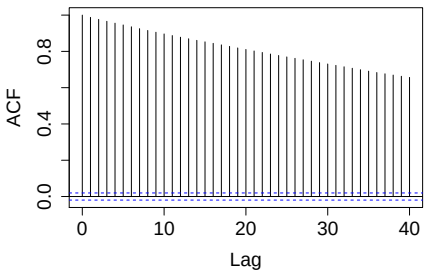
(c) Trace plot $m = 10$



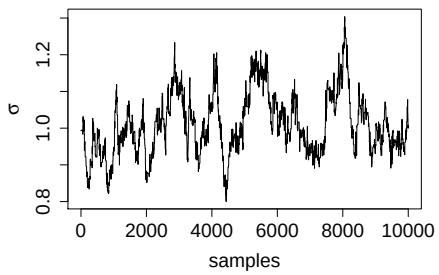
(d) Autocorrelation $m = 10$



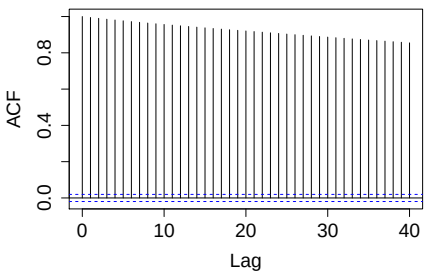
(e) Trace plot $m = 20$



(f) Autocorrelation $m = 20$



(g) Trace plot $m = 30$



(h) Autocorrelation $m = 30$

Figure 5.10: Output of Random Walk algorithm for σ applied to the one dimensional model in Eq. (5.27) with $N = 101$ observations, interobservation time $\Delta = 0.1$ and fixed $\alpha = 1.0$. The true value was $\sigma = 1.0$. On the right are the corresponding autocorrelation functions. Note that the Modified Bridge Sampler was used to impute missing data (see Table 5.1).

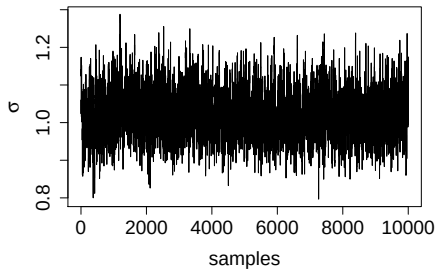
increase as more missing data is added. We choose to use the Innovation Scheme in our applied work in Chapter 8. However, the method still has faults. The main difficulty is due to having to update all of the missing data every time the diffusion parameters are sampled. This can result in a low acceptance rate and the step size must be scaled down. This is particularly apparent when applied to high dimensional problems with a reasonably large interobservation time. Figure 5.12 shows the output of the Random Walk and Innovation Scheme algorithms applied to a six dimensional model with large interobservation time $\Delta = 1.0$ and $N = 101$ observations. The value of $\alpha = 1.0$ was held fixed and the missing data was imputed using the Modified Bridge algorithm. All six of the parameters entering into the diffusion function were updated but only the trace plots of σ_1 are shown. In this case the mixing time of the Innovation Scheme becomes comparable to that of the Random Walk. The problem is that the autocorrelation is large even for low values of m. This could be due to discretisation error of the map $\boldsymbol{g}$. The algorithm is proved rigorously to work in continuous time [Dargatz, 2010]. With large discretisation error the function $\boldsymbol{g}^{-1}$ would not accurately map to a continuous time process that is of unit diffusion. There is certainly scope for an improved method of sampling the diffusion function but since the Innovation Scheme is very general in applicability and not too complicated to implement, we determine to use it in our applications.

We apply the Innovation Scheme as detailed in Algorithm 4.1 to the SDE in Eq. (5.1) in two dimensions. We have already described in the introduction the structure of the diffusion function and that the number of parameters increases as $d(d+1)^2/2$. For clarity we give the diffusion function explicitly for a two dimensional problem

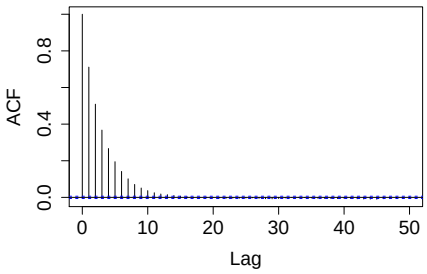
$$\boldsymbol{a}(\boldsymbol{X}, \boldsymbol{\sigma}) = \begin{pmatrix} \sigma_1 & 0 \\ \sigma_2 & \sigma_3 \end{pmatrix} + \begin{pmatrix} \sigma_4 & 0 \\ \sigma_5 & \sigma_6 \end{pmatrix} X_{t,1} + \begin{pmatrix} \sigma_7 & 0 \\ \sigma_8 & \sigma_9 \end{pmatrix} X_{t,2}\,. \tag{5.28}$$

The parameters will not be identifiable unless we restrict the domain using the prior. At the moment there is degeneracy under the mapping $(\sigma_1, \sigma_2, \sigma_3) \to -(\sigma_1, \sigma_2, \sigma_3)$. and other degeneracies are possible. The parameters are made identifiable by requiring $\sigma_1, \sigma_2, \sigma_3 > 0$. All other parameters can take negative values.

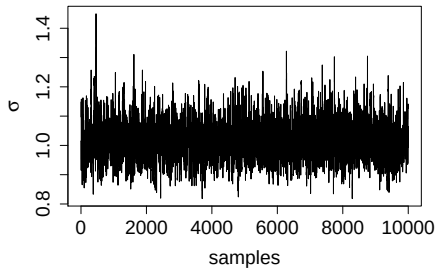
We performed a simulation study to test the data requirements for inferring the diffusion parameters. We used a two dimensional model, as in Eq. (5.28), and varied the interobservation time $\Delta = \{0.1, 0.01, 0.001\}$ and total time $T = \{1, 10, 100\}$ to understand how much data, and at what frequency, is required for accurate inference. We used randomly generated true values for the parameters in the diffusion and fixed the drift function parameters with values in Table 5.3. There



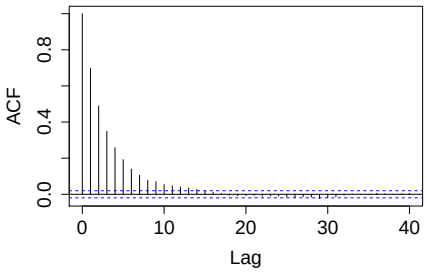
(a) Trace plot $m = 2$



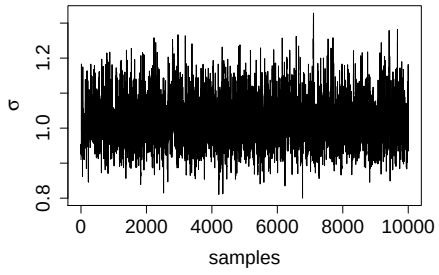
(b) Autocorrelation $m = 2$



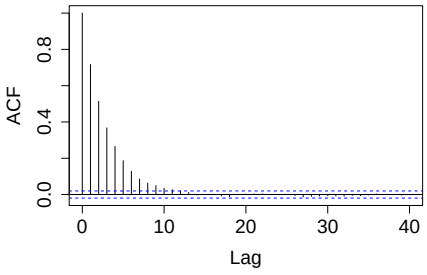
(c) Trace plot $m = 10$



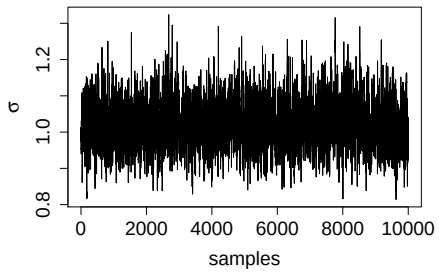
(d) Autocorrelation $m = 10$



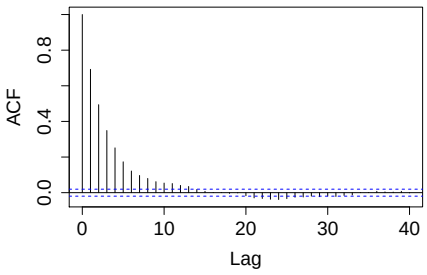
(e) Trace plot $m = 20$



(f) Autocorrelation $m = 20$

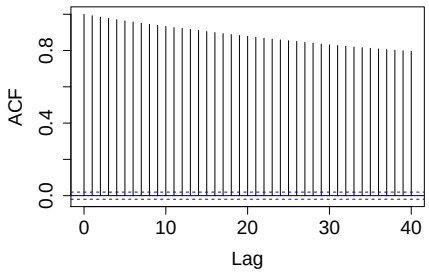


(g) Trace plot $m = 30$

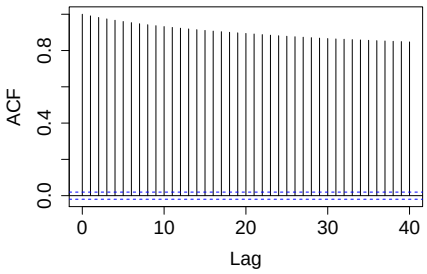


(h) Autocorrelation $m = 30$

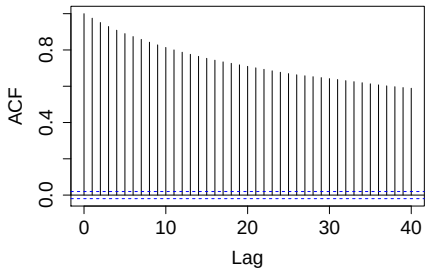
Figure 5.11: Output of Innovation Scheme for σ using the change of variables in Eq. (4.38) and Eq. (4.39) applied to the same data set used in Figure 5.10. The Modified Bridge sampler was used to impute missing data (see Table 5.1).



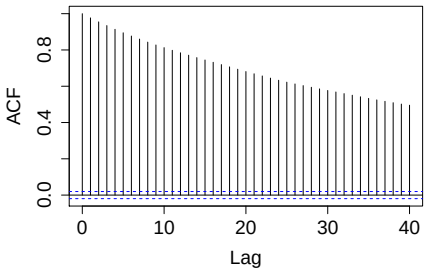
(a) Random Walk, $m = 2$



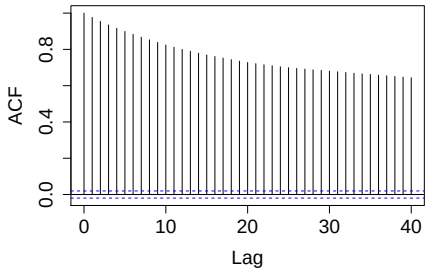
(b) Innovation Scheme, $m = 2$



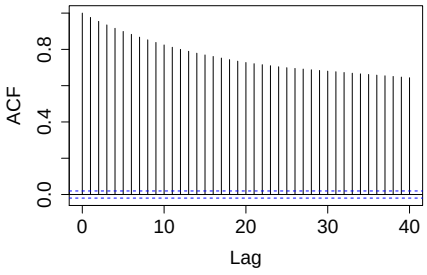
(c) Random Walk, $m = 10$



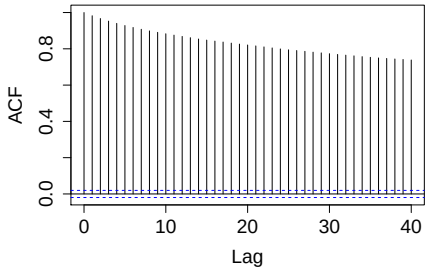
(d) Innovation Scheme, $m = 10$



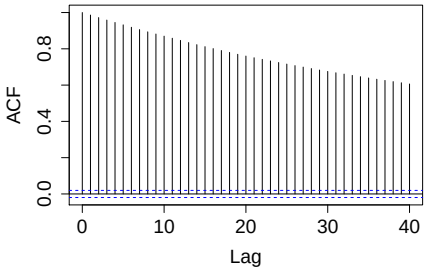
(e) Random Walk, $m = 20$



(f) Innovation Scheme, $m = 20$



(g) Random Walk, $m = 30$



(h) Innovation Scheme, $m = 30$

Figure 5.12: Autocorrelation functions of Random Walk and Innovation Scheme for σ_1 applied to the six dimensional model in Eq. (5.27) with observation interval $\Delta = 1.0$.

	$T = 10$			$T = 100$			$T = 1000$		
	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001
$A_{00} = 0.17$	2.16 (0,6.34)	0.31 (0,0.75)	0.25 (0.06,0.45)	0.01 (0,0.04)	0.15 (0.13,0.16)	0.17 (0.17,0.18)	0.37 (0.11,0.39)	0.28 (0.16,0.28)	0.17 (0.17,0.17)
$A_{01} = 2.23$	6.1 (0.22,11.97)	1.81 (1.09,2.55)	1.99 (1.77,2.22)	1.22 (0.66,1.76)	1.88 (1.69,2.03)	2.15 (2.1,2.22)	0.99 (0.62,1.73)	1.48 (1.3,2.24)	2.23 (2.23,2.23)
$A_{02} = 1.46$	2.44 (0,7.97)	2.01 (1.4,2.53)	1.46 (1.35,1.57)	2.09 (1.76,2.47)	1.63 (1.54,1.72)	1.54 (1.47,1.57)	2.37 (1.58,2.51)	2.13 (1.46,2.17)	1.46 (1.46,1.46)
$A_{03} = 2.58$	6.02 (0.84,14.02)	2.92 (2.26,3.79)	2.69 (2.39,2.97)	2.49 (2.16,2.87)	2.87 (2.53,2.99)	2.7 (2.59,2.73)	8.52 (2.58,8.89)	4.35 (2.59,4.43)	2.58 (2.58,2.58)
$A_{04} = -0.13$	5.09 (−4.26,15.96)	−0.8 (−1.63,0.02)	−0.4 (−0.65,−0.16)	−1.16 (−2.01,−0.27)	−0.33 (−0.5,−0.16)	−0.16 (−0.19,−0.08)	−1.41 (−3.04,0.28)	−1.17 (−1.36,−0.12)	−0.13 (−0.13,−0.13)
$A_{05} = 0.38$	3.25 (−3.42,12.65)	1.06 (0.34,1.6)	0.39 (0.27,0.5)	0.71 (0.34,1.11)	0.4 (0.29,0.5)	0.37 (0.34,0.4)	0.35 (0.23,0.47)	0.55 (0.3,0.59)	0.38 (0.38,0.38)
$A_{06} = 0.32$	−0.01 (−2.08,2.18)	0.41 (0.15,0.7)	0.4 (0.29,0.51)	0.26 (0.22,0.3)	0.36 (0.32,0.37)	0.35 (0.33,0.35)	1.19 (0.36,1.25)	0.56 (0.33,0.57)	0.32 (0.32,0.32)
$A_{07} = 0.1$	3.19 (−2.12,8.87)	−0.11 (−0.64,0.42)	−0.04 (−0.2,0.11)	0.36 (−0.04,0.74)	0.3 (0.21,0.4)	0.19 (0.13,0.23)	0.17 (−0.01,0.81)	0.34 (0.09,0.52)	0.09 (0.09,0.09)
$A_{08} = 0.01$	1.44 (−3.17,5.95)	0.23 (−0.08,0.52)	0.02 (−0.05,0.09)	0.14 (−0.15,0.4)	−0.07 (−0.14,0)	−0.04 (−0.07,−0.01)	0.06 (−0.38,0.18)	−0.03 (−0.3,0.01)	0.01 (0.01,0.01)
	6.41	0.08	0.01	0.11	0.01	0.00	1.21	0.13	0.00

Table 5.2: Diffusion parameter estimates for a two dimensional cubic model with fixed drift function parameters given in Table 5.3. On the left is the true value of the parameter. The length of the data set used for the inference is labelled as T and the observation interval is $\Delta = \{0.1, 0.01, 0.001\}$. There was no missing data in this study. The posteriors were estimated using 3×10^6 samples from three MCMC chains. In each cell the parameter is estimated from the posterior mean and in brackets is shown the 10-90 percentiles of the posterior. The bottom of the table shows the Posterior Expected Loss of Eq. (5.6).

was no missing data in this study. The results estimated from 10^6 MCMC samples are shown in Table 5.2.

A close look at the numbers in Table 5.2 indicates that the estimates are converging to the true values for increasing T and decreasing Δ. Estimates using $T = 10$ and $\Delta = 0.001$ are accurate. Other values $\Delta = \{0.1, 0.01\}$ give posterior estimates that do not reflect this sampling property correctly. Estimates for $T = 100$ are close to the true value but do not reflect the correct sampling. This is true for $T = 1000$ until Δ is as small as 0.001.

5.3.1 Low dimensional noise

In Section 3.5.2 the homogenisation procedure applied to the triad model resulted in a two dimensional model driven by a one dimensional Brownian motion. The model was of the form

$$\begin{aligned} dX_1 &= (\gamma X_1 + \alpha X_1 X_2^2)dt + (\sigma X_1 + \alpha\lambda X_2)dB_t \\ dX_2 &= (\gamma X_2 + \beta X_1^2 X_2)dt + (\beta\lambda X_1 + \sigma X_2)dB_t \end{aligned} \tag{5.29}$$

This form of model causes problems for the method presented above. The problem stems from the diffusion coefficient not being invertible, which means that we can

not write down a unique likelihood function using Girsanov's theorem. As this model arose as one of our toy problems, and could be encountered more generally, we present some simple inference methods for it. We focus on inferring parameter σ and use the short, high frequency data set, shown in Figure 5.13. Firstly notice

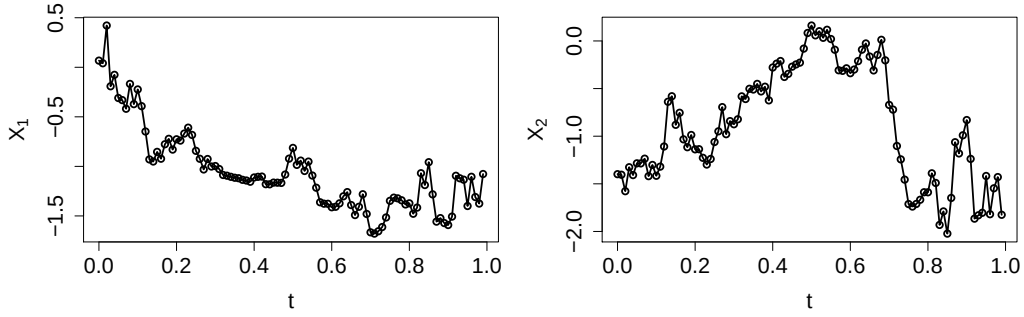


Figure 5.13: Data set used for inference from Eq. (5.29).

that the model can be transformed to a one dimensional process using Ito's formula. Making the change of variables

$$Y_t = \beta X_1^2 - \alpha X_2^2$$

leads to the SDE

$$dY_t = (2\gamma + \sigma^2 - \alpha\beta\lambda^2)Y_t dt + 2\sigma Y_t dB_t \,. \tag{5.30}$$

We implemented Algorithm 4.1 for Eq. (5.30) to infer σ from high frequency observations ($\Delta = 0.01$ and $N = 100$). The estimated posteriors for increasing m, shown in Figure 5.14a, converge towards $\sigma = 1$. Figure 5.14b shows the estimated posteriors for σ using Eq. 5.29, except now replacing the 1d Brownian motion for 2d Brownian motion. The results show that this is not an acceptable approximation in this case as the estimates are incorrect. We will need to consider this problem when applying the Innovation Scheme to models derived using the homogenisation procedure from Chapter 3.

5.4 Inference for Drift Parameters

In this section we give details of the computational implementation of the sampling of parameters in the drift function of Eq. (5.1). We describe the algorithm for a general cubic model and use the Euler approximation for the likelihood function.

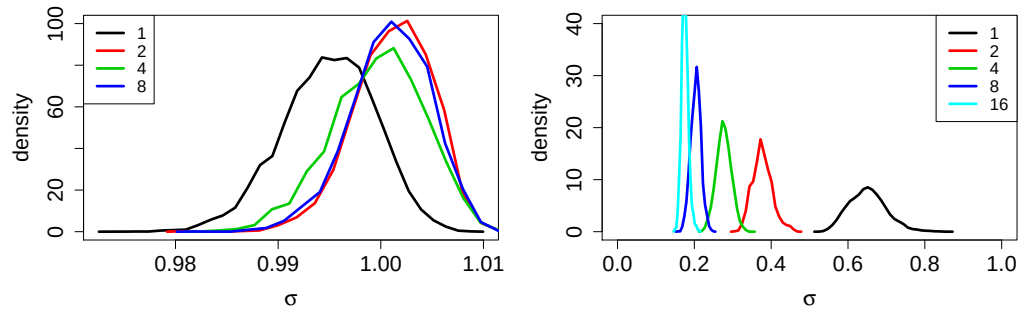


Figure 5.14: Estimates of the posterior distributions of σ. Each curve is an estimate of the posterior for a different amount of imputed data m. A high frequency data set with $N = 100$ observations and interobservation time $\Delta = 1.0$ from Eq. (5.29) was used. The true value is $\sigma = 1.0$. On the left are the results from fitting the 1 dimensional model Eq. (5.30) and on the right are those estimated from Eq. (5.29) with 2-dimensional Brownian motion.

At the start of this chapter we noted that the number of parameters in the drift function of Eq. (5.1) increases as the fourth power of the dimension. Initial work (not shown) inferring these parameters using a Metropolis-Hastings step and random walk proposal indicated that this approach was completely unpractical. Since the drift parameters enter linearly we can construct a Gibbs sampler where their conditional posterior is Gaussian. This greatly improves the mixing of the Markov Chain.

5.4.1 Gibbs Sampler

Consider the general form of a D dimensional cubic SDE with linear noise in Eq. (5.1). The parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are of interest in this section. We write the parameters as one matrix $\boldsymbol{A} \in \mathbb{R}^{D\times P}$. We allow for the inclusion of all possible linear, quadratic and cubic terms in the model with linear terms entering first, followed by quadratic then cubic. We include them into matrix $\boldsymbol{A}$ as $A_{i,1} = \alpha_i$, $A_{i,j+1} = \beta_{i,j}$, $A_{i,j(j-1)/2+k+D+1} = \gamma_{i,j,k}$ and $A_{i,f(j,k,l)} = \lambda_{i,j,k,l}$, where $i, j \in \{1, \ldots, D\}$, $k \in \{1, \ldots, j\}$ and $l \in \{1, \ldots, k\}$. The index function for the cubic terms is

$$f(j,k,l) = 1 + D + D(D+1)/2 + j(j-1)(j+1)/6 + k(k-1)/2 + l\,. \qquad (5.31)$$

This gives the total number of parameters in a row as

$$P = \frac{11D}{6} + D^2 + \frac{D^3}{6}\,.$$

Consider N observations of the discretised system with time interval δ. We set $Y_t = X_{t+1} - X_t$ and let $U \in \mathbb{R}^{N-1\times P}$ be the design matrix of the data, scaled by δ. The columns of U are indexed in the same way as the columns of parameter matrix A. For example, a two dimensional system would have $P = 10$ and design matrix

$$U = \delta \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & X_{1,1}^2 & X_{1,1}X_{1,2} & X_{1,2}^2 & X_{1,1}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N-1,1} & X_{N-1,2} & X_{N-1,1}^2 & X_{N-1,1}X_{N-1,2} & X_{N-1,2}^2 & X_{N-1,1}^3 \end{pmatrix}$$

$$\begin{pmatrix} X_{1,1}^2X_{1,2} & X_{1,1}X_{1,2}^2 & X_{1,2}^3 \\ \vdots & \vdots & \vdots \\ X_{N-1,1}^2X_{N-1,2} & X_{N-1,1}X_{N-1,2}^2 & X_{N-1,2}^3 \end{pmatrix}$$

The log likelihood can be written

$$L(A;X) = -\frac{1}{2}\sum_{t=1}^{N-1}|\Sigma_t| - \frac{1}{2}\sum_{t=1}^{N-1}\sum_{i,j=1}^{D}\left(Y_{ti} - \sum_{k=1}^{P}U_{tk}A_{ik}\right)\Sigma_{tij}^{-1}\left(Y_{tj} - \sum_{k=1}^{P}U_{tk}A_{jk}\right)$$

where the instantaneous covariance matrix $\mathbf{\Sigma}_t$ is computed from $\Sigma_{t,j,k}^{1/2} = (a_{j,k} + \sum_{l=1}^{D} b_{l,j,k}X_{t,j})\Delta^{1/2}$.

We have DP parameters to infer in the matrix $\boldsymbol{A}$. We set zero mean Gaussian prior with covariance matrix $\mathbf{\Gamma} \in \mathbb{R}^{DP\times DP}$. Let $\Lambda \in \mathbb{R}^{DP\times DP}$ be a matrix with components

$$\Lambda_{(i-1)P+j,(k-1)P+l} = \sum_{t=1}^{N-1}U_{tj}\Sigma_{tik}^{-1}U_{tl} + \Gamma_{(i-1)P+j,(k-1)P+l}^{-1}$$

where $i,k = 1\ldots D$ and $j,l = 1\ldots P$. Let $b \in \mathbb{R}^{DP}$ with components

$$b_{(i-1)P+j} = \sum_{t,k}U_{t,j}\Sigma_{tik}^{-1}Y_{tk}.$$

The posterior mean $\mu_{(i-1)P+j}$ of $A_{i,j}$ is given by the solution of $\mathbf{\Lambda\mu} = \boldsymbol{b}$ and the posterior covariance is $\mathrm{Cov}(A_{i,j}, A_{k,l}) = \Lambda_{(i-1)P+j,(k-1)P+l}^{-1}$.

We applied the above Gibbs sampler to a large data set from a two dimen-

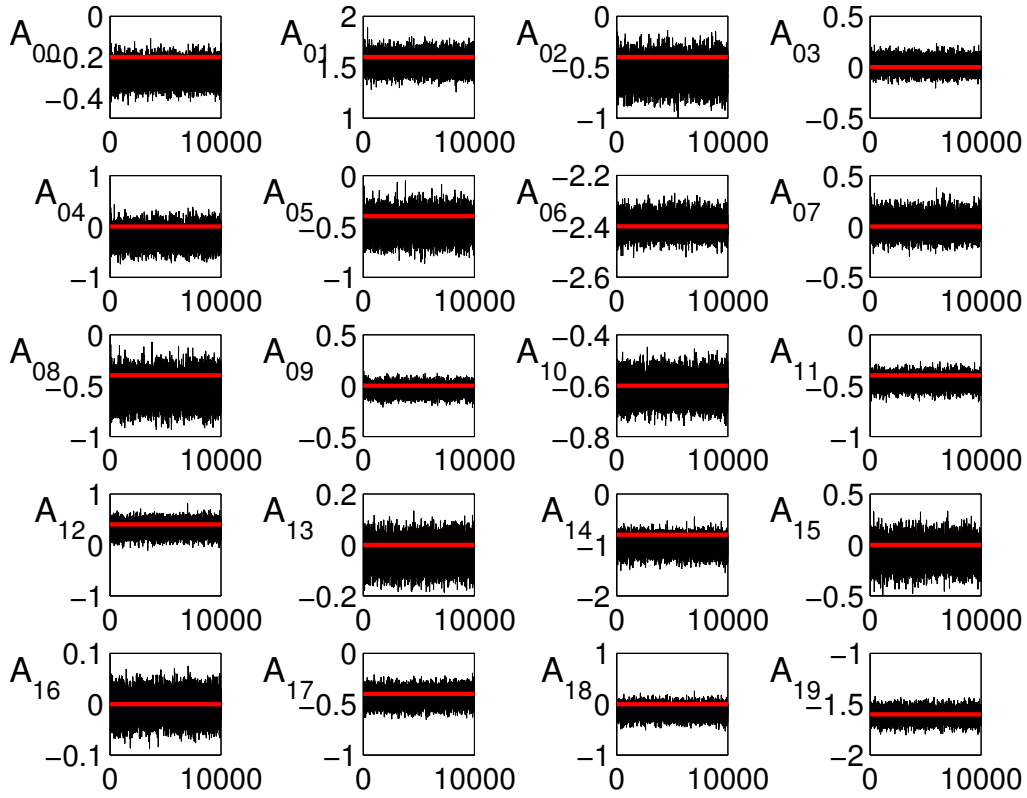


Figure 5.15: Output of Gibbs sampler for 20 drift parameters of two dimensional model from Eq. (5.1). The observation interval is $\delta = 10^{-3}$ and $T = 10,000$. The true values are shown in red.

sional model with random values for the diffusion parameters. We chose a fine observation interval of $\delta = 10^{-3}$ and long observation period $T = 10,000$. Figure 5.15 displays the trace plots for all 20 parameters (note that the indices are from 0 rather than 1 as in the text). Using this large data set the algorithm is able to reproduce the true values shown in red.

We performed a further simulation study to test the dependence of the posterior estimates upon the data set used. We inferred all of the drift parameters for a simple two dimensional cubic model using data sets of length $T = \{10, 100, 1000\}$ and with observation interval $\Delta = \{0.1, 0.01, 0.001\}$. Note that the diffusion parameters are fixed in this model to the values in Table 5.2 and there is no missing data. The results are shown in Table 5.3. For each parameter we estimated the posterior mean and the posterior 10-90 percentile. The error of the estimates was quantified using the Posterior Expected Loss (PEL) of Eq. (5.6). Table 5.3 indicates that all of the estimated posterior distributions of the parameters are of similar range and

	$T = 10$			$T = 100$			$T = 1000$		
	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001
$A_{00} = 0$	2.18 (−2.5,6.89)	−0.44 (−4.79,3.74)	−1.95 (−6.07,2.2)	−0.13 (−1.07,0.77)	0.65 (−0.29,1.6)	0.64 (−0.28,1.58)	0.02 (−0.13,0.17)	0.01 (−0.13,0.17)	0.02 (−0.13,0.18)
$A_{01} = 5$	−2.04 (−7.23,3.04)	4.86 (0.84,9.04)	6.21 (2.31,10.12)	2.54 (1.78,3.28)	4.73 (3.93,5.51)	5.22 (4.43,6.03)	2.84 (2.7,2.97)	4.59 (4.45,4.73)	4.86 (4.73,5)
$A_{02} = 0$	2.63 (−0.16,5.39)	2.15 (−0.17,4.49)	1.63 (−0.66,4)	0.45 (−0.52,1.41)	0.58 (−0.42,1.6)	0.13 (−0.89,1.13)	−0.08 (−0.22,0.05)	−0.14 (−0.27,0)	−0.19 (−0.32,−0.04)
$A_{03} = 0$	1.54 (−2.75,5.84)	−0.17 (−3.12,2.76)	0.26 (−2.62,3.2)	0.27 (−0.38,0.93)	0.01 (−0.72,0.73)	−0.17 (−0.88,0.54)	−0.01 (−0.08,0.06)	0 (−0.07,0.06)	0 (−0.07,0.06)
$A_{04} = 0$	−3.51 (−8.17,1.13)	0.61 (−3,4.3)	1.3 (−2.32,5.07)	−0.96 (−2.09,0.2)	−0.58 (−1.8,0.68)	−0.07 (−1.3,1.16)	−0.01 (−0.04,0.01)	−0.02 (−0.04,0.01)	−0.02 (−0.05,0.01)
$A_{05} = 0$	0.32 (−2.19,2.81)	0 (−2.31,2.31)	0.19 (−2.11,2.49)	0.66 (−0.11,1.43)	0.02 (−0.79,0.84)	−0.27 (−1.05,0.55)	0 (−0.06,0.07)	0.01 (−0.06,0.07)	0 (−0.07,0.07)
$A_{06} = -3$	−0.88 (−2.96,1.24)	−2.58 (−4.3,−0.87)	−3.17 (−4.83,−1.49)	−1.8 (−2,−1.6)	−2.92 (−3.12,−2.71)	−3.05 (−3.25,−2.85)	−1.71 (−1.77,−1.65)	−2.75 (−2.81,−2.69)	−2.91 (−2.97,−2.85)
$A_{07} = 0$	0.58 (−1.87,3.06)	−1.32 (−3.32,0.7)	−1.48 (−3.54,0.52)	0.05 (−0.49,0.58)	−0.07 (−0.66,0.51)	−0.22 (−0.79,0.37)	−0.02 (−0.07,0.04)	−0.01 (−0.06,0.05)	0 (−0.05,0.06)
$A_{08} = 0$	−0.83 (−3,1.31)	−0.28 (−2.31,1.76)	−0.18 (−2.21,1.86)	−0.48 (−1.03,0.07)	−0.37 (−0.96,0.23)	−0.16 (−0.75,0.42)	−0.02 (−0.08,0.03)	−0.01 (−0.06,0.04)	−0.01 (−0.07,0.04)
$A_{09} = 0$	0.71 (−0.11,1.53)	−0.4 (−1.16,0.38)	−0.49 (−1.26,0.27)	0.33 (−0.02,0.67)	0.04 (−0.3,0.39)	0.09 (−0.25,0.44)	0.08 (0.02,0.13)	0.1 (0.04,0.16)	0.12 (0.05,0.17)
$A_{10} = 0$	−0.38 (−5.19,4.24)	−2.21 (−6.44,2.03)	−1.17 (−5.21,2.9)	−0.31 (−1.22,0.61)	−0.18 (−1.1,0.75)	0.06 (−0.89,0.99)	0.02 (−0.13,0.17)	0.05 (−0.1,0.2)	0.05 (−0.1,0.2)
$A_{11} = 0$	−2.93 (−8.11,2.15)	2.42 (−1.64,6.48)	1.67 (−2.29,5.64)	0.88 (0.1,1.64)	0.39 (−0.4,1.19)	0.33 (−0.48,1.12)	−0.06 (−0.2,0.08)	0.11 (−0.03,0.25)	0.11 (−0.04,0.24)
$A_{12} = 5$	5.96 (3.22,8.68)	4.2 (1.89,6.63)	4.72 (2.33,7.09)	1.68 (0.72,2.65)	4.56 (3.52,5.57)	5.09 (4.08,6.12)	2.88 (2.74,3.02)	4.7 (4.56,4.84)	4.98 (4.84,5.11)
$A_{13} = 0$	2.68 (−1.67,6.93)	−1.41 (−4.36,1.57)	−1.88 (−4.74,1.02)	−1.05 (−1.72,−0.38)	−0.61 (−1.33,0.1)	−0.74 (−1.46,−0.02)	−0.01 (−0.08,0.05)	−0.03 (−0.1,0.03)	−0.03 (−0.09,0.04)
$A_{14} = 0$	−1.13 (−5.84,3.6)	4.67 (1.01,8.45)	4.23 (0.47,7.92)	1.39 (0.24,2.55)	0.78 (−0.44,2.03)	0.91 (−0.33,2.14)	0.02 (−0.01,0.05)	0.03 (0,0.05)	0.03 (0,0.05)
$A_{15} = 0$	0.42 (−2.07,2.94)	−0.6 (−2.88,1.77)	−0.84 (−3.13,1.45)	−0.22 (−1.01,0.56)	−0.11 (−0.9,0.7)	−0.24 (−1.06,0.55)	0 (−0.07,0.06)	−0.01 (−0.08,0.05)	−0.02 (−0.09,0.04)
$A_{16} = 0$	−0.82 (−2.94,1.29)	0.12 (−1.61,1.83)	0.39 (−1.26,2)	−0.08 (−0.28,0.12)	−0.04 (−0.24,0.17)	−0.02 (−0.22,0.19)	−0.03 (−0.09,0.03)	−0.08 (−0.14,−0.02)	−0.09 (−0.15,−0.03)
$A_{17} = 0$	−0.46 (−2.95,2.04)	−3.21 (−5.28,−1.22)	−2.97 (−4.99,−0.93)	−0.71 (−1.25,−0.17)	−0.66 (−1.24,−0.08)	−0.78 (−1.35,−0.19)	0.01 (−0.05,0.06)	−0.01 (−0.06,0.04)	0 (−0.05,0.05)
$A_{18} = 0$	0.66 (−1.5,2.86)	1.33 (−0.71,3.35)	1.63 (−0.41,3.6)	0.65 (0.11,1.22)	0.45 (−0.13,1.04)	0.58 (−0.01,1.18)	0.06 (0.01,0.12)	0.02 (−0.03,0.08)	0.03 (−0.02,0.09)
$A_{19} = -3$	−2.61 (−3.44,−1.78)	−3.23 (−4.02,−2.45)	−3.47 (−4.24,−2.69)	−1.46 (−1.8,−1.12)	−2.73 (−3.07,−2.38)	−3.04 (−3.39,−2.7)	−1.75 (−1.81,−1.69)	−2.81 (−2.87,−2.75)	−2.98 (−3.04,−2.92)
	8.48	5.25	4.96	1.19	0.37	0.36	0.43	0.02	0.01

Table 5.3: Drift parameter estimates for a two dimensional cubic model with fixed diffusion function parameters given by the values in Table 5.2 and no missing data. On the left is the true value of the parameter. The length of the data set used for the inference is labelled as T and the observation interval is $\Delta = \{0.1, 0.01, 0.001\}$. The posteriors were estimated using 3×10^6 samples from three MCMC chains. In each cell the parameter is estimated from the posterior mean and in brackets is shown the 10-90 percentiles of the posterior. The bottom of the table shows the Posterior Expected Loss of Eq. (5.6).

the PEL should not be biased towards any one parameter but serve as a measure for the whole parameter vector.

The table demonstrates the reduction in PEL as the sample length T and sampling frequency $1/\Delta$ increases. There are large errors for these estimates when using either a short or sparsely sampled data set. Although the PEL for the case where $T = 1000$ and $\Delta = 0.1$ do not seem large, the estimates are still far from the true value. The estimate converge for data sets with $T = 1000$ and $\Delta = \{0.01, 0.001\}$. Note that a few days of computation time was needed to draw 10^6

	$T = 10$			$T = 100$			$T = 1000$		
	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001
$A_{00} = 0$	2.18 (−2.5,6.89)	−2.46 (−8.46,3.52)	−2.59 (−8.35,3.24)	−0.13 (−1.07,0.77)	−0.68 (−1.84,0.5)	−0.69 (−1.85,0.49)	0.02 (−0.13,0.17)	−0.08 (−0.27,0.09)	−0.08 (−0.26,0.09)
$A_{01} = 5$	−2.04 (−7.23,3.04)	4.99 (−0.7,10.61)	4.84 (−0.49,10.22)	2.54 (1.78,3.28)	4.89 (3.83,5.97)	4.76 (3.74,5.79)	2.84 (2.7,2.97)	5.04 (4.87,5.21)	4.88 (4.71,5.03)
$A_{02} = 0$	2.63 (−0.16,5.39)	0.61 (−2.94,4.05)	0.39 (−2.95,3.75)	0.45 (−0.52,1.41)	−0.06 (−1.42,1.3)	−0.2 (−1.49,1.13)	−0.08 (−0.22,0.05)	−0.26 (−0.44,−0.08)	−0.28 (−0.45,−0.09)
$A_{03} = 0$	1.54 (−2.75,5.84)	3.25 (−1.23,8.07)	3.26 (−1.07,7.77)	0.27 (−0.38,0.93)	0.67 (−0.34,1.67)	0.59 (−0.41,1.59)	−0.01 (−0.08,0.06)	0.02 (−0.06,0.1)	0.02 (−0.06,0.09)
$A_{04} = 0$	−3.51 (−8.17,1.13)	−2.53 (−7.76,2.61)	−2.53 (−7.61,2.41)	−0.96 (−2.09,0.2)	−1.6 (−3.35,0.18)	−1.41 (−3.14,0.27)	−0.01 (−0.04,0.01)	−0.03 (−0.05,0)	−0.02 (−0.05,0)
$A_{05} = 0$	0.32 (−2.19,2.81)	1.85 (−1.6,5.2)	1.85 (−1.43,5.11)	0.66 (−0.11,1.43)	1.18 (0.03,2.32)	1.12 (−0.02,2.25)	0 (−0.06,0.07)	0.05 (−0.04,0.13)	0.05 (−0.04,0.13)
$A_{06} = -3$	−0.88 (−2.96,1.24)	−4.06 (−6.63,−1.76)	−4.02 (−6.5,−1.79)	−1.8 (−2,−1.6)	−3.19 (−3.44,−2.94)	−3.06 (−3.3,−2.83)	−1.71 (−1.77,−1.65)	−3.01 (−3.08,−2.94)	−2.92 (−2.98,−2.85)
$A_{07} = 0$	0.58 (−1.87,3.06)	0.44 (−2.27,3.17)	0.53 (−2.05,3.23)	0.05 (−0.49,0.58)	0.43 (−0.37,1.27)	0.39 (−0.42,1.19)	−0.02 (−0.07,0.04)	0.03 (−0.04,0.09)	0.03 (−0.03,0.09)
$A_{08} = 0$	−0.83 (−3,1.31)	−2.22 (−4.98,0.62)	−2.12 (−4.84,0.65)	−0.48 (−1.03,0.07)	−1 (−1.86,−0.14)	−0.92 (−1.76,−0.09)	−0.02 (−0.08,0.03)	−0.02 (−0.09,0.05)	−0.02 (−0.09,0.05)
$A_{09} = 0$	0.71 (−0.11,1.53)	1.29 (0.06,2.61)	1.33 (0.1,2.6)	0.33 (−0.02,0.67)	0.7 (0.17,1.2)	0.71 (0.19,1.2)	0.08 (0.02,0.13)	0.14 (0.05,0.23)	0.15 (0.06,0.23)
$A_{10} = 0$	−0.38 (−5.19,4.24)	−4.63 (−10.46,0.84)	−4.46 (−10.07,0.86)	−0.31 (−1.22,0.61)	0.66 (−0.55,1.87)	0.6 (−0.51,1.75)	0.02 (−0.13,0.17)	−0.02 (−0.2,0.16)	−0.01 (−0.2,0.16)
$A_{11} = 0$	−2.93 (−8.11,2.15)	2.18 (−3.8,8.3)	2.24 (−3.53,8.27)	0.88 (0.1,1.64)	0.1 (−0.89,1.09)	0.15 (−0.82,1.14)	−0.06 (−0.2,0.08)	0.08 (−0.09,0.26)	0.11 (−0.06,0.29)
$A_{12} = 5$	5.96 (3.22,8.68)	8.28 (4.62,11.9)	7.82 (4.39,11.36)	1.68 (0.72,2.65)	5.8 (4.52,7.07)	5.57 (4.36,6.81)	2.88 (2.74,3.02)	5.05 (4.88,5.22)	4.89 (4.72,5.05)
$A_{13} = 0$	2.68 (−1.67,6.93)	4.03 (−0.84,9.1)	4.08 (−0.63,8.96)	−1.05 (−1.72,−0.38)	−0.83 (−1.76,0.11)	−0.81 (−1.72,0.1)	−0.01 (−0.08,0.05)	−0.01 (−0.09,0.08)	−0.01 (−0.09,0.08)
$A_{14} = 0$	−1.13 (−5.84,3.6)	0.2 (−5.79,6.51)	0.02 (−5.85,6.12)	1.39 (0.24,2.55)	0.12 (−1.44,1.66)	0.07 (−1.43,1.61)	0.02 (−0.01,0.05)	0.03 (0,0.06)	0.03 (0,0.06)
$A_{15} = 0$	0.42 (−2.07,2.94)	1.49 (−1.84,5)	1.59 (−1.64,4.85)	−0.22 (−1.01,0.56)	0.15 (−0.89,1.18)	0.2 (−0.8,1.19)	0 (−0.07,0.06)	0.01 (−0.07,0.08)	0 (−0.07,0.08)
$A_{16} = 0$	−0.82 (−2.94,1.29)	−3.07 (−5.88,−0.39)	−3.14 (−5.86,−0.5)	−0.08 (−0.28,0.12)	−0.15 (−0.45,0.15)	−0.15 (−0.45,0.15)	−0.03 (−0.09,0.03)	−0.07 (−0.15,0.02)	−0.08 (−0.17,0)
$A_{17} = 0$	−0.46 (−2.95,2.04)	−1.37 (−4.65,1.87)	−1.17 (−4.45,1.91)	−0.71 (−1.25,−0.17)	−0.59 (−1.36,0.19)	−0.57 (−1.31,0.17)	0.01 (−0.05,0.06)	0.05 (−0.02,0.12)	0.05 (−0.03,0.12)
$A_{18} = 0$	0.66 (−1.5,2.86)	0.3 (−2.65,3.15)	−0.01 (−2.73,2.67)	0.65 (0.11,1.22)	0.24 (−0.51,0.99)	0.18 (−0.54,0.92)	0.06 (0.01,0.12)	0.02 (−0.04,0.08)	0.02 (−0.04,0.08)
$A_{19} = -3$	−2.61 (−3.44,−1.78)	−4.44 (−5.58,−3.38)	−4.1 (−5.16,−3.13)	−1.46 (−1.8,−1.12)	−3.09 (−3.53,−2.66)	−2.94 (−3.36,−2.55)	−1.75 (−1.81,−1.69)	−3.06 (−3.14,−2.99)	−2.97 (−3.04,−2.9)
	8.48	11.06	10.32	1.19	0.76	0.68	0.43	0.01	0.01

Table 5.4: Drift parameter estimates for a two dimensional cubic model with diffusion function parameters given by the values in Table 5.2. On the left is the true value of the parameter. The data used is the same as that of Table 5.3 sampled at the $\Delta = 0.1$ interval. In this case data is imputed to obtain the intervals $\Delta = \{0.01, 0.001\}$. The Modified Bridge sampler was used to impute data (see Table 5.1). The posteriors were estimated using 3×10^6 samples from three MCMC chains. The bottom of the table shows the Posterior Expected Loss of Eq. (5.6).

samples for the case $T = 1000$ and $\Delta = 0.001$: this is a total of $N = 10^6$ data points in the time series.

We also performed a test with both the Gibbs sampler and data imputation. Table 5.4 is the same as Table 5.3 except now in all cases the data is observed at interval $\Delta = 0.1$. The smaller intervals $\Delta = \{0.01, 0.001\}$ are obtained by imputing data with $m = \{10, 100\}$ respectively. For this study we used the Modified Bridge sampler (see Table 5.1) to impute missing data. The table shows that imputing data approximately doubles the Posterior Expected Loss compared to Table 5.3. As expected the confidence intervals are broader but with more imputed data the algo-

rithm can recover the true values. This is a reassuring test of the data imputation strategy.

The aim of this thesis is to infer models that can be used for prediction. This can be problematic when dealing with non-linear models as some (generally unknown) regions of the parameter space will give solutions that explode to infinity with probability 1. This is a particular problem when, as exemplified by Table 5.3, large amounts of data are needed to regain the true values.

To demonstrate this problem we performed an inference on a two dimensional cubic model using $N = 1,000$ observations at $\Delta = 0.1$. For each inferred parameter value we then simulated the solution for $T = 100$. After this time we recorded whether the solution retained finite values or had exploded. The marginal posterior distributions of the cubic parameters are plotted in Figure 5.16. Each plot shows two histograms: one in blue records the distribution of stable parameter values and in mauve are those that exploded. The overlapping region is shown in purple. Notice that, when looking at the marginal distributions, the stable and unstable regions largely overlap; it is difficult to separate the two regions. In this case 40% of values were unstable. Tests (not shown) indicate that this is an even bigger problem in higher dimensions. Therefore, it is worth constructing some constraints on the parameter space to enforce the solutions to remain finite. In Chapter 7 we propose a solution to this problem.

5.5 GPU Computing

In this section we describe the implementation of the Algorithms 4.2 and 4.1 for sampling parameters and missing data on a Graphics Processing Unit (GPU) using NVIDIA's CUDA parallel programming environment. Recently statisticians have realised the potential for reduced computational time that can be achieved using massively parallel computation using GPUs. The GPU, having developed from increasing computational demand on graphics rendering within the gaming and video editing industry, is specialised at Single Instruction Multiple Data (SIMD) tasks. Compared with the conventional CPU it devotes more transistors to arithmetic instructions and less to data caching and flow control and so are suitable for algorithms with high arithmetic intensity, few branching statements or calls to memory.

The GPU consists of a number of Streaming Multiprocesses (SM) each of which has a limited amount of fast on-chip shared memory. An SM is capable of performing an operation on 32 threads simultaneously whilst holding the remaining threads in memory. For the purpose of compatibility with different GPU configura-

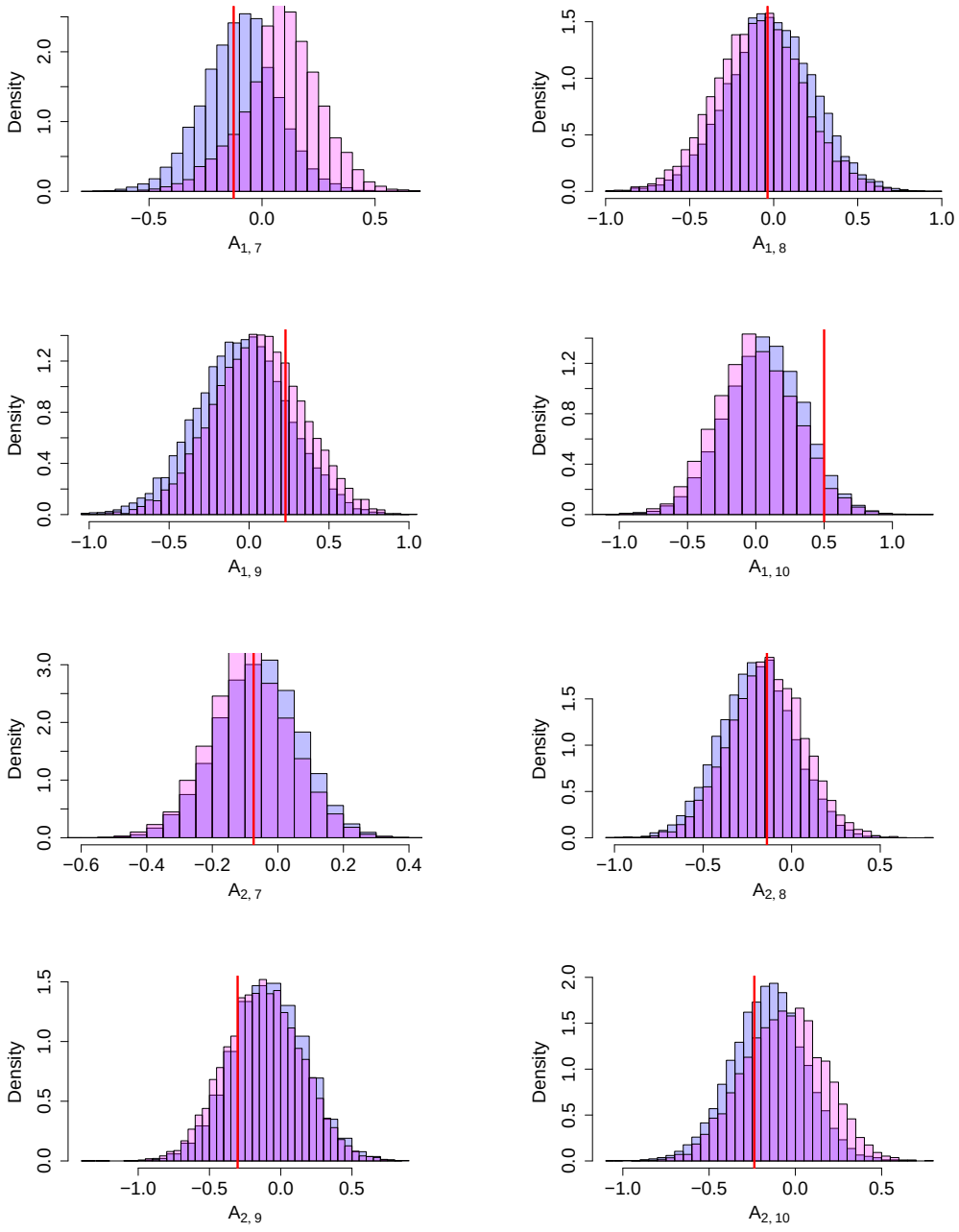


Figure 5.16: Marginal distributions of cubic parameters inferred for a two dimensional model of form Eq. 5.1. A data set with $N = 1,000$ observations at interval $\Delta = 0.1$ was used. The diffusion parameters were fixed and there was no missing data. The blue histogram shows the parameters that gave stable solutions to the SDE, while the mauve is for those that gave unstable solutions. The purple shows the overlap between the two regions of the marginal distributions. The true values are given by the red lines.

tions the CUDA programming model groups threads into thread blocks. Currently up to 1024 threads can be contained in a single thread block. Each block is sent to a SM and instructions on 32 threads executed simultaneously. The execution of these groups of threads hides the latency associated with memory request operations.

Thread Blocks are organised into a Grid. Whilst threads within a block have a limited amount of fast local on-chip memory, blocks within a grid only share access to global device memory which is relatively slow with latency at the order of 100 clock cycles. Threads within a block can be synchronised and communication between them is fast. Blocks communicate by transferring through the CPU which is slow.

Statisticians considering implementing their algorithm on a GPU should take these hardware factors into account when designing there parallel code. Another consideration is the use of single or double floating point arithmetic. GPUs were originally designed to use single precision but with the recent demand for general purpose GPU computation more recent models have double precision capability. However, single precision remains at least 3-4 times faster, although this may come down in the future.

When converting a statistical algorithm for massively parallel computation one should consider how to decompose the problem into identical operations that can be performed with little dependence between them. Many data intensive applications in statistics are amenable to this sort of alteration. For example, Suchard et al. [2010] demonstrate the gains of using a GPU on a Bayesian mixture problem. Given a Gaussian mixture density they estimate the mean, variance and weight of each component. The inference algorithm is simplified by using a data augmentation strategy which structures the problem to be soluble by Gibbs sampling. Each data point is assigned a configuration variable. At each stage of the algorithm the posterior configuration probabilities are computed. For a lot of data and many mixture components the number of configuration probabilities becomes very large. They implement a fine grained parallelisation strategy where each data point-configuration pair is assigned a dedicated thread. They describe their choice of execution plan to optimise the use of shared memory and minimise latency associated with transfers between global and shared memory. Given that the amount of shared memory is only 16KB they describe the efficient method of memory transfers to global memory by coalescing transactions into multiples of 16. After considering these hardware details they report a 120 speed up over the standard algorithm implemented on a single CPU.

A different approach to dividing an algorithm for parallelisation is described

by Lee et al. [2010]. They describe various Monte Carlo methods that can be parallelised instead of parallelising the data as in the previous example. They show how easy it is to implement an importance sampling algorithm for the GPU by computing the importance weight of a sample by a single thread. They note that the standard Metropolis-Hastings algorithm does not gain much by parallelisation as it is an inherently serial algorithm although population MCMC and particle samplers work well. They split a population MCMC algorithm so that each thread samples a different distribution with reversible swaps between chains. They have thousands of chains simulating from tempered distributions with only a single chain sampling the target density. Applied to a mixture model they show the improved mixing of the chain between widely separated modes. They also demonstrate a Sequential Monte Carlo algorithm where, like the importance sampling example, each thread updates the weights for each particle. The authors note that there was little reduction in accuracy by using only single precision. For large numbers of Monte Carlo samples they report a speed increase of approximately 280 over the CPU implementation.

The inference procedure in this thesis transfers naturally to a GPU implementation. The Markov nature of SDE data implies that the data set can be divided into independent blocks. In our implementation each thread is responsible for a single observation interval. The imputed data within that interval is sampled using the independence sampler proposal by a single thread. Each thread has an ID and uses this to reference its section of data.

The algorithm is split into two steps. Firstly the update of missing data and secondly the sampling of parameters. For perfect observation of the process, each thread in the first step can run without communication with threads responsible for neighbouring data intervals. If there is observation error then the data at the observation time needs to be passed between threads causing a potential bottleneck in this step of the algorithm. We only consider the case of perfect observation.

The sampling of parameters is a global operation as it involves all of the data in the likelihood function. However, again due to the Markov property, each thread can compute the likelihood for a single data block. When this is done the threads need to synchronise before all the values can be added to form a single likelihood value. This is an example of a parallel reduction algorithm and is computed using a tree structure. Each evenly numbered thread receives a value from its neighbouring thread and adds it to its own. Then every four threads sum their values and so on until there is just a single likelihood value. This is then added to the prior which can be computed by a single thread. The pseudo code for the update of missing data and parameters using the innovation scheme is shown below.

Algorithm 5.1 Parallel SDE inference with perfect observations. For each step Y has $m+1$ components and is stored in local memory, unique to each thread. For the second step σ^* is stored in shared memory so is accessible to all threads.

τ=blockDim.x blockIdx.x + threadIdx.x
$Y_0 = X_{\tau m}$, $Y_m = X_{\tau m+m}$
$\alpha = 0$
for $i = 0$ to $m-2$ **do**
  $Y_{i+1} \sim q(Y_{i+1}|Y_i, X_{\tau m+m}, \sigma)$ where $q(\cdot|\cdot)$ is one of the bridge distributions discussed in Table 5.1, Section 5.2
  $\alpha = \alpha + L(Y_{[0:m]}|\sigma) - L(X_{\tau m:\tau m+m}|\sigma)$
end for
where $L(Y_{[0:m]}|\sigma)$ is the log likelihood function.
$Y_{[0:m]}$ is accepted with probability $\exp(\alpha)$.
if $\tau = 0$ **then**
  $\sigma^* = \sigma + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \eta)$ and η is a tuning parameter.
end if
τ=blockDim.x blockIdx.x + threadIdx.x
$Y_0 = X_{\tau m}$
$B_0 = 0$, $Y_0 = X_{\tau m}$
for $i = 0$ to $m-2$ **do**
  $B_{i+1} = f^{-1}(Y_{i+1}, \sigma)$ where $f(\cdot)$ is one of the transformations for the innovation scheme discussed in Section 5.3.
  $Y_{i+1} = f(W_{i+1}, \sigma^*)$
end for
$\alpha_\tau = L(Y_{0:m}|\sigma^*) - L(X_{\tau m:\tau m+m}|\sigma) + |J(f(X,\sigma))| - |J(f(Y,\sigma))|$ where $J(\cdot)$ is the Jacobian for $f()$.
SYNCTHREADS
for $i = 1$ to BlockDim.x -1 **do**
  if $(\tau = 0) \mod 2i$ **then**
    $\alpha_\tau = \alpha_\tau + \alpha_{\tau+i}$
  end if
  SYNCTHREADS
end for
$\tau = 0$
$\alpha_0 = \alpha_0 + \pi(\sigma^*) - \pi(\sigma)$, the prior distributions. Set $\sigma = \sigma^*$ and $X = Y$ with probability $\exp(\alpha)$.

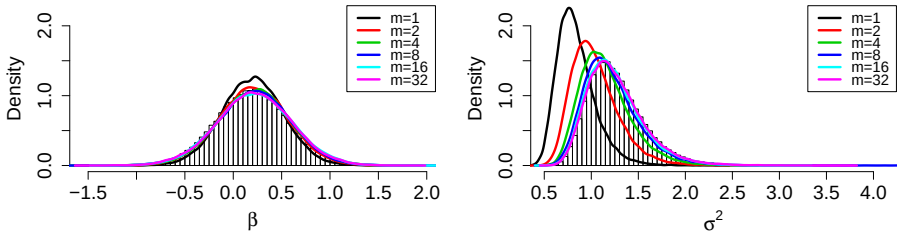


Figure 5.17: Posterior distributions for parameters from the O-U process Eq. (4.9) output from the GPU implementation of Algorithm 5.1 (solid lines) compared with the exact posterior distributions (histograms). Parameters were estimated using a data set with $N = 100$ observations and interobservation time $\Delta = 0.1$. A single long run of 10^5 MCMC samples were used to compute the posteriors.

At present the algorithm only applies to univariate processes but could easily be generalised. Each thread of the algorithm proposes new data $\boldsymbol{Y}_{[0:m]}$ for a sequence of missing data $X_{\tau m+[0:m]}$ indexed by a parameter τ. This is calculated as $\tau = blockDim.x blockIdx.x + threadIdx.x$, where $blockIdx.x$ indexes the thread block of the data, $blockDim.x$ is the size of the thread block and $threadIdx.x$ is the thread identifier. The value $\tau = 0$ is the master thread and performs global operations that need to be computed only once.

The first part of the algorithm, for imputing missing data, divides into independent threads so there is almost a linear increase in computational efficiency with number of threads for any given value of m. However, this is limited by the number of threads per block. The second stage, updating parameters, is slower as each thread requires access to some shared memory to read the updated parameters and there is a reduction step to calculate the global likelihood value.

Initially we tested our implementation of Algorithm 5.1 on a GPU by applying it to the one dimensional OU-process model of Eq. (4.9). We used a data set with $N = 100$ observations and interobservation time $\Delta = 0.1$. We used the Modified Bridge proposal of Table 5.1 to impute the missing data. We compared the parameter estimates with those of the exact posterior distributions. The results, shown in Figure 5.17, demonstrate that the estimated posteriors converge to the true distributions for increasing m.

Figure 5.18 compares the real computational time of the GPU with the CPU implementations. Each plot shows how the running time increases with the amount of imputed data m. Notice that for small amounts of data, $N < 65$, the CPU implementation is faster. This is because the small number of threads does not

compensate for the increased overheads and reduced clock speed of the GPU implementation. The potential of the parallel algorithm is demonstrated for values $N > 65$. Here, although both algorithms are linear in m, the GPU implementation is much faster. This is particularly true for large m with speed increases of factor 5 or more. On this particular GPU (in a standard laptop) the speed increase are not realised for $N > 257$. This is because, as mentioned previously, the algorithm will have to use multiple thread blocks so the threads would not have access to the same shared memory. As scientific computing expands its use of GPUs the number of threads per block should rise and the amount of shared memory increase.

5.6 Summary and Conclusions

In this chapter we have focussed on practical issues related to sampling the parameters from a model of the form Eq. (5.1). We focussed upon developing one specific method, namely the Innovation Scheme of Dargatz [2010] and Golightly and Wilkinson [2008]. The theoretical motivation for this algorithm was discussed in Chapter 4. In particular it is one approach to overcoming the degeneracy issue of SDE inference algorithms as the amount of missing data increases. However, many practical issues remain when estimating posterior distributions for parameters in such a large model as Eq. (5.1). Basic Markov Chain Monte Carlo algorithms are inefficient at exploring the mass of the posterior distributions due to the high dimensional space of missing data which has a complicated correlation structure. The standard method of imputing this missing data would be to use the Modified Bridge of Durham and Gallant [2002]. We found this insufficient when working with multivariate problems. Therefore, we developed the Linear Bridge sampler, which starts from the linearised equation and calculates a bridge distribution using recent work of Barczy and Kern [2010]. We demonstrated that this was significantly more efficient than the standard Modified Bridge.

In Section 5.3 we demonstrated that the Innovation scheme is significantly more efficient at inferring diffusion parameters than a standard random walk. However, we also showed that this algorithm deteriorates rapidly with increasing dimension and so this remains a problem if one wanted to infer parameters for a climate model with greater than 5 dimensions. In Section 5.3.1 we remarked upon an issue for likelihood based inference for models where the Brownian motion has less components than the observed variables. In this case it is not always possible to calculate the likelihood function. This arises as a problem for our applications in Chapter 8.

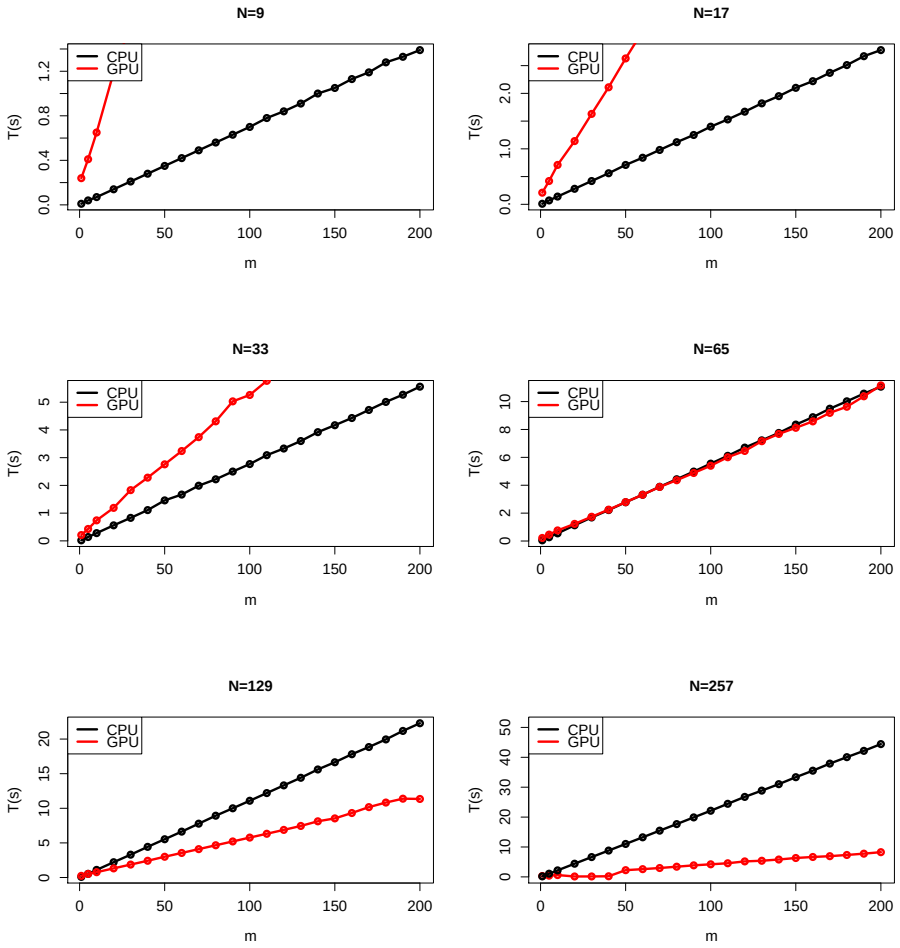


Figure 5.18: Real computation times to draw 1000 MCMC samples from the posterior distribution of the OU process for various size data sets. The time in seconds is plotted versus the amount of missing data for an implementation of the algorithm on a CPU and GPU.

The number of parameters in the system increases rapidly with increasing dimension. We implemented a Gibbs sampling algorithm in Section 5.4.1, which removes the issue of having to tune parameters in a high dimensional Random Walk algorithm. We showed that the MCMC algorithm can regain the true parameter values as the data interval and sampling frequency increase. There remains an issue relating to the stability of the resulting SDE model. Some parameter vectors, sampled from the posterior, correspond to SDEs with unstable solutions. This is obviously a major issue for prediction. In Chapter 7 we suggest a solution to this stability problem, which leads to some novel MCMC sampling problems. Further research could be done in this direction.

Significant computing power is needed for the MCMC algorithms discussed in this chapter. One alternate approach is to use the structure of the data to implement the algorithm on a Graphics Processing Unit (GPU). In Section 5.5 we implemented one version of the basic inference algorithm and observed significant reductions in computation time. It would be useful to profit from new GPU technology that are now installed in compute clusters, the challenge being that code has to be rewritten and only some algorithms parallelise as effectively as the basic algorithm implemented in Section 5.5

In Chapter 3 we discussed the dependence of reduced climate models on time scale separation between resolved and unresolved variables. One possible method of removing the need for perfect time scale separation is to use models with red noise. This amounts to introducing latent, unobserved variables into the system. In the next chapter we derive an inference method for this type of model.

Chapter 6

Models with Latent Variables

6.1 Models with Latent Variables

In this chapter we consider models where one of the variables is unobserved. A general framework for this problem was given in the thesis of Dargatz [2010] and the papers Golightly and Wilkinson [2008]. Initial experiments on parameter inference for missing components indicate that the data requirements and computation time to get meaningful estimates are enormous. However, we find it feasible when the unobserved process is linear as then this missing data can be sampled directly from the posterior. We describe an inference algorithm for a model of the form

$$\begin{aligned} d\boldsymbol{X}_t &= \boldsymbol{f}(\boldsymbol{X}_t)\,dt + \boldsymbol{Y}_t\,dt \\ d\boldsymbol{Y}_t &= -\boldsymbol{\gamma}\boldsymbol{Y}_t\,dt + \boldsymbol{\sigma}\,d\boldsymbol{B}_t\,, \end{aligned} \tag{6.1}$$

where $\boldsymbol{f}(\boldsymbol{x})$ is a general non-linear function and only $\boldsymbol{X}$ is observed. This is motivated by the need to include memory effects in models with low time scale separation as was discussed in Chapter 3. As well as this, the $\boldsymbol{X}$ component will be a smooth process and will be driven by autocorrelated red noise, which is more realistic for modelling real physical systems. We demonstrate that this model is useful in Chapter 8, where we apply it to the reduced Lorenz model.

We assume that the matrices $\boldsymbol{\gamma}$ and $\boldsymbol{\sigma}$ are diagonal, so there are no interactions between unobserved components and only one unobserved component forces each observed. The usual Euler approximation is insufficient as the resulting covariance matrix is singular (due to the zeroes on the leading diagonal). This type of process is known as a hypoelliptic diffusion and was studied by Pokern et al. [2009], Pavliotis and Stuart [2008]. It consists of both smooth, with zero quadratic variation, and rough components. To write down a likelihood for this process we

use an approximation of higher order than the usual Euler method. Pokern et al. [2009] suggest expanding such that a noise term enters directly into the observed component to give a model with non-singular covariance matrix. This leads to an order $O(\Delta^2)$ in the observed variable and $O(\Delta^{3/2})$ in the unobserved, where we use a fixed interobservation time Δ. However, the inconsistency could lead to a bias in the parameter inference so we recommend expanding to $O(\Delta^2)$ in both components. This is then the Milstein expansion given in Eq. (2.38). Recalling the notation of earlier chapters for X_{ij}, where i indexes the observation time and j the component, we have the following statistical model for Eq. (6.1)

$$\begin{aligned}\begin{pmatrix} X_{i+1,j} \\ Y_{i+1,j} \end{pmatrix} &= \begin{pmatrix} X_{ij} \\ Y_{ij} \end{pmatrix} + \Delta \begin{pmatrix} f(\boldsymbol{X}_i) + Y_{ij} \\ -\gamma_j Y_{ij} \end{pmatrix} \\ &\quad + \sigma_j \begin{pmatrix} \int_{i\Delta}^{(i+1)\Delta} (B_j(s) - B_j(i\Delta))ds \\ -\gamma_j \int_{i\Delta}^{(i+1)\Delta} (B_j(s) - B_j(i\Delta))ds + \int_{i\Delta}^{(i+1)\Delta} dB_j(s) \end{pmatrix}. \end{aligned} \tag{6.2}$$

Using the rules from Chapter 2, we can calculate the covariance matrix with

$$\begin{aligned}&\mathbb{E}\left[\left(\int_{i\Delta}^{(i+1)\Delta} (B_j(s) - B_j(i\Delta))ds\right)^2\right] \\ &= \mathbb{E}\left[\left(\int_0^\Delta ds \int_0^\Delta dB_j(s) - \int_0^\Delta s dB_j(s)\right)^2\right] \\ &= \Delta^2 \mathbb{E}\left[\left(\int_0^\Delta dB_j(s)\right)^2\right] - 2\Delta \mathbb{E}\left[\int_0^\Delta dB_j(s) \int_0^\Delta s dB_j(s)\right] + \mathbb{E}\left[\left(\int_0^\Delta s dB_j(s)\right)^2\right] \\ &= \Delta^2 \int_0^\Delta ds - 2\Delta \int_0^\Delta s ds + \int_0^\Delta s^2 ds \\ &= \frac{\Delta^3}{3}\end{aligned}$$

and similarly

$$\begin{aligned}
&\mathbb{E}\left[\int_{i\Delta}^{(i+1)\Delta}(B_j(s)-B_j(i\Delta))ds\int_{i\Delta}^{(i+1)\Delta}dB_j(s)\right]\\
&=\mathbb{E}\left[\left(\int_0^\Delta ds\int_0^\Delta dB_j(s)-\int_0^\Delta sdB_j(s)\right)\int_0^\Delta dB_j(s)\right]\\
&=\Delta\mathbb{E}\left[\left(\int_0^\Delta dB_j(s)\right)^2\right]-\mathbb{E}\left[\int_0^\Delta sdB_j(s)\int_0^\Delta dB_j(s)\right]\\
&=\Delta\int_0^\Delta ds-\int_0^\Delta sds\\
&=\frac{\Delta^2}{2}\,.
\end{aligned}$$

Using these results we can write down the block diagonal covariance matrix, with block

$$\boldsymbol{\Sigma}=\sigma^2\begin{pmatrix}\Delta^3/3 & \Delta^2/2-\gamma\Delta^3/3\\ \Delta^2/2-\gamma\Delta^3/3 & \Delta+\gamma^2\Delta^3/3-\gamma\Delta^2\end{pmatrix}$$

and inverse

$$\boldsymbol{\Sigma}^{-1}=\frac{1}{\sigma^2}\begin{pmatrix}12/\Delta^3+4\gamma^2/\Delta-12\gamma/\Delta^2 & -6/\Delta^2+4\gamma/\Delta\\ -6/\Delta^2+4\gamma/\Delta & 4/\Delta\end{pmatrix}.$$

The posterior is quadratic in $\boldsymbol{Y}$ and so can be sampled directly. Similar formula to the following were calculated by Pokern [2007]. Consider sampling a single component $Y_{\cdot,j}$ of $\boldsymbol{Y}$. To compute $\mathbb{E}(Y_{ij})$, $\mathrm{Var}(Y_{ij})$ and $\mathrm{Cov}(Y_{ij},Y_{i+1,j})$ the relevant part of the likelihood function is

$$\begin{aligned}
L(Y_{ij}|\boldsymbol{X},\boldsymbol{Y})\propto&-\frac{1}{2}(X_{i+1,j}-X_{ij}-\Delta X_{ij}-\Delta f(\boldsymbol{X}_i))^2(12/\Delta^3+4\gamma_j^2/\Delta-12\gamma_j/\Delta^2)/\sigma_j^2\\
&-\frac{1}{2}(Y_{i+1,j}-Y_{ij}+\Delta\gamma_jY_{ij})^2\frac{4}{\sigma_j^2\Delta}\\
&+(X_{i+1,j}-X_{ij}-\Delta X_{ij}-\Delta f(\boldsymbol{X}_i))(Y_{i+1,j}-Y_{ij}+\Delta\gamma_jY_{ij})(6/\Delta^2-4\gamma_j/\Delta)/\sigma_j^2\\
&-\frac{1}{2}(Y_{ij}-Y_{i-1,j}+\Delta\gamma_jY_{i-1,j})^2\frac{4}{\sigma_j^2\Delta}\\
&+(X_{ij}-X_{i-1,j}-\Delta Y_{i-1,j}-\Delta f(\boldsymbol{X}_{i-1}))(Y_{i,j}-Y_{i-1,j}+\Delta\gamma_jY_{i-1,j})\\
&\times(6/\Delta^2-4\gamma_j/\Delta)/\sigma_j^2\,.
\end{aligned}$$

We update blocks of missing data $Y_{im,j}:Y_{(i+1)m,j}$. For a block away from

the boundaries of the data set the precision matrix $\mathbf{\Lambda}_j \in \mathbb{R}^{m+1\times m+1}\,, j = 1\ldots D$ is

$$\mathbf{\Lambda}_j = \frac{1}{\sigma^2}\begin{pmatrix} \frac{8}{\Delta} & & \cdots & & 0 \\ & \ddots & \frac{2}{\Delta} & & \\ \vdots & \frac{2}{\Delta} & \frac{8}{\Delta} & \frac{2}{\Delta} & \vdots \\ & & \frac{2}{\Delta} & \ddots & \\ 0 & & \cdots & & \frac{8}{\Delta} \end{pmatrix}.$$

If the block is the first in the data set then replace $\Lambda_{j00} = 4/\Delta + \tau_j$, where τ_j is the prior precision of Y_{0j}. If it is the last block then $\Lambda_{jmm} = 4/\Delta$. Inverting $\mathbf{\Lambda}_j$ gives the covariance matrix

$$\mathrm{Cov}(Y_{im,j} : Y_{(i+1)m,j}) = \mathbf{\Lambda}_j^{-1}\,.$$

To compute the mean first define the vector $\boldsymbol{b}_k \in \mathbb{R}^{m+1}\,, k = 1\ldots D$ as

$$\begin{aligned} b_{kj} = &\left(\frac{6}{\Delta^2} - \frac{2\gamma}{\Delta}\right)(X_{im+j+1,k} - X_{im+j,k} - \Delta f(\boldsymbol{X}_{im+j})) \\ &+ \left(\frac{6}{\Delta^2} - \frac{4\gamma}{\Delta}\right)(X_{im+j,k} - X_{im+j-1,k} - \Delta f(\boldsymbol{X}_{im+j-1}))\,, \end{aligned}$$

where $j = 1\cdots m-1$. For $j = 0$

$$\begin{aligned} b_{k0} = &\left(\frac{6}{\Delta^2} - \frac{2\gamma}{\Delta}\right)(X_{im+1,k} - X_{im,k} - \Delta f(\boldsymbol{X}_{im,k})) \\ &+ \left(\frac{6}{\Delta^2} - \frac{4\gamma}{\Delta}\right)(X_{im,k} - X_{im-1,k} - \Delta f(\boldsymbol{X}_{im-1,k})) \\ &+ (4\gamma - 6/\Delta)Y_{im-1,k} + 4/\Delta(Y_{im-1,k} - \Delta\gamma Y_{im-1,k}) \end{aligned}$$

and if $j = m$

$$\begin{aligned} b_{km} = &\left(\frac{6}{\Delta^2} - \frac{2\gamma}{\Delta}\right)(X_{im+m+1,k} - X_{im+m,k} - \Delta f(\boldsymbol{X}_{im+m,k})) \\ &+ \left(\frac{6}{\Delta^2} - \frac{4\gamma}{\Delta}\right)(X_{im+m,k} - X_{im+m-1,k} - \Delta f(\boldsymbol{X}_{im+m-1,k})) \\ &+ (4\gamma - 6/\Delta)Y_{im+m+1,k} - 4/\Delta(\Delta\gamma - 1)Y_{im+m+1,k}\,. \end{aligned}$$

The first block of data where $i = 0$ and $j = 0$ then

$$b_{k0} = \left(\frac{6}{\Delta^2} - \frac{2\gamma}{\Delta}\right)(X_{1k} - X_{0k} - \Delta f(\boldsymbol{X}_0))\,.$$

For the last block $i = N-2$ and $j = m$

$$b_{km} = \left(\frac{6}{\Delta^2} - \frac{4\gamma}{\Delta}\right)\left(X_{(N-2)m+m,k} - X_{(N-2)m+m-1,k} - \Delta f(\boldsymbol{X}_{(N-2)m+m-1,k})\right).$$

Then the mean of the block is

$$\boldsymbol{\mu}_k = \mathbb{E}(Y_{im,k} : Y_{(i+1)m,k}) = \boldsymbol{\Lambda}_k^{-1}\boldsymbol{b}_k\,.$$

The $m+1$ block of missing data are then sampled from $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1})\,, k = 1 \ldots D$ as an extra step in the Algorithms 4.1 and 4.2.

We also impute data between observations for the observed component X. This is not as easy as when noise acts directly on $\boldsymbol{X}$ as now the imputed paths need to be smooth. The acceptance rate of proposed bridges becomes very low. To update $\boldsymbol{X}_{j+1}$ conditioned on $\boldsymbol{X}_j$ and $\boldsymbol{X}_m$ we first consider the covariance

$$\begin{aligned}
\mathrm{Cov}(X_{j+1,k}, X_{mk}) &= \sigma_k^2\mathbb{E}\left[\int_{j\Delta}^{(j+1)\Delta}(B_k(s) - B_k(j\Delta))ds \int_{j\Delta}^{m\Delta}(B_k(s) - B_k(j\Delta))ds\right] \\
&= \sigma_k^2\mathbb{E}\left[\int_0^{\Delta} B_k(s)ds \int_0^{(m-j)\Delta} B_k(s)ds\right] \\
&= \sigma_k^2\mathbb{E}\left[\left(\Delta\int_0^{\Delta} dB_k(s) - \int_0^{\Delta} sdB_k(s)\right)\right. \\
&\left.\times\left((m-j)\Delta\int_0^{(m-j)\Delta} dB_k(s) - \int_0^{(m-j)\Delta} sdB_k(s)\right)\right] \\
&= \sigma_k^2\Delta^3(\frac{m-j}{2} - \frac{1}{6})\,.
\end{aligned}$$

Then the covariance matrix for $X_{j+1,k}$ and X_{mk} conditioned on X_{jk} is

$$\boldsymbol{M}_k = \sigma_k^2\Delta^3\begin{pmatrix} 1/3 & (m-j)/2 - 1/6 \\ (m-j)/2 - 1/6 & (m-j)^3 \end{pmatrix} \; k = 1 \ldots D\,.$$

We make the approximation

$$\begin{aligned}
X_{j+1,k} &= X_{jk} + \Delta f(\boldsymbol{X}_j) + \Delta Y_{jk} + \xi_{1k} \\
X_{mk} &= X_{jk} + (m-j)\Delta f(\boldsymbol{X}_j) + (m-j)\Delta Y_{jk} + \xi_{2k}\,,
\end{aligned}$$

where $\xi_{\cdot,k} \sim \mathcal{N}(0, \boldsymbol{M}_k)$, then condition on the observed value of X_{mk} using the multivariate Normal theory.

We test this algorithm by inferring parameters γ and σ for the following

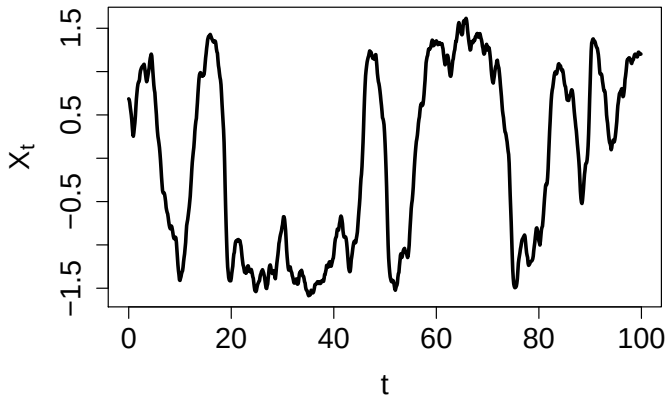


Figure 6.1: Data set from model Eq. (6.3) with $N = 1000$ points with observation interval $\Delta = 0.1$.

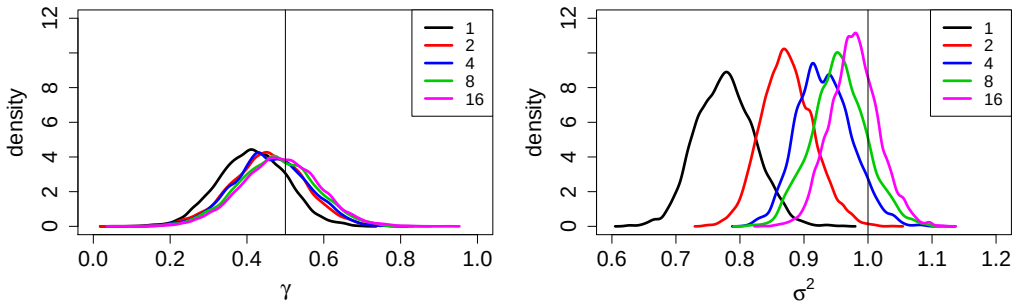


Figure 6.2: Estimated posterior distributions for parameters from the latent process model Eq. (6.3) for various amounts of imputed data m.

model

$$\begin{aligned} dX_t &= (X_t - X_t^3)dt + Y_t dt \\ dY_t &= -\gamma Y_t dt + \sigma dB_t \end{aligned} \tag{6.3}$$

given only observations of X_t. The data set we use is shown in Figure 6.1. The true values are $\gamma = 0.5$ and $\sigma = 1$. We used Gamma priors for both parameters $\gamma, \sigma^2 \sim \Gamma(1, 1)$. We estimated posterior distributions using $20,000$ samples from the algorithm after discarding 5000 as burn in. The results are shown in Figure 6.2 for various amounts of imputed data m. The true values are within the posterior mass and as m increases the posterior mean converges to the true values. The mixing of

the algorithm becomes poor with more imputed data. It is likely that this algorithm would be insufficient in higher dimensions and it may be more practical (though not as rigorous) to use a simulation based approach to inference as discussed in Chapter 4.

Chapter 7

Prediction for Models with Cubic Drift and Linear Diffusion

In this chapter we move from inference to prediction for models of the form in Eq. (5.1). The posterior estimates for parameters, obtained using the inference methods discussed in the previous chapter, are input to the model, which can be used as an approximating stochastic climate model as discussed in Chapter 3. The model can be forward simulated using different random number seeds to obtain a probabilistic prediction of the evolution of the system. The statistical properties of the stochastic model can be compared to the full model to validate that it is a useful approximation. This is done in Chapter 8 for the models introduced in Chapter 3.

In the current chapter we address some practical issues that arise when building predictive models that include parameter estimates. As discussed in Section 5.4.1, the inferred model may not be stable and thus solutions may explode to infinity in finite time. The posterior distribution of parameters may include regions of stability but in general this is not known and there can be significant overlap between stable and unstable regions of marginal distributions as demonstrated in Figure 5.16. If a subset of predictions explode to infinity they need to be removed before any estimates are calculated. This is equivalent to restricting the parameter space by using a prior and is an advantage of the Bayesian approach. However, practical experience indicates that for higher dimensional models the unstable region of the parameter space becomes larger so this rejection method becomes inefficient and unpractical. We propose an alternative method of including prior information to restrict the parameter space. The idea is to ensure that the energy associated with the cubic terms in the model is non-increasing. This places constraints on the parameters entering the cubic terms. This approach was applied to a simple cubic

model by Majda et al. [2009]. Here we develop this for general cubic models. The condition of non increasing energy can be translated to a restriction on the cubic parameters by ensuring a certain matrix is negative definite. We refer to this as the stability matrix. It is likely that there are other methods that restrict the space less severely but at least this approach is guaranteed to give stable solutions. In Section 7.1.1, through an example, we study how the parameter space is restricted by the stability matrix in comparison to the rejection method mentioned above.

The derivation of the stability matrix in the general case is given in Section 7.1.2. The constraint has consequence for the inference method used such that the Gibbs sampler of Section 5.4.1 can now only be applied directly to sample the non-cubic parameters after conditioning on the remainder. There are several options to sample the cubic parameters in the stability matrix which we introduce in Section 7.2. This leads to an interesting MCMC problem of sampling negative/positive definite matrices, details of which we give in Section 7.2. We propose novel solutions to this problem which use the Wishart and non-central Wishart distributions as proposals for the transition density. In Section 7.2 we validate these algorithms and assess their efficiency in comparison to component-wise sampling of the stability matrix. Firstly we give an intuitive derivation of a stability matrix for a simple two dimensional model.

7.1 Derivation of the Stability Matrix

7.1.1 Simple Models

Here we give an explicit derivation of the stability matrix for a one then two dimensional model. The approach is motivated by the work of Majda et al. [2009] but can be cast in terms of the theory of Lyapunov stability discussed in Section 2.6. Recalling Theorem 1, we seek a proper, twice differentiable function V and numbers $K > 0, c > 0$ and $\epsilon \geq 0$ such that for $|x| > K$ we have $LV \leq cV + \epsilon$, where L is the infinitesimal generator. We find that it is easiest to use the squared Euclidean norm for the Lyapunov function $V(\boldsymbol{x}) = |\boldsymbol{x}|^2$. Consider first the 1 dimensional example

$$dX_t = (a_1 + a_2 X_t + a_3 X_t^2 + a_4 X_t^3)dt + (\sigma_1 + \sigma_2 X_t)dB_t\,,$$

with $V(x) = x^2$. Stability requires

$$LV(x) = 2x(a_1 + a_2 x + a_3 x^2 + a_4 x^3) + (\sigma_1 + \sigma_2 x)^2 \leq cx^2 + \epsilon\,. \tag{7.1}$$

Outside of a ball of radius K the quartic term will dominate. The above condition will be satisfied for any ϵ if $a_4 < 0$. In multiple dimensions the restrictions on the parameter space become more complicated. Consider the two dimensional system

$$\begin{aligned} dX_t &= (a_1 X_t^3 + a_2 X_t Y_t^2 + a_3 X_t^2 Y_t)dt + \sigma_1 dB_t^1 \\ dY_t &= (a_4 Y_t^3 + a_5 X_t Y_t^2)dt + \sigma_2 dB_t^2 \,. \end{aligned}$$

Arguing as before the system will be stable if

$$x(a_1 x^3 + a_2 x y^2 + a_3 x^2 y) + y(a_4 y^3 + a_5 x y^2) \leq 0 \,. \tag{7.2}$$

This can be written in matrix form

$$\begin{pmatrix} x^2 & xy & y^2 \end{pmatrix} \begin{pmatrix} a_1 & a_3/2 & 0 \\ a_3/2 & a_2 & a_5/2 \\ 0 & a_5/2 & a_4 \end{pmatrix} \begin{pmatrix} x^2 \\ xy \\ y^2 \end{pmatrix} \leq 0 \,. \tag{7.3}$$

This will hold if the matrix is negative definite. We can ensure that a matrix is negative definite using the following property.

Theorem 2. *A $n \times n$ matrix M is negative definite if and only if all $k \leq n$ leading principal minors obey $|M^{(k)}|(-1)^k > 0$. The kth principal minor is the determinant of the upper left $k \times k$ sub-matrix.*

In the case of Eq. (7.3) this gives the conditions

$$a_1(a_2 a_4 - \frac{a_5^2}{4}) - \frac{a_3^2}{4} a_4 < 0, \;\; a_1 a_2 - \frac{a_3^2}{4} > 0, \;\; a_1 < 0 \,.$$

We know that the parameters along the diagonal must be negative. This leads to the following bounds on a_3 and a_5

$$-\sqrt{4a_1 a_2 - \frac{a_1 a_5^2}{a_4}} < a_3 < \sqrt{4a_1 a_2 - \frac{a_1 a_5^2}{a_4}} \,.$$

$$-\sqrt{4a_2 a_4 - \frac{a_4 a_3^2}{a_1}} < a_5 < \sqrt{4a_2 a_4 - \frac{a_4 a_3^2}{a_1}} \,.$$

Or we can write this as the stability boundary

$$\frac{a_1}{4} a_5^2 + \frac{a_4}{4} a_3^2 = a_1 a_2 a_4 \,.$$

Values of a_3 and a_5 inside this ellipse give an SDE with stable solutions. Outside

the solutions may be stable but this can not be guaranteed.

7.1.2 General Case

We generalise the above to the model in Eq. (5.1) and look for a matrix $\boldsymbol{M}$ to use as a stability criteria. The Lyapunov function constraint implies that the energy associated with the cubic terms should induce damping which means that the associated energy should have negative time derivative. This energy equation is

$$\frac{1}{2}\frac{dE}{dt} = \sum_{i=1}^{D} X_i \frac{dX_i}{dt} = \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{j}\sum_{l=1}^{k} A_{i,f(j,k,l)} X_i X_j X_k X_l \,, \tag{7.4}$$

where $f(j,k,l)$ is given by Eq.(5.31). From this we determine a quadratic form that is negative definite, similar to Majda et al. [2009]. Consider the vector $\boldsymbol{v}$ with $(D+1)D/2$ components of the form $V_{(i-1)i/2+j} = X_i X_j$ with $1 \leq j \leq i \leq D$. For example, for a two dimensional system $\boldsymbol{v} = (X_1X_1, X_1X_2, X_2X_2)$. Now if we can find a negative definite matrix $\boldsymbol{M}$ such that

$$\boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v} = \frac{1}{2}\frac{dE}{dt} \tag{7.5}$$

then the time derivative will be negative.

One possible solution is as follows. Let matrix $\boldsymbol{M} \in \mathbb{R}^{(D+1)D/2 \times (D+1)D/2}$ then assign its components as

$$M_{(i-1)i/2+j,(k-1)k/2+l} = \begin{cases} A_{k,f(i,j,l)}, \text{ if } k > j \text{ and } l \leq j \\ 0, \text{ if } k > j \text{ and } l > j \\ A_{k,f(i,j,l)} + A_{l,f(i,j,k)}, \text{ if } k \leq j \text{ and } l < k \\ A_{k,f(i,j,l)}, \text{ if } k \leq j \text{ and } l = k \end{cases} , \tag{7.6}$$

where $1 \leq j \leq i \leq D$ and $1 \leq l \leq k \leq D$. For example, for $D = 2$ we have

$$\begin{aligned} \boldsymbol{M} &= \begin{pmatrix} A_{1,f(1,1,1)} & A_{2,f(1,1,1)} & 0 \\ A_{1,f(2,1,1)} & A_{2,f(2,1,1)} & 0 \\ A_{1,f(2,2,1)} & A_{2,f(2,2,1)} + A_{1,f(2,2,2)} & A_{2,f(2,2,2)} \end{pmatrix} \\ &= \begin{pmatrix} A_{1,7} & A_{2,7} & 0 \\ A_{1,8} & A_{2,8} & 0 \\ A_{1,9} & A_{2,9} + A_{1,10} & A_{2,10} \end{pmatrix} \end{aligned} \tag{7.7}$$

and for $D = 3$

$$\boldsymbol{M} = \begin{pmatrix} A_{1,11} & A_{2,11} & 0 & A_{3,11} & 0 & 0 \\ A_{1,12} & A_{2,12} & 0 & A_{3,12} & 0 & 0 \\ A_{1,13} & A_{2,13} + A_{1,14} & A_{2,14} & A_{3,13} & A_{3,14} & 0 \\ A_{1,15} & A_{2,15} & 0 & A_{3,15} & 0 & 0 \\ A_{1,16} & A_{2,16} + A_{1,17} & A_{2,17} & A_{3,16} & A_{3,17} & 0 \\ A_{1,18} & A_{2,18} + A_{1,19} & A_{2,19} & A_{3,18} + A_{1,20} & A_{3,19} + A_{2,20} & A_{3,20} \end{pmatrix} .$$

Then when updating the drift parameters one ensures that $\boldsymbol{M}$ is negative definite. It suffices to check that the symmetric part $(\boldsymbol{M} + \boldsymbol{M}^T)/2$ is negative definite.

We now give a careful derivation of $\boldsymbol{M}$. We write the quadratic form component-wise and equate it to the energy equation Eq. (7.4):

$$\begin{aligned} \sum_{i=1}^{D}\sum_{j=1}^{i}\sum_{k=1}^{D}\sum_{l=1}^{k} M_{(i-1)i/2+j,(k-1)k/2+l} X_i X_j X_k X_l &= \sum_{i=1}^{D}\sum_{j=1}^{D}\sum_{k=1}^{j}\sum_{l=1}^{k} A_{i,f(j,k,l)} X_i X_j X_k X_l \\ &= \sum_{j=1}^{D}\sum_{i=1}^{D}\sum_{k=1}^{i}\sum_{l=1}^{k} A_{j,f(i,k,l)} X_i X_j X_k X_l \\ &= \sum_{k=1}^{D}\sum_{i=1}^{D}\sum_{j=1}^{i}\sum_{l=1}^{j} A_{k,f(i,j,l)} X_i X_j X_k X_l \,, \end{aligned}$$

where in the second line we have just renamed $i \leftrightarrow j$ and in the third $j \leftrightarrow k$. This implies that we can write

$$\begin{aligned} \sum_{k=1}^{D}\sum_{l=1}^{k} M_{(i-1)i/2+j,(k-1)k/2+l} X_k X_l &= \sum_{k=1}^{D}\sum_{l=1}^{j} A_{k,f(i,j,l)} X_k X_l \\ &= \sum_{k=j+1}^{D}\sum_{l=1}^{j} A_{k,f(i,j,l)} X_k X_l + \sum_{k=1}^{j}\sum_{l=1}^{k} A_{k,f(i,j,l)} X_k X_l \\ &\quad + \sum_{k=1}^{j}\sum_{l=k+1}^{j} A_{k,f(i,j,l)} X_k X_l \,. \end{aligned}$$

Elements of the first two terms can be easily assigned to components of $\boldsymbol{M}$, however

we must first rearrange the last term as

$$\sum_{k=1}^{j}\sum_{l=k+1}^{j} A_{k,f(i,j,l)}X_kX_l = \sum_{l=1}^{j}\sum_{k=1}^{l-1} A_{k,f(i,j,l)}X_kX_l$$
$$= \sum_{k=1}^{j}\sum_{l=1}^{k-1} A_{l,f(i,j,k)}X_kX_l$$

The result in Eq. (7.6) now follows.

7.2 Sampling the Stability Matrix

7.2.1 Basic Algorithms

The parameters in all terms but the cubic can be sampled using the Gibbs sampler. However, the cubic parameters must obey the negative definite requirement of matrix $\boldsymbol{M}$. They will therefore have a normal distribution subject to the negative definite constraint. This results in a Truncated Normal distribution with a complicated truncation boundary. In this section we compare five methods for sampling the density of a $n \times n$ matrix $\boldsymbol{M}$, with normally distributed components, subject to the constraint that it is negative (or equivalently positive) definite. The algorithms are summarised in Table 7.1. The first two are simple and described here. The other three are in the following subsections.

The first algorithm simply proposes a sample from the conditional posterior, just the normal distribution, constructs the matrix $\boldsymbol{M}$ as in Eq. (7.6) then accepts or rejects according to whether the negative definite property holds. This can be checked by simply computing a Cholesky decomposition of the negated matrix. We will refer to this approach as the **Rejection** algorithm.

For the second algorithm we use a random walk with Gaussian innovations to update the cubic parameters, where the proposal covariance will be chosen such that it is proportional to the posterior covariance computed using a preliminary run. Again $\boldsymbol{M}$ is constructed as in Eq. (7.6). At each step $\boldsymbol{M}$ needs to be checked for negative definiteness. We refer to this as the **Random Walk** algorithm.

7.2.2 Component-wise Sampling

The third algorithm updates $\boldsymbol{M}$ component-wise and is based on the property of the principal minors given in Theorem 2. As shown in Eq. (7.6) parameters can enter $\boldsymbol{M}$ as a linear combination. For example, in a two dimensional model as in

Algorithm	Section	Summary
Rejection	7.2.1	Sample from the Normal distribution and reject if not negative definite
Random Walk	7.2.1	Use Normal innovations to propose a value and reject if not negative definite
Component-wise	7.2.2	Calculate the upper and lower bounds and sample the Truncated Normal
Central Wishart	7.2.3	Propose values from the Wishart distribution
Non-Central Wishart	7.2.4	Multiply current matrix by a Matrix Normal and use non-Central Wishart

Table 7.1: Summary of Monte Carlo algorithms, discussed in this section, to sample negative/positive definite matrices with Normally distributed components.

Eq. (7.7), $M_{3,2} = (A_{2,9} + A_{1,10})/2$. Therefore after sampling $M_{3,2}$ one parameter, $A_{2,9}$ say, is sampled from its conditional posterior of the Gibbs sampler and the other calculated as $A_{1,10} = 2A_{3,2} - A_{2,9}$.

We describe the component-wise sampling of a negative definite matrix, separating the discussion into describing the algorithms for diagonal and off-diagonal parameters. Consider the parameters on the diagonal. As they only enter $\boldsymbol{M}$ once each will have an associated upper bound. The Algorithm 7.1 calculates the upper bound associated with the constraint from each principal minor. It does this to find the least upper bound and thereby the truncation point of the normal distribution.

Algorithm 7.1 Sample parameters along diagonal of Stability Matrix

for $i = 1$ to n **do**
  U_i=0
  for $j = i$ to n **do**
    $x = -\left(\sum^{j}_{\substack{k \neq i \\ k=1}} (-1)^{i+k} M_{ik} |M^{(j)}_{\{-i\},\{-k\}}|\right) / |M^{(j)}_{\{-i\},\{-i\}}|$
  end for
  if $x < U_i$ **then**
    $U_i = x$
  end if
  $M_{ii} \sim \mathcal{N}_{-}(\mu_i, U_i, \sigma_i^2)$
end for

Here, $\mathcal{N}_{-}(\mu, u, \sigma^2)$ is the right truncated normal distribution with mean μ, standard deviation σ and upper bound u.

The off-diagonal parameters enter twice so there will be a quadratic function determining their upper and lower bounds for each leading principal minor. For parameter in element $M^{(k)}_{ij}$ there will be an associated quadratic $a^{(k)}_{ij} M^2_{ij} + b^{(k)}_{ij} M_{ij} +$

$c_{ij}^{(k)} = 0$ where the coefficients are functions of the other parameters. These coefficients are found to be

$$
\begin{aligned}
a_{ij}^{(k)} &= -|M^{(k)}_{/\{i,j\},/\{i,j\}}| \\
b_{ij}^{(k)} &= (-1)^{i+j} \sum_{\substack{k \neq i \\ k=1}}^{j-1} M_{jk}(-1)^{j-1+k}|M^{(k)}_{/\{i,j\},/\{j,k\}}| \\
&\quad + (-1)^{i+j} \sum_{\substack{k \neq i \\ k=j+1}}^{N} M_{jk}(-1)^{j+k}|M^{(k)}_{/\{i,j\},/\{j,k\}}| \\
&\quad + (-1)^{i+j} \sum_{\substack{k \neq j \\ k=1}}^{i-1} M_{ik}(-1)^{i-1+k}|M^{(k)}_{/\{i,j\},/\{i,k\}}| \\
&\quad + (-1)^{i+j} \sum_{\substack{k \neq j \\ k=i+1}}^{N} M_{ik}(-1)^{i+k}|M^{(k)}_{/\{i,j\},/\{i,k\}}| \\
c_{ij}^{(k)} &= \sum_{\substack{k \neq j \\ k=1}}^{N} M_{ik}(-1)^{i+k} \left(\sum_{\substack{l \neq i \\ l=1}}^{k-1} M_{jl}(-1)^{j-1+l}|M^{(k)}_{/\{i,j\},/\{l,k\}}| \right. \\
&\quad \left. + \sum_{\substack{l \neq i \\ l=k+1}}^{N} M_{jl}(-1)^{j+l}|M^{(k)}_{/\{i,j\},/\{l,k\}}| \right),
\end{aligned} \tag{7.8}
$$

where $|M^{(k)}_{/\{i,j\},/\{l,k\}}|$ represents the kth principal minor with rows i and j and columns l and k removed. For each component M_{ij} this quadratic can be solved to give upper and lower bounds on the parameter. The matrix M can be cycled through updating each parameter in turn. Algorithm 7.2 describes the sampling of off-diagonal elements using the coefficients in Eq. (7.8). Here, the notation, $\mathcal{N}^{+}_{-}(\mu, u^{-}, u^{+}, \sigma^2)$ refers to the doubly truncated normal distribution with mean μ, left truncation u^{-}, right truncation u^{+} and standard deviation σ.

To simulate from truncated normal distributions one has the option of just proposing from the full distribution and then rejecting the sample if it falls outside the permitted region. However, this can be extremely inefficient if the truncated region is in the tail of the distribution. Most proposals will then be rejected. Alternatively one can consider using the inverse Cumulative Density Function (CDF) method. One simply calculates the corresponding CDF of the lower and upper

Algorithm 7.2 Sample parameters off diagonal

for $i = 1$ to n **do**
  for $j = i + 1$ to n **do**
    $u^+ = \infty$
    $u^- = -\infty$
    for $k = j$ to n **do**
      Calculate $a_{ij}^{(k)}$, $b_{ij}^{(k)}$ and $c_{ij}^{(k)}$ and solve $a_{ij}^{(k)} x^2 + b_{ij}^{(k)} x + c_{ij}^{(k)} = 0$.
      Set mn $= \min(x_1, x_2)$ and mx $= \max(x_1, x_2)$
    end for
    if $mx < u^+$ **then**
      $u^+ = mx$
    end if
    if $mn > u^-$ **then**
      $u^- = mn$
    end if
    $M_{ij} \sim \mathcal{N}_-^+(\mu_{ij}, u^-, u^+, \sigma_{ij}^2)$
  end for
end for

boundaries and then draws a uniform random variable between these numbers. Inverting the CDF gives a random variable from the Normal distribution restricted to this region. This also becomes computationally inefficient in tail regions. The problem is due to the large number of terms needed in the numerical approximation of the inverse CDF of the Normal distribution. In these low probability regions the numerical error can be large compared to the estimated value.

For our problem we use the rejection sampler methods proposed by Robert [1995].

Definition 2. *Rejection sampling from a distribution $h(x)$ is based on a proposal distribution $g(x)$ such that $h(x) \leq Cg(x)$ holds for some constant C and all of the support of $h(x)$.*

For a one sided truncated Normal the exponential distribution is a good proposal. First it is translated to coincide with the truncation point then the rate parameter is optimised in order to closely match the tail of the Normal distribution. The proposal is

$$g(z; \alpha, \mu^-) = \alpha \exp(-\alpha(z - \mu^-)) \mathbb{I}_{z \geq \mu^-} . \tag{7.9}$$

The optimal value of α is calculated by maximising the expected acceptance probability and is shown to be

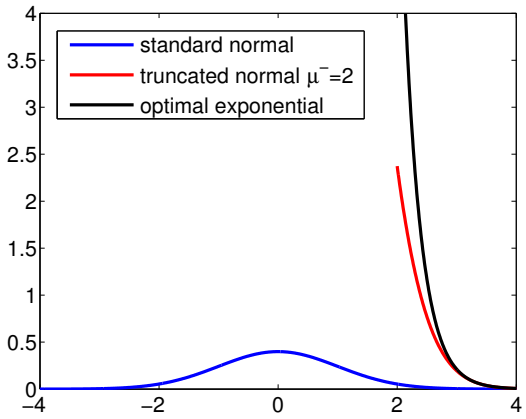


Figure 7.1: Left truncated normal distribution with $\mu^- = 2$ compared with scaled optimal exponential proposal of Eq. (7.9) used for rejection sampling.

$$\alpha^*(\mu^-) = \frac{\mu^- + \sqrt{(\mu^-)^2 + 4}}{2}$$

More details are given in Robert [1995].

Figure 7.1 shows a standard Normal distribution with left truncation $\mu^- = 2$ with the truncated distribution and the optimal exponential approximation. We did a numerical study to compare the standard Normal and Exponential proposals. The efficiency of proposing x from the standard normal then accepting if $x > \mu^-$ falls to approximately 0.023 while for the optimised exponential proposal is approximately 0.5.

For the doubly truncated Normal one uses either an exponential or uniform distribution, as a proposal, depending upon the size of the truncated region. If the following holds

$$u^+ > u^- + \frac{2\sqrt{e}}{u^- + \sqrt{(u^-)^2 + 4}} \exp\left(\frac{(u^-)^2 - u^-\sqrt{(u^-)^2 + 4}}{4}\right)$$

then it can be shown that the exponential is more efficient, otherwise the uniform is better [Robert, 1995]. Figure 7.2 shows both the uniform and exponential approximations for both cases.

We use Algorithms 7.1 and 7.2, along with the methods of sampling truncated Normal variables, to sample the stability matrix $\boldsymbol{M}$. We call this the **Component-wise Algorithm**. To compare the efficiency of this algorithm to the others we use two model problems, details of which are given in Table 7.2.

Figure 7.3 shows the autocorrelation functions estimated from the output of

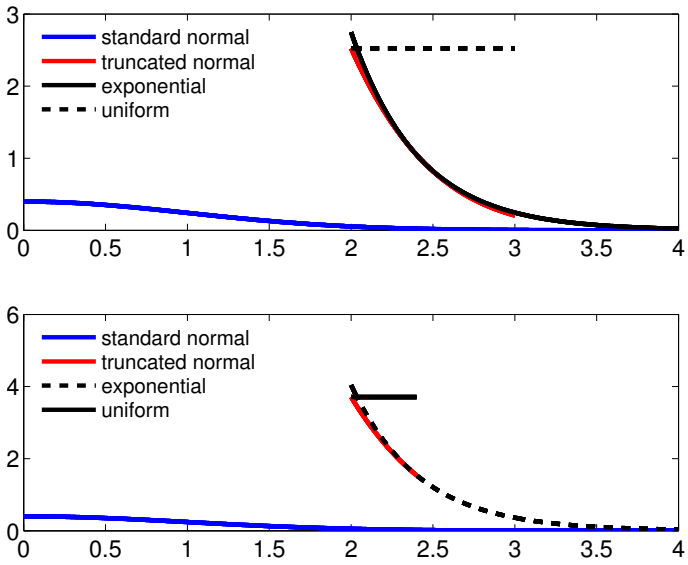


Figure 7.2: Doubly truncated normal distribution. The top figure has $u^- = 2$ and $u^+ = 3$ and is better approximated with the exponential distribution. The bottom figure has $u^- = 2$ and $u^+ = 2.5$ and the uniform is more efficient.

the **Component-wise** algorithm to sample $\boldsymbol{W}$ from Model Problem 1 with $d = 3$. There are 6 independent parameters. The ACFs indicate that the Markov Chain is mixing reasonably well.

7.2.3 Central Wishart Algorithm

Another approach is to consider algorithms that sample the whole matrix $\boldsymbol{M}$ at once. A convenient algorithm can be developed based on the *Wishart distribution*. This is a probability distribution over the space of positive definite matrices and is a matrix generalisation of the chi-squared distribution.

Model Problem	**Mean $\boldsymbol{\mu}$**	**Covariance $\boldsymbol{\Gamma}$**
1	$\mu_i = 0, i = 1, \ldots, d$	$\boldsymbol{\Gamma} = \boldsymbol{I}_d$
2	$\mu_i = 5, i = 1, \ldots, d$	$\Gamma_{ij} = 5(1 - (j - i)/d)$

Table 7.2: Model Problems for efficiency tests. Both are normal densities Truncated Normal densities. The Normal distribution from which they are derived has mean $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\boldsymbol{\Gamma} \in \mathbb{R}^{d \times d}$. The components of these densities are entered into the upper triangle of a matrix $\boldsymbol{W}$ in row major order. Then it is required that $\boldsymbol{W} \in \mathbb{R}^{p \times p}$ is negative definite. Here we set $d = p(p+1)/2$ to be the number of independent components.

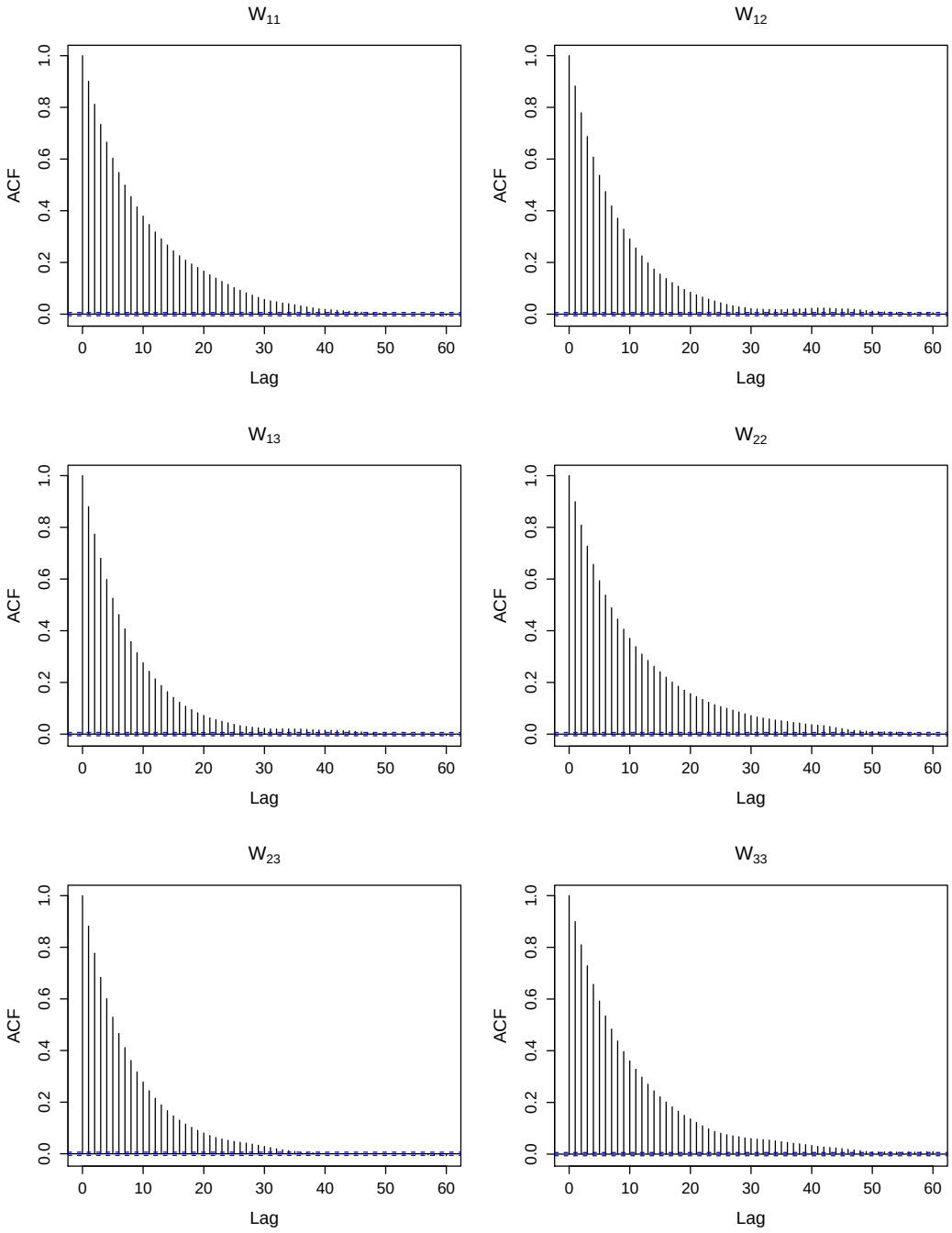


Figure 7.3: Autocorrelation functions estimated from the output of the **Component-wise** algorithm applied to Model Problem 1. They were estimated using 10^5 MCMC samples after discarding an initial burn in of 10^4.

Definition 3. *If* $\boldsymbol{X}_i \sim \mathcal{N}_p(0, \Sigma)\, i = 1, \ldots, n$ *are independent normally distributed* $p-$*vectors, then the matrix*

$$\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T$$

is Wishart distributed *with positive definite scale matrix* $\boldsymbol{\Sigma}$*, dimension* p *and degrees of freedom* n*. We write* $\boldsymbol{S} \sim \mathcal{W}(\boldsymbol{\Sigma}, p, n)$*.*

The density of $\boldsymbol{S}$ is

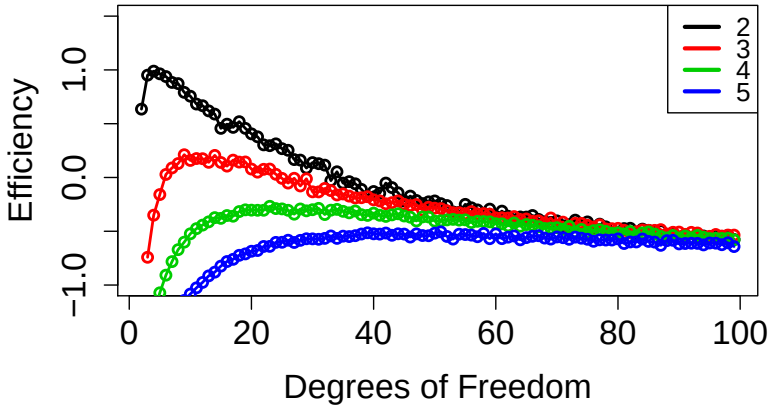
$$p(\boldsymbol{S}; \boldsymbol{\Sigma}) = \frac{(\det \boldsymbol{S})^{(n-p-1)/2}}{2^{pn/2}\Gamma_p(n/2)(\det \boldsymbol{\Sigma})^{n/2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right)\right). \tag{7.10}$$

We also use the construction via the matrix normal distribution on $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. If $\boldsymbol{X} \sim \mathcal{N}_{p,n}(0, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$ is matrix normal distributed then $\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X} \sim \mathcal{W}(\boldsymbol{\Sigma}, p, n)$.

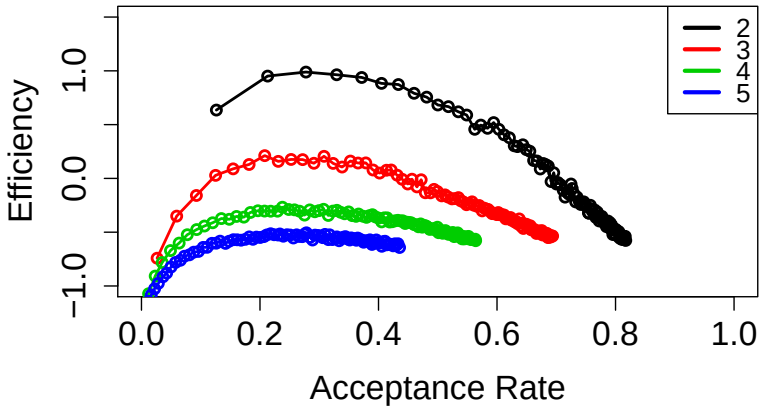
The algorithm we consider generates a proposal from $\boldsymbol{M}^* \sim \mathcal{W}(\boldsymbol{M}, p, n)/n$. In this case $\mathbb{E}(\boldsymbol{M}^*) = \boldsymbol{M}$ and $\mathrm{Cov}(\boldsymbol{M}^*) = 2/n\boldsymbol{M} \otimes \boldsymbol{M}$. Therefore, the expectation has the desired property of equalling the current state. There is one free parameter n in this algorithm that we can use to control the magnitude of the covariance. However, we have no freedom in the structure of the covariance. We will refer to this algorithm as the **Central Wishart** algorithm.

We investigated the optimal value of n in the Central Wishart Algorithm by applying it to two artificial problems. We performed the experiment for $p = 2, 3, 4, 5$. For each case we discarded a burn in of 10^3 samples and computed the autocorrelation function ρ using 10^6 samples. As discussed in Section 5.1 we can quantify the efficiency of an algorithm using the autocorrelation function. We estimate the efficiency as in Definition 1.

Figure 7.4a shows the log efficiency plotted against varying degrees of freedom n in the proposal distribution applied to the second model problem. The results indicate that for low dimensional problems there is a maximum efficiency at $n = 6$ for $p = 2$ and $n = 10$ for $p = 3$. For higher dimensions the log efficiency appears to assymptote at 0.5 independently of the dimension of the problem. Figure 7.4b is a plot of the log efficiency versus acceptance rate. For all dimensions the optimal acceptance rate is approximately 0.25. This is in qualitative agreement with the theoretical predictions of Roberts et al. [1997]. In that paper the authors prove that for a target distribution with independent components, as the dimension of the problem goes to infinity, the optimal acceptance rate for random walk style algorithms is 0.234. They show that the scale of random walk proposals should increase as $O(1/d)$. They argue that the asymptotic result can apply even in low dimensions: this agrees with our result in Figure 7.4b. However, it is an open



(a) Efficiency as a function of degrees of freedom.



(b) Efficiency as a function of acceptance rate.

Figure 7.4: Efficiency of the Wishart proposal distribution sampling the standard normal distribution restricted to positive definite matrices (see text). The dimension of the matrix M ranges from 2 to 5.

problem to derive optimal asymptotic acceptance rates when the target is not a product distribution. Our results here indicate that the prediction of 0.234 still holds approximately even when there is correlation between components. The results (not shown) for the first model problem are similar.

7.2.4 Non-Central Wishart Algorithm

As a fifth and final algorithm to sample negative/positive definite matrices we consider the **non-central Wishart** distribution.

Definition 4. *Let* $\boldsymbol{X} \sim \mathcal{N}_{n,k}(\boldsymbol{\Pi}, \boldsymbol{I}_k \otimes \boldsymbol{\Sigma}), k \geq n$ *be matrix normal distributed. A random* $n \times n$ *positive definite matrix* $\boldsymbol{S} = \boldsymbol{X}^T\boldsymbol{X}$ *has a* non-central Wishart distribution *with parameters* $\boldsymbol{\Sigma}$*,* n*,* k *and* $\boldsymbol{\Delta} = \boldsymbol{\Pi}^T\boldsymbol{\Pi}$*. In this case we write* $\boldsymbol{S} \sim \mathcal{W}(\boldsymbol{\Sigma}, n, k; \boldsymbol{\Delta})$ *and call* k *the degrees of freedom and* $\boldsymbol{\Delta}$ *the non centrality matrix.*

The density of $\boldsymbol{S}$ is given by

$$p(\boldsymbol{S}; k, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}) = \frac{(\det \boldsymbol{S})^{(k-n-1)/2}}{2^{nk/2}\Gamma_n(k/2)(\det \boldsymbol{\Sigma})^{k/2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{S} + \boldsymbol{\Omega}\right)\right) {}_0F_1\left(\frac{k}{2}; \frac{1}{4}\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}\right), \tag{7.11}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Pi}^T\boldsymbol{\Pi}$ and ${}_0F_1(\cdot)$ is a hypergeometric function with a matrix argument (see Muirhead [1982]).

One possible way of using the non-central Wishart distribution to sample the space of $\boldsymbol{M}$ is by drawing

$$\boldsymbol{A} \sim \mathcal{N}_{n,n}(\boldsymbol{\Pi}, \boldsymbol{\Phi} \otimes \boldsymbol{\Sigma}) \tag{7.12}$$

from the $n \times n$ matrix normal distribution with mean $\boldsymbol{\Pi} \in \mathbb{R}^{n\times n}$, row covariance $\boldsymbol{\Phi} \in \mathbb{R}^{n\times n}$ and column covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$. The proposal is then constructed as $\boldsymbol{M}^* = \boldsymbol{A}^T\boldsymbol{M}\boldsymbol{A}$. If $\boldsymbol{M}$ is a positive definite matrix then $\boldsymbol{M}^*$ will also be positive definite [Eaton, 2007]. It happens that if we choose the matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ appropriately then we can calculate the forward $p(\boldsymbol{M}^*|\boldsymbol{M})$ and backward $p(\boldsymbol{M}|\boldsymbol{M}^*)$ transition densities.

Theorem 3 (Eaton [2007])**.** *Consider* $\boldsymbol{A} \sim \mathcal{N}_{n,n}(\boldsymbol{\Pi}, \boldsymbol{\Phi} \otimes \boldsymbol{\Sigma})$ *and let* $\boldsymbol{M}^* = \boldsymbol{A}^T\boldsymbol{M}\boldsymbol{A}$*, where* $\boldsymbol{M} \geq 0$ *is* $n \times n$*. If* $\boldsymbol{\Phi} = \boldsymbol{M}^{-1}$ *then*

$$\boldsymbol{M}^* \sim \mathcal{W}(\boldsymbol{\Sigma}, n, n; \boldsymbol{\Pi}^T\boldsymbol{M}\boldsymbol{\Pi}).$$

Proof 4. *Write* $\boldsymbol{M} = \boldsymbol{C}^2$ *with* $\boldsymbol{C} \geq 0$*. Let* $\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{A}$ *then* $\boldsymbol{Y} \sim \mathcal{N}_{n,n}(\boldsymbol{C}\boldsymbol{\Pi}, (\boldsymbol{C}\boldsymbol{\Phi}\boldsymbol{C}) \otimes \boldsymbol{\Sigma})$*, which implies*

$$\boldsymbol{Y} \sim \mathcal{N}_{n,n}(\boldsymbol{C}\boldsymbol{\Pi}, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$$

Clearly $\boldsymbol{M}^* = \boldsymbol{Y}^T\boldsymbol{Y}$ *and so*

$$\boldsymbol{M}^* \sim \mathcal{W}(\boldsymbol{\Sigma}, n, n; \boldsymbol{\Pi}^T\boldsymbol{M}\boldsymbol{\Pi})$$

from the definition of the non-Central Wishart.

The **non-Central Wishart** algorithm samples a matrix Normal $\boldsymbol{A}$ from Eq. (7.12) using $\boldsymbol{\Phi} = \boldsymbol{M}^{-1}$ then constructs the proposal as $\boldsymbol{M}^* = \boldsymbol{A}^T\boldsymbol{M}\boldsymbol{A}$. We need only then choose appropriate values for $\boldsymbol{\Pi}$ and $\boldsymbol{\Sigma}$. From the properties of the non-central Wishart (see Muirhead [1982]) we know that the expectation of $\boldsymbol{M}^*$ is

$$\mathbb{E}[\boldsymbol{M}^*] = n\boldsymbol{\Sigma} + \boldsymbol{\Pi}^T\boldsymbol{M}\boldsymbol{\Pi}\,.$$

This is equal to $\boldsymbol{M}$ if we choose $\boldsymbol{\Pi} = \sqrt{\alpha}\boldsymbol{I}_n$ and $\boldsymbol{\Sigma} = (1-\alpha)/n\boldsymbol{M}$. Then we only have one free parameter α to set.

We tested the algorithm on the model problems in Table 7.2. Output from initial tests of the algorithm (with no optimisation using $\alpha = 0.5$), applied to model problem 2 for $n = 3$, is shown in Figure 7.5. The histograms estimated from the MCMC output are close to those from samples drawn directly from the distribution (using the **Rejection** algorithm). This demonstrates that the Markov Chain converges to the target distribution. Figure 7.6 shows the autocorrelation functions for the MCMC output. It shows that the algorithm is well mixing, even with no tuning.

7.2.5 Efficiency of the Algorithms

We performed a direct comparison of all five algorithms discussed in this section by running them for 10^6 iterations on the two model problems in Table 7.2 for dimensions 2-5. For each, we record the time t in seconds for the simulation to complete and estimate the number of independent samples per second. The **Random Walk**, **Central Wishart** and **non-Central Wishart** algorithms are Markov Chains and so successive samples are correlated. Therefore, we first estimate the efficiency η as in Definition 1 and report the number of independent samples per second as $\nu = (\eta/100)10^6/t$. For the **Rejection** algorithm we run the simulation until 10^6 samples have been accepted and report $\nu = 10^6/t$ and for the **Component-wise** algorithms $\nu = 10^6/t$.

Table 7.3 shows the results from the model problem 1, where there was no correlation between components and Table 7.4 shows the results for model problem 2 where there is significant correlation. The results demonstrate that using the **Rejection** algorithm becomes extremely inefficient as the dimension increases. Although it is a simple algorithm it takes a very long time to draw 10^6 samples as so few proposals are accepted. It is worse for model problem 2 where the density is correlated (Table 7.4). The **Random Walk** algorithm is simple and very quick to run though the efficiency decreases rapidly with dimension and so η reduces to the

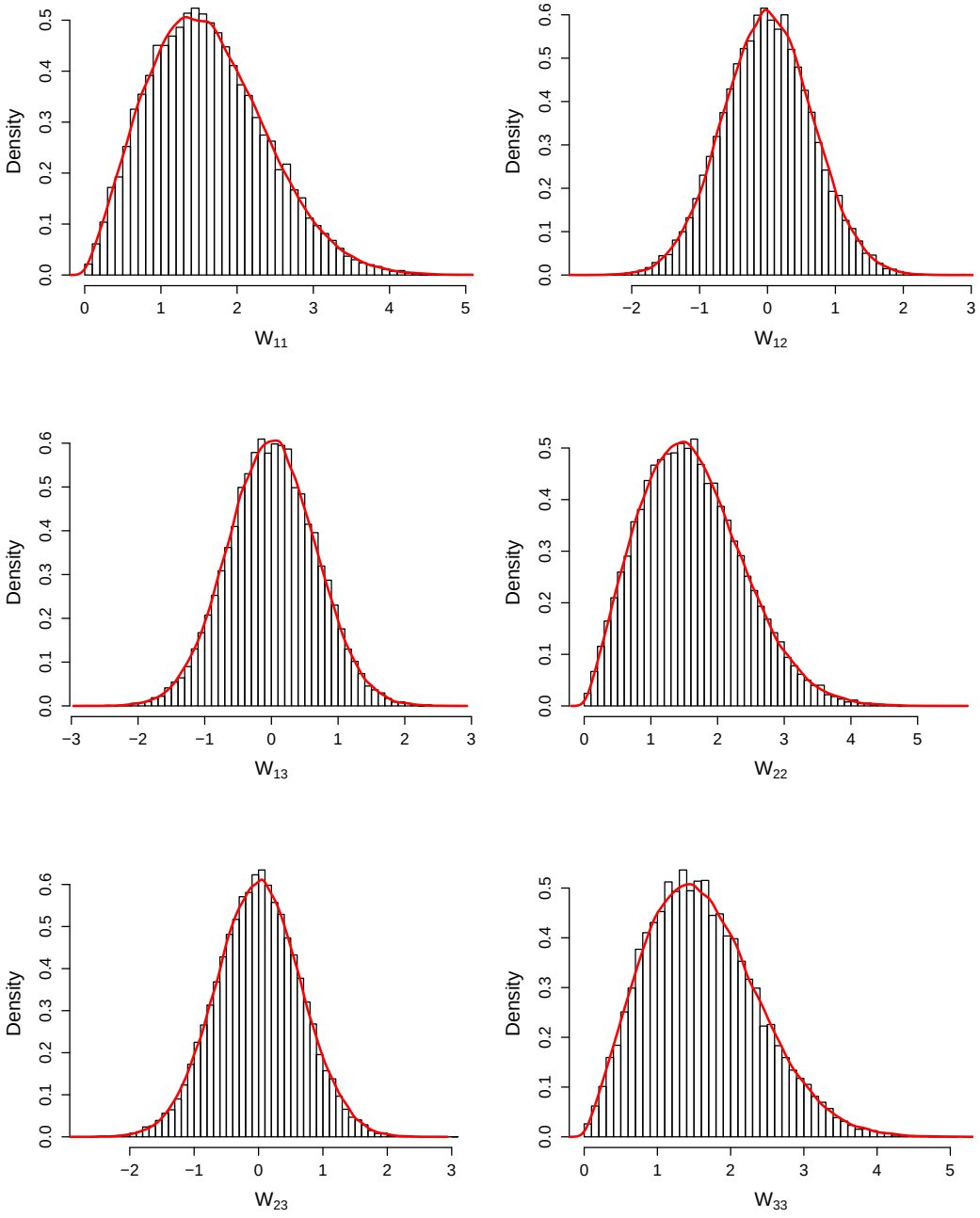


Figure 7.5: Output of **non-central Wishart** algorithm applied to model problem 2 in Table 7.2. The histograms are estimated from 10^5 MCMC samples and the density in red from 10^5 samples drawn directly from the distribution using the **Rejection** algorithm.

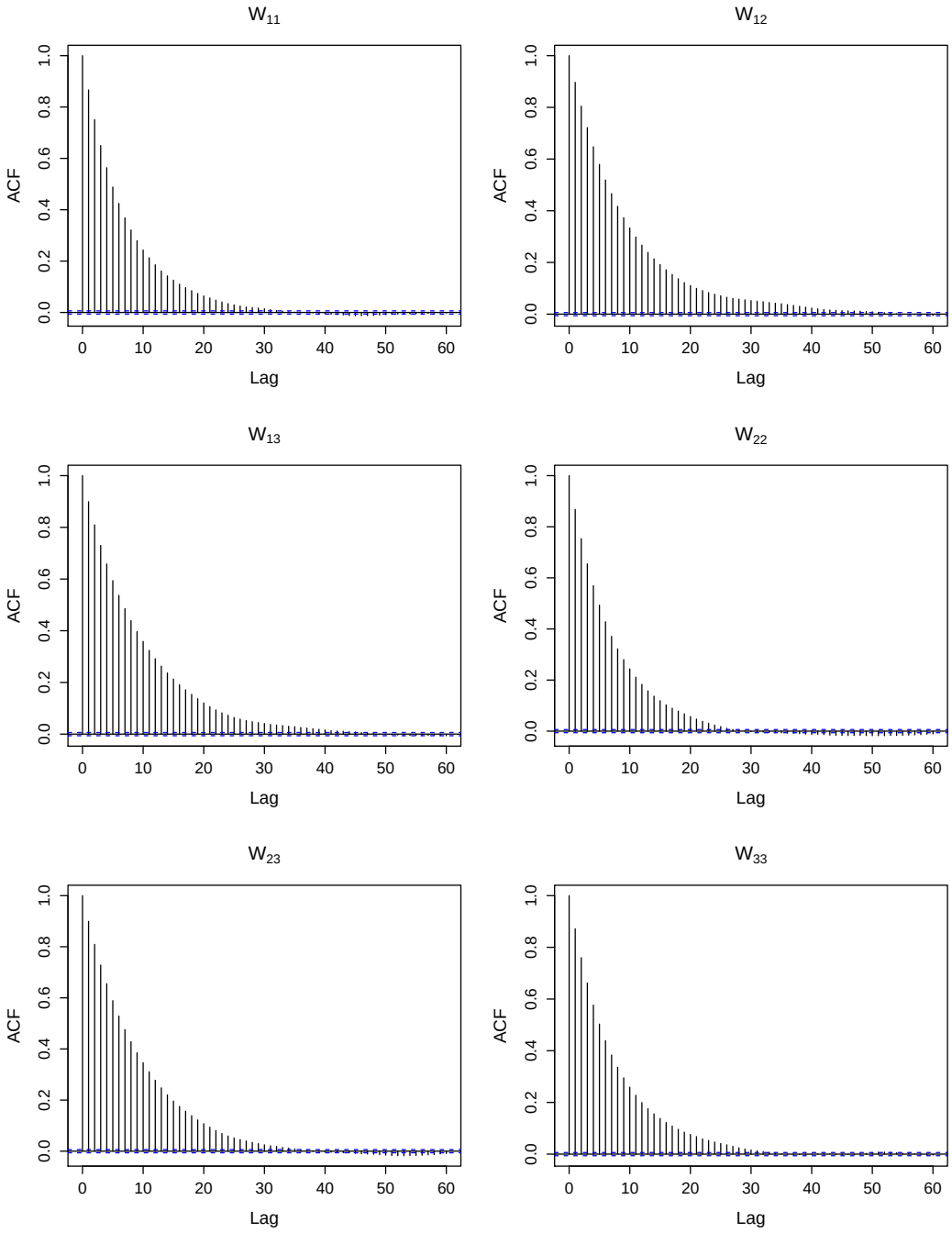


Figure 7.6: Autocorrelation functions of **non-Central Wishart** algorithm applied to model problem 2 from Table 7.2.

order of 10-100 independent samples per second. In practical terms this may still

	Rejection	Random Walk	Component-wise	Central Wishart	Non-Central Wishart
2	50000	2214	58823	346	418
3	2212	475	16667	120	51
4	59	203	4950	53	10
5	1	75	1558	22	3

Table 7.3: The number of independent samples per second for the Monte Carlo algorithms of Table 7.1 applied to model problem 1 of Table 7.2. The results for the Rejection and Component-wise algorithm are calculated from the time taken to draw 10^6 samples; the remainder are Markov Chain algorithms and so also include the efficiency factor as described in the text.

be acceptable. However, with this type of algorithm there is a risk that the chain is not exploring the full space and so the efficiency would be overestimated. The **Component-wise** algorithm draws uncorrelated samples directly from the target density. Although it can become quite slow it performs well compared to the other algorithms. Tables 7.3 and 7.4 show that it is slower on model problem 2. Although it takes the same time to calculate the upper and lower bounds of the Normal densities in this algorithm it seems that a larger number of proposals are needed within the rejection sampler for model problem 2. T that there is his could be because the truncation boundaries are more often within the tails of the distribution and there is a more complex boundary though further work would be needed to quantify this.

The time to draw 10^6 samples from the **Central Wishart** is less than for the **Component-wise** algorithm as the dimension increases. It is very simple to make proposals from a Wishart distribution, the algorithm is only impaired by the longer time needed to compute the transition densities using Eq. (7.10). It is just affected by poor efficiency in higher dimensions as reflected in Figure 7.4. Table 7.4 show that the performance is only slightly worse for model problem 2 Note that here we adjusted the degrees of freedom in the proposal to achieve an acceptance probability of approximately $0.2 - 0.3$, though as shown in Figure 7.4b this maximum efficiency remains low.

The **non-Central Wishart** algorithm takes a lot longer to run due to the need to compute the hypergeometric function in Eq. (7.11) and so the number of samples per second becomes very low as shown in both Tables 7.3 and 7.4. To compute the hypergeometric function we used code made available by Koev and Edelman [2006]. This was the major bottleneck in the compute time as the number of terms needed for the series to converge could be large and variable. Further study of the hypergeometric function could lead to an efficient method by which it can be approximated, greatly reducing the compute time. Note that further work could also be undertaken to understand how the free parameters $\mathbf{\Pi}$ and $\mathbf{\Sigma}$ affect

	Rejection	Random Walk	Component-wise	Central Wishart	Non-Central Wishart
2	4716	7105	28571	760	396
3	200	382	8620	267	31
4	5	154	3175	79	3
5	1	57	1192	33	1

Table 7.4: The number of independent samples per second for the Monte Carlo algorithms of Table 7.1 applied to model problem 2 of Table 7.2. The results for the Rejection and Component-wise algorithm are calculated from the time taken to draw 10^6 samples; the remainder are Markov Chain algorithms and so also include the efficiency factor as described in the text.

the efficiency of this algorithm. Here, we have used an ad hoc method of setting these parameters without any tuning. In conclusion, it seems that the **Central** and **non-Central Wishart** algorithms are novel and could even be useful for the right problem, potentially some complex matrix distribution where there is no means of sampling components individually like the Truncated Normals studied here.

In our applications, when we require the inference to be constrained, we use the **Component-wise** algorithm. This is because it does not require any problem specific tuning and, although it is complicated to implement, it performs reasonably well.

7.3 Using a Stability Matrix as Prior

The aim of deriving a stability matrix and developing efficient ways to sample it was so that it can be used as prior information for parameters in the cubic models studied in Chapter 5. Here we demonstrate its effect on the posterior distribution of parameters estimated for an example problem of the form Eq. (5.1), with randomly generated parameters. As in Figure 5.16 we compute the full posterior for parameters but also estimate the density for just those parameters that give stable solutions to the resultant SDE. This was calculated numerically by simulating the SDE for each parameter vector and recording whether the solution remained bounded. Using data from an arbitrary two dimensional model of the form Eq. (5.1) with $N = 100$ and $\Delta = 0.1$ we estimated all 20 of the parameters entering the drift term. Those parameters in the diffusion function were fixed. The **Component-wise** algorithm was used to sample the 8 cubic parameters using a stability matrix of the form Eq. (7.7) while the others were sampled using the standard Gibbs sampler of Section 5.4.1. We estimated the posterior distributions using 3×10^6 MCMC samples taken from 3 chains after checking each had converged to the same distribution.

The results, in Figure 7.7, compare the full posteriors, stable posteriors and

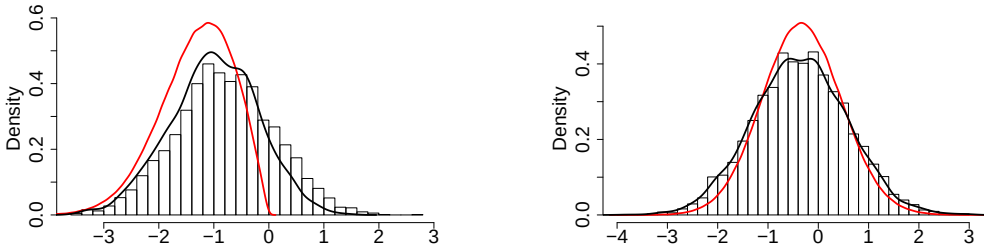
posteriors which use a stability matrix as prior. In this case the subset of stable parameters is very similar to the full posterior: the stable parameters account for 80% of the whole distribution. The posterior which includes the stability matrix prior is close to the full posterior but has some different features, particularly for those parameters that enter the diagonal components of the stability matrix. The stability matrix restricts them to be negative, which is evidently much too strong a constraint in this case. Work would need to be done to remove this constraint. In general, the prior information, in its current format, is too restrictive but there are several possibilities for relaxing the constraints while ensuring stable SDEs are inferred.

The form of matrix given in Eq. (7.6) is not the only possible way of deriving a matrix that satisfies Eq. (7.5). One could think of a method of entering components into $\boldsymbol{M}$ such that they are all off diagonal. Also it would be useful to make the matrix larger so that no two parameters enter the same component. Of course with a larger matrix there will be some redundant components that are equal to 0. This may cause a problem for the sampling strategy, particularly those algorithms based upon the Wishart distribution. The probability of proposing a matrix with one component set at a definite value is 0 so these algorithms might not be applicable. However, it would still be possible to use the **Component-wise** algorithm without much alteration.

Further study would lead to a greater understanding of the minimally restrictive conditions that can be derived to enforce stochastic stability. This could either be developed using the same Lyapunov function used here, namely the simple squared Euclidean norm, or could involve research into other Lyapunov functions. Still using the same Lyapunov function we could learn how to implement the stability bound in Theorem 1 that includes other parameters besides the cubic terms. In particular this would include the parameters that enter the stochastic terms. This would be a departure from the approach of Majda et al. [2009] and may lead to a more general approach of using prior information to infer non-linear SDEs.

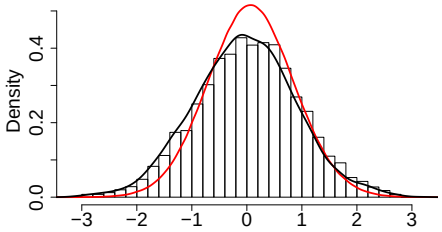
7.4 Summary and Conclusions

In this chapter we have addressed the problem of stochastic stability for SDEs of the form Eq. (5.1) inferred from data. We have proposed a solution, motivated by the work of Majda et al. [2009], which implements an energy constraint on the system. In Section 7.1.2 we derived a means of casting this energy constraint into the requirement that a certain matrix be negative definite. This matrix's components

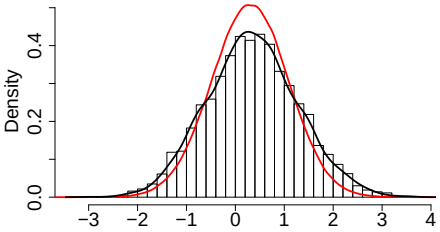


(a) $A_{1,7}$

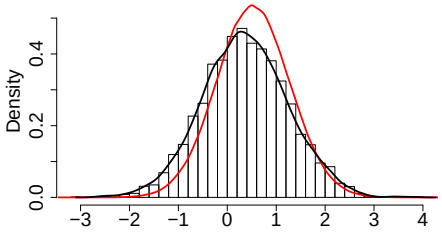
(b) $A_{1,8}$



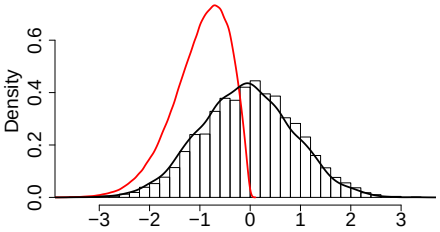
(c) $A_{1,9}$



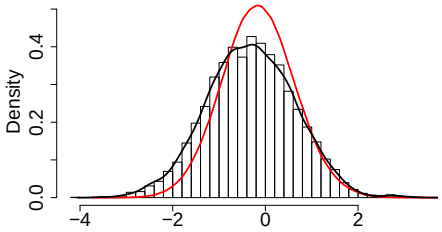
(d) $A_{1,10}$



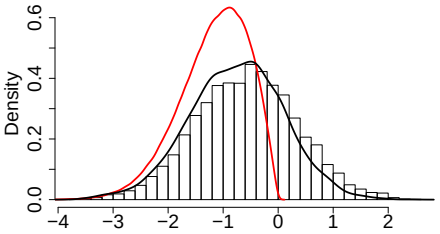
(e) $A_{2,7}$



(f) $A_{2,8}$



(g) $A_{2,9}$



(h) $A_{2,10}$

Figure 7.7: Estimate posterior distributions for parameters from a two dimensional model of the form Eq. (5.1) with $N = 100$ and $\Delta = 0.1$. The parameters, which are randomly generated, are written in the matrix notation introduced in Section 5.4.1. The histograms are the posterior distributions with uninformative prior, in red are the posterior distributions for parameters with stable SDEs and in black are the posterior distributions which include the stability matrix prior information derived in this chapter.

are the parameters entering into the cubic terms of the SDE. Requiring this Stability Matrix to be negative definite places bounds on the domain of these parameters. This is included in the Bayesian framework as prior information.

This novel use of prior information has consequences for the MCMC algorithm used for inference; the Gibbs sampler of Section 5.4.1 is no longer applicable. In Section 7.2 we considered five different algorithms to sample the Stability Matrix. These included basic rejection and random walk sampling, which were found to be inefficient compared to a component-wise algorithm. This Component-wise algorithm is complicated to implement as it involves solving a quadratic equation to compute the upper and lower bounds of each parameter. It then implements a rejection algorithm to sample truncated Normal distributions. However, it was found to be more efficient than algorithms based upon the Central and Non-Central Wishart distributions. As far as we are aware, these distributions have not been used as proposals in a Metropolis-Hastings algorithm and the work here is new. We studied how to select the parameters of the Central Wishart distribution in Section 7.2.3 and found that the optimal efficiency corresponded to an acceptance rate close to 0.234, which corresponds to a broad class of Metropolis-Hastings algorithms [Roberts et al., 1997]. We derived the Non-Central Wishart algorithm in Section 7.2.4. This is a novel use of this distribution. However, it is not clear how to tune the parameters and the algorithm is very slow computationally due to the need to calculate matrix Hypergeometric functions for the proposal density. Further work needs to be done to understand how to optimise this algorithm.

Based on the theory of stochastic stability discussed in Section 2.6 we know that negative definiteness of the Stability Matrix is a sufficient condition to ensure the inferred parameters lead to SDEs whose solutions remain bounded. However, as discussed in Section 7.3, in its present form, it is likely overly restrictive on the space of parameters. There are many ways in which the constraints could be relaxed while ensuring stochastic stability. The matrix could be enlarged such that any increase in dimension adds no further restriction on the parameter space. This limiting matrix, if it exists, would then be minimally restrictive. The Component-wise algorithm, derived in Section 7.2.2 would still be able to sample this matrix. Further work, involving more detailed study of matrix spaces could be pursued in this direction.

The methods in this Chapter could be developed into a very general framework for including stability as prior information for SDE inference. Different Lyapunov functions could be tested to see if this leads to any practical algorithms that can be derived. For the inference problems in this thesis we find it useful to implement the Component-wise algorithm to sample the Stability Matrix as we find

that the advantage of being guaranteed a stable SDE outweigh the fact that the prior has an unquantifiable influence on the posterior and in some cases may affect the estimates. In the next Chapter we apply the methods developed here, and the previous two chapters, to fit SDEs of the form Eq. (5.1) to the dynamical systems discussed in Chapter 3.

Chapter 8

Applications to Geophysical Models

In this chapter we demonstrate the inference methods developed in Chapters 4 and 5 by applying them to the simple toy models discussed in Chapter 3. We investigate the effects of discrete time observation of a system by performing inferences with various amounts of imputed data. We show that errors due to low frequency observation can cause significant error in predictive skill of a model by computing the autocorrelation functions and comparing them to the full system. We also demonstrate that models with a latent noise process can offer an improvement over the basic model structures predicted by the standard homogenisation method when there is no significant time scale separation.

8.1 Chaotic Lorenz System

The first system we consider is the cubic model coupled to the chaotic Lorenz system. It is fully deterministic and consists of a slow variable representing a climate process and three fast variables representing weather fluctuations. The slow variable moves inside a double well potential and is perturbed by the chaotic Lorenz system, which acts as noise. It has been shown by Mitchell and Gottwald [2012] that, in the limit of complete time scale separation, the resolved variable in this system can be modelled by a one dimensional, cubic SDE. In this case no approximations are needed to go from the deterministic to the reduced stochastic model so it will serve as a good test case for other aspects of the model reduction. Therefore, we use this example to explore the effects of lack of time scale separation, which would likely be the case for real atmospheric variables. We also test the methods developed in Chapter 6 by

fitting a model with a latent process. We see that this has potential for capturing memory effects in data with lack of time scale separation.

This section is also a first test of the likelihood based inference for SDE models. This method of inference is not routinely applied in the atmospheric sciences and the results here demonstrate that this principled approach is especially useful in situations where the data is not sampled at a high frequency. We compare the results from the Bayesian data imputation methods of this thesis with a previous non-parametric method that relies on high frequency data. The model equations are as follows

$$\begin{aligned}
\frac{dx}{dt} &= x - x^3 + \frac{4}{90\epsilon} y_2 \\
\frac{dy_1}{dt} &= \frac{10}{\epsilon^2}(y_2 - y_1) \\
\frac{dy_2}{dt} &= \frac{1}{\epsilon^2}(28y_1 - y_2 - y_1 y_3) \\
\frac{dy_3}{dt} &= \frac{1}{\epsilon^2}(y_1 y_2 - \frac{8}{3} y_3)\,.
\end{aligned} \tag{8.1}$$

In Chapter 3 we discussed that the homogenised equation for the slow variable alone is given by

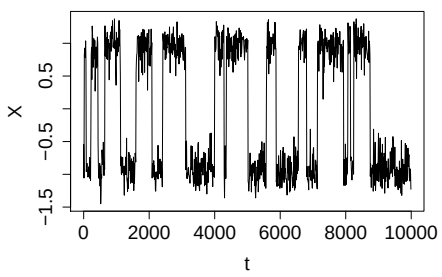
$$dX = X(1 - X^2)dt + \sigma dB_t\,. \tag{8.2}$$

This equation was used by Mitchell and Gottwald [2012] as a test for data assimilation in reduced systems: estimating the unknown state of a system given partial, noisy observations. Firstly, however, an estimate for σ is needed. [Mitchell and Gottwald, 2012] estimate drift and diffusion functions from their definition as conditional averages (see Chapter 2):

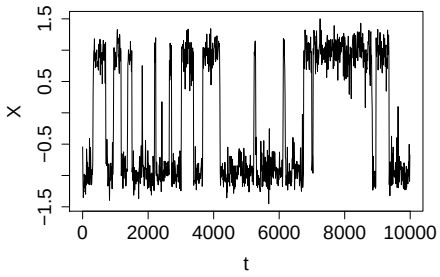
$$A(x) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left\langle X(t + \Delta t) - x \right\rangle \Big|_{X(t)=x}$$

$$B(x)^2 = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \left\langle (X(t + \Delta t) - x)^2 \right\rangle \Big|_{X(t)=x}$$

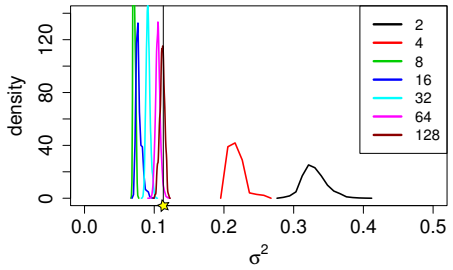
These estimates are obtained by dividing the space into bins $[X, X + \Delta X]$ and using a fixed observation interval Δt. It is not easy to estimate the errors in this method [Sura and Barsugli, 2002]. One approach is to just repeat the procedure for varying Δt to check consistency. The method has been applied several times to estimate models from time series data in atmospheric/ocean sciences: Sura [2003] fit a one dimensional model with multiplicative noise to sea surface wind data; Stemler et al.



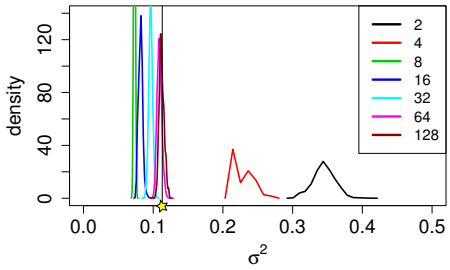
(a) Path of x from Eq. (8.1) for $\epsilon = 0.1$



(b) Path of x from Eq. (8.1) for $\epsilon = 0.01$



(c) Posterior distributions for $\epsilon = 0.1$

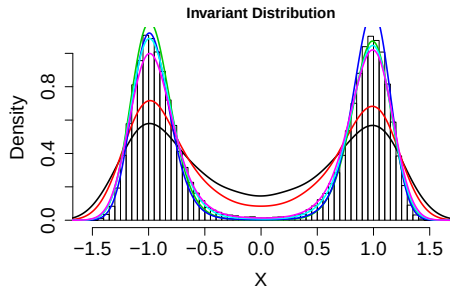


(d) Posterior distributions for $\epsilon = 0.01$

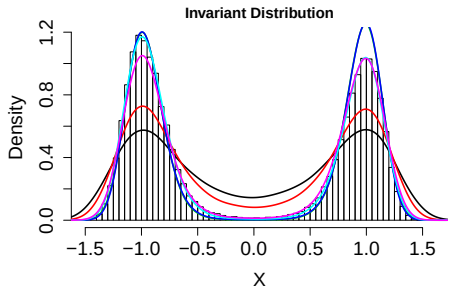
Figure 8.1: Inference for the reduced double well model coupled to chaotic Lorenz system: Eq. (8.1) for two values of ϵ. The stars show the value $\sigma^2 = 0.113$ obtained by Mitchell and Gottwald [2012].

[2007] demonstrate that the method is useful to estimate low dimensional models when there is no explicit time scale separation in the system; and Berner [2005] fit a non-linear model to planetary waves. As discussed extensively in Chapter 4, we use likelihood based parametric estimation. In a Bayesian context, parameters obtained using MCMC have error estimates readily available. The particular advantage that we demonstrate here is that they can be used when the observation interval Δ is large.

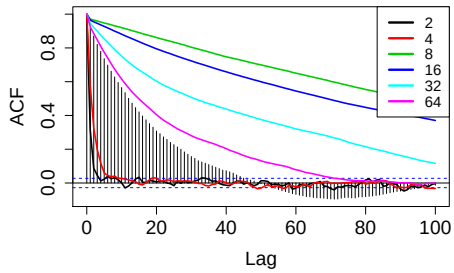
We applied Algorithms 4.1 and 4.2 using the Modified Bridge (see Table 5.1), to the reduced model Eq. (8.2) with $N = 1000$ observations and observation interval $\Delta = 10$ for two different time scale separations: $\epsilon = 0.01$ and $\epsilon = 0.1$. The data used for the inference is shown in Figures 8.1a and 8.1b. In the first instance we assume that the drift is known and just do the inference for σ with an uninformative prior ($\sigma \sim \mathcal{N}(0, 10)$). The posterior distributions were estimated using 10^5 MCMC samples from three chains after discarding 10^4 samples as burn in. The estimated posterior distributions, for various amounts of imputed data m, are shown in Figures 8.1c and 8.1d



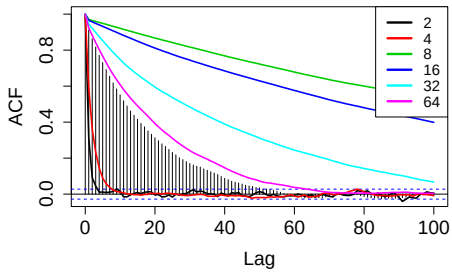
(a) Invariant distribution with $\epsilon = 0.1$. Histogram is for the full system, the lines correspond to different m.



(b) Invariant distribution with $\epsilon = 0.01$. Histogram is for the whole system, the lines correspond to different m.



(c) Autocorrelation function for $\epsilon = 0.1$. Bars are for the full system.

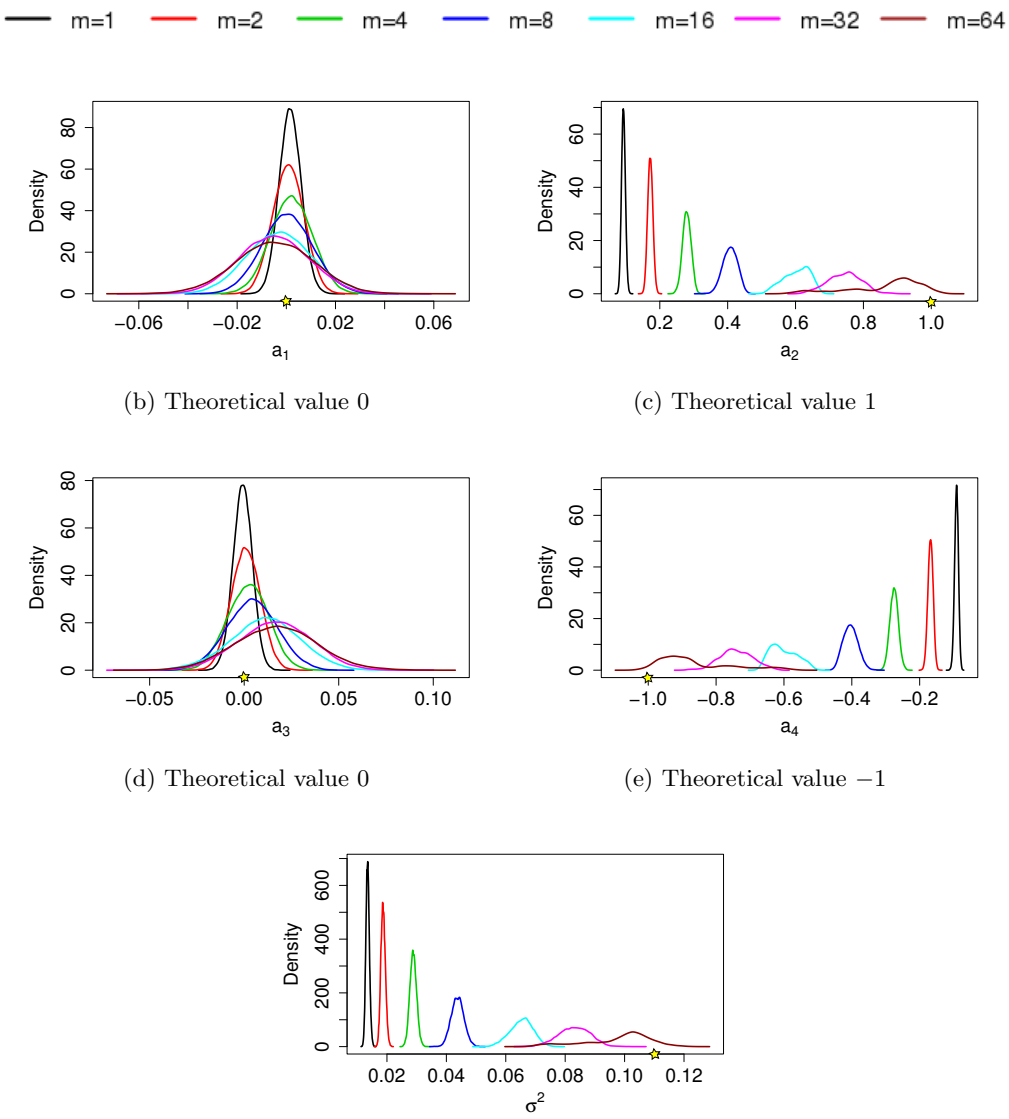


(d) Autocorrelation function for $\epsilon = 0.01$. Bars are for the full system.

Figure 8.2: Predictive statistics for the reduced double well model coupled to chaotic Lorenz system: Eq. (8.1) for two values of ϵ. In each plot the lines correspond to the inferred one dimensional model for different m.

With the large Δ we use here, the posterior distributions only start to converge to a consistent value for $m \geq 64$. The star on the horizontal axis indicates the value $\sigma^2 = 0.113$ obtained by Mitchell and Gottwald [2012] using the observation interval $\Delta = 0.0005$. It is encouraging that our method, applied to discretely observed data, can reproduce this value, obtained from effectively continuous time observation and the non-parametric method discussed above.

Figure 8.2 shows the predictive skill of the one dimensional reduced model using values for σ estimated for various m. Figures 8.2a and 8.2b show that the reduced model can reproduce the double well distribution of the full although the separation and depth of each well is underestimated for $m = 2$ and $m = 4$ due to the larger noise. For $m \geq 8$ the model reproduces well the full models marginal distribution for x. It is not clear whether there is much difference between $m = 8$ and $m = 64$. However, observing Figures 8.2c and 8.2d we see that the autocorrelation function for the full model is much better approximated when $m = 64$. Specifi-



(b) Theoretical value 0

(c) Theoretical value 1

(d) Theoretical value 0

(e) Theoretical value -1

(f) Theoretical value 0.11

Figure 8.3: Posterior distribution estimates from MCMC output applied to a sparse data set ($\Delta = 10$). Distributions correspond to different amounts of missing data m between observations with the key shown at the top. The distribution in brown, for $m = 64$, agrees with the theoretical values predicted by the homogenisation procedure.

cally, one dimensional models can not capture the negative correlation at large lags exhibited by the full model.

We now fit a one dimensional cubic SDE to the data without any assumptions from the homogenisation procedure. We just consider the general cubic form argued

for in Chapter 3

$$dX_t = (a_1 + a_2 X_t + a_3 X_t^2 + a_4 X_t^3)dt + \sigma dB_t \tag{8.3}$$

and estimate all of the parameters $\{a_1, a_2, a_3, a_4, \sigma\}$ from sparse observations of the system: again using $\Delta = 10.0$ and $N = 1000$. To update the drift parameters we use the Gibbs sampler of Section 5.4.1. The estimated posterior distributions are shown in Figure 8.3. A lot of imputed data is needed before the estimates start to converge towards the values predicted by homogenisation but the inference demonstrates that there is enough information in the sparse data set if the likelihood is well approximated.

The large amount of data imputation needed can be understood by considering the quadratic variation of the process. The quadratic variation is directly related to the diffusion function of the process as seen in Eq. 4.2. For a process X_t observed over a fixed time interval $[0, T]$, it can be estimated as a function of the observation interval $\Delta = T/(N-1)$ by

$$QV(\Delta) = \frac{1}{\Delta} \sum_{i=0}^{N-1} (X_{i+\Delta} - X_i)^2 \,, \tag{8.4}$$

where N is the number of observations. For a diffusion process one would expect the quadratic variation to be independent of the sampling frequency for a range of time scales. This is due to the scale invariant property of diffusions. Below an upper limit one should see the quadratic variation "plateau" at a constant value. This can be seen in Figure 2.1b, where the quadratic variation is estimated for Brownian motion. Figure 8.4 shows the quadratic variation of x from Eq. (8.1) calculated for a range of observation frequencies Δ for different ϵ. It shows that there is a maximum around the values $0.1 - 0.3$. This is the time scale best modelled by a diffusion process and agrees with the $m \geq 64$ data imputation when observing at interval $\Delta = 10$.

The full model with $\epsilon = 0.1, 0.01$ is well approximated by the reduced model. This may not be the case when there is no real time scale separation: if $\epsilon = 0.5$ or even $\epsilon = 1.0$. This means that y_2 in Eq. (8.1) is not well approximated by a white noise process. Instead we consider a red noise: a latent linear process that has a non-zero autocorrelation time. The approximating equation then has the form

$$\begin{aligned} dX_t &= (a_1 + a_2 X_t + a_3 X_t^2 + a_4 X_t^3)dt + Y_t dt \\ dY_t &= -\gamma Y_t dt + \sigma dB_t \,. \end{aligned} \tag{8.5}$$

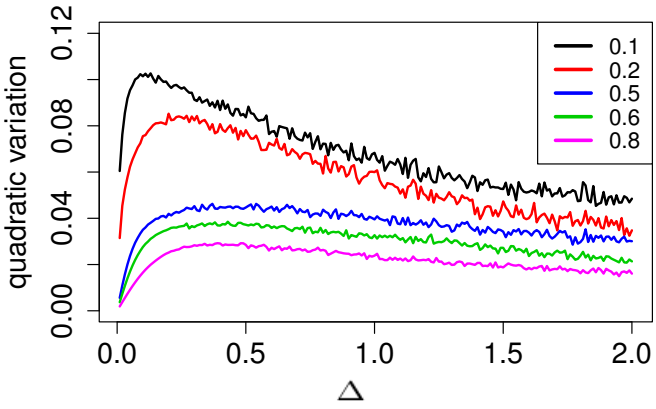


Figure 8.4: Quadratic variation, calculated as Eq. (8.4), for the process x_t from the model Eq. (8.1). The curves represent different time scale separations ϵ.

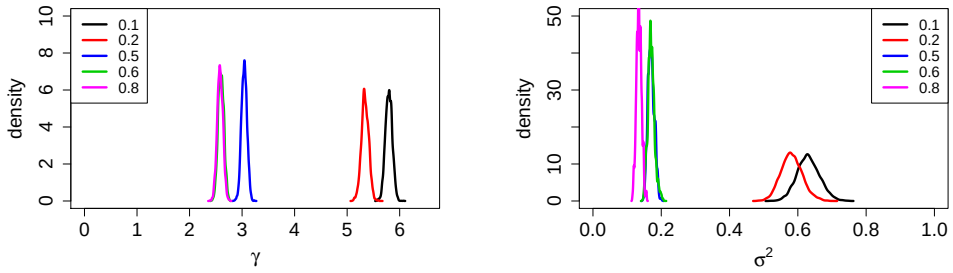
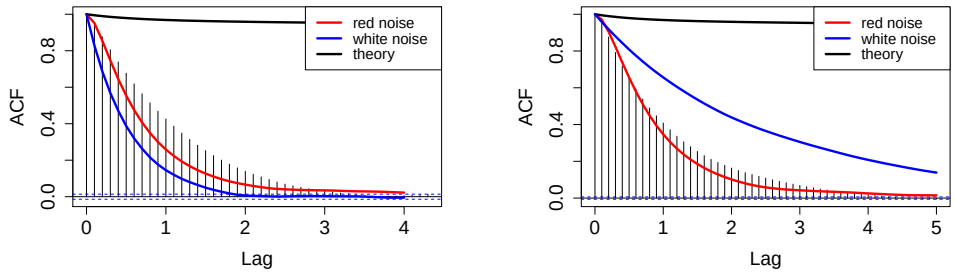


Figure 8.5: Posterior estimates for parameters from the model Eq. 8.5 using $N = 1000$ observations of Eq. (8.1) with observation interval $\Delta = 0.01$. The posteriors were estimated using 10^5 samples from 3 chains after discarding a burn in of 10^4 samples. The different posteriors are for varying time scale separation ϵ.

The Brownian motion acts on the unobserved process Y_t, which then forces the observed variable X_t. The parameters in this model can be inferred according to the algorithm in Section 6.1. We used $N = 1000$ high frequency observations with $\Delta = 0.01$ for different time scales ϵ. We did not impute missing data between observations, just the latent noise process Y_t. We used 10^5 MCMC samples from three chains after discarding a burn in of 10^4. We used the fixed theoretical values for the diffusion function, namely $a_1 = 0$, $a_2 = 1$, $a_3 = 0$ and $a_4 = -1$ and estimated the diffusion parameters. Figure 8.5 shows the posterior estimates for the unknown parameters γ and σ for varying ϵ.

We compared the predictions of the latent noise model to the standard model in Eq. (8.3) for $\epsilon = 0.6$ and $\epsilon = 0.8$. We used the posterior mean values to



(a) $\epsilon = 0.6$ (b) $\epsilon = 0.8$

Figure 8.6: Plots comparing the autocorrelation function of the full chaotic Lorenz model with reduced models. The bars are for the full model; red is the latent noise process; blue is the standard empirical model and black is the theoretical model predicted by homogenisation.

produce predictive simulations. The resulting autocorrelation plots are shown in Figure 8.6 and compared to those from the full model, the theoretical model and the one dimensional empirical model. In this case the latent noise model does well in reproducing the short time autocorrelations of the full model much better than the standard or theoretical models. This is likely due to the smoothness of paths simulated from the latent noise process.

This is a useful result and encourages further work into models with latent noise processes to model systems such as Eq. (8.1). This approach is a good example of an approach motivated jointly by a theoretical result (the derivation of the drift function) and an empirical approach. This empirical work depended upon an inference method that made estimating the parameters γ and σ. However, this is still a challenging inference problem. The Markov Chain does not mix as well as for the fully observed models and finer observation intervals are needed.

Further work could be focussed upon a systematic model comparison by calculating Bayes' factors between models Eq. (8.3) and Eq. (8.5). This is a challenging problem but has been attempted for a restricted class of SDE models (see Polson and Roberts [1994]).

8.2 Model Reduction for Triad Systems

In this section we apply the Empirical Mode Reduction Strategy to the Burgers model, introduced in Section 3.5.3. This is a deterministic system with 52 components and the aim here is to reduce it to a two dimensional stochastic system that accurately represents some of the features of the original. There are two aspects to

this problem. The first step is to represent the effects of the unresolved modes as a stochastic process. The second is to follow the homogenisation procedure for the triad model, detailed in Section 3.5.2.

Strictly the homogenisation procedure relies upon a complete time scale separation between the resolved and unresolved modes. As in the previous section, where we studied the three dimensional Lorenz model, we control this time scale separation explicitly with parameter ϵ. Complete time scale separation is given by $\epsilon \to 0$. in this section we assess the sensitivity of the results to finite values of ϵ.

As an alternative we fit a general cubic model of the form Eq. (5.1) to assess if an empirical approach, where there are a lot of parameters to infer, compares to the homogenisation method. Particularly we are interested in the case where there is no time scale separation.

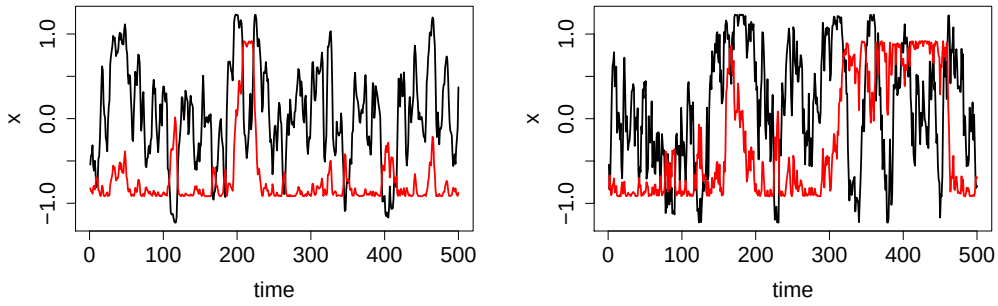
In Chapter 3, Section 3.5.3 we calculated the Fourier transform of the Burgers PDE. Through a single component we coupled this to a triad model of the form studied in Section 3.5.2. The resulting system is given by

$$\begin{aligned}
\frac{dx_1}{dt} &= \frac{b_1}{\epsilon} x_2 y_1 \\
\frac{dx_2}{dt} &= \frac{b_2}{\epsilon} x_1 y_1 \\
\frac{dy_k}{dt} &= \frac{b_3}{\epsilon} x_1 x_2 \delta_{1,k} - \mathrm{Re}\frac{ik}{2\epsilon^2} \sum_{p+q+k=0} \hat{u}_p^* \hat{u}_q^* \\
\frac{dz_k}{dt} &= -\mathrm{Im}\frac{ik}{2\epsilon^2} \sum_{p+q+k=0} \hat{u}_p^* \hat{u}_q^* \,.
\end{aligned} \tag{8.6}$$

This system retains finite values provided that $b_1 + b_2 + b_3 = 0$ (see Majda et al. [2002]). We use the values $\boldsymbol{b} = \{0.9, -0.5 - 0.4\}$ and, as in Section 3.5.3, we choose a spherical cut off $\Lambda = 50$.

8.2.1 Stochastic Mode Reduction

We are interested in eliminating $\boldsymbol{y}$ from Eq. (8.6) leaving equations for just x_1 and x_2. The small parameter ϵ represents the time scales within the system. The variables $\boldsymbol{y}$ have fastest time scale of order $O(1/\epsilon^2)$ compared to $O(1/\epsilon)$ for x_1 and x_2. As $\epsilon \to 0$ we can use the method of homogenisation for SDEs (see Chapter 3,



(a) $\epsilon = 1.0$ (b) $\epsilon = 0.01$

Figure 8.7: Example of solution x_1 (black) and x_2 (red) from the Burgers model in Eq. (8.6) for two values of ϵ.

Sections 3.5.2 and 3.5.3) to eliminate the fast variables and give

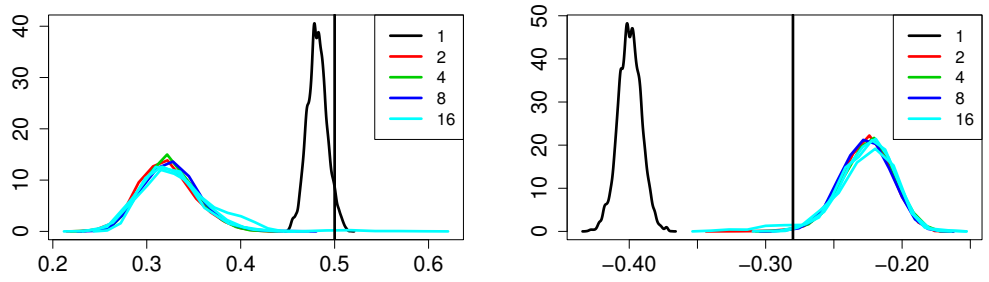
$$\begin{aligned} dx_1(t) &= \frac{b_1}{\gamma}(b_3 x_2^2(t) + \frac{\sigma^2}{2\gamma}b_2)x_1(t)dt + \frac{\sigma}{\gamma}b_1 x_2(t)dB_t \\ dx_2(t) &= \frac{b_2}{\gamma}(b_3 x_1^2(t) + \frac{\sigma^2}{2\gamma}b_1)x_2(t)dt + \frac{\sigma}{\gamma}b_2 x_1(t)dB_t\,, \end{aligned} \tag{8.7}$$

where unknown parameters σ and γ have been introduced. Majda et al. [2002] estimate these parameters using the full system. Here we estimate them using the Algorithms 4.1 and 4.2 from observations of the climate variables alone. Unfortunately, this presents a problem: the reduced model in Eq. (8.7) is a two dimensional system with a one dimensional Brownian motion. This means that the likelihood, given by Girsanov's theorem and discussed in Chapter 4, can not be written down. An alternative is to consider a transformation of variables as undertaken in Section 5.3.1 that results in a univariate SDE. However, it appears that for this specific case, no such transformation exists. In this instance we consider the inference problem where the model Eq. (8.7) is driven by two independent Brownian motions.

We apply the inference to a data set with total time $T = 500$ and observation interval $\Delta = 0.1$. We simulate the system for $\epsilon = \{0.01, 0.1, 0.25, 0.5, 0.8, 1.0\}$. Example data for $\epsilon = 1.0$ and $\epsilon = 0.01$ are shown in Figure 8.7.

We retained 10^5 MCMC samples from three chains after discarding a burn in of 10^4 samples. We used Algorithm 4.1 to infer γ and σ and Algorithm 4.2 to impute missing data using the Modified Bridge proposal (see Table 5.1). The values of $b_i, i = 1, 2, 3$ are assumed known. Posterior estimates for the case $\epsilon = 0.8$ are shown in Figure 8.8. The vertical lines are the mean values of the posterior distributions

for the case $\epsilon = 0.01$. Note that, as expected, there is a small discrepancy between the estimates for $\epsilon = 0.8$ and $\epsilon = 0.1$.



(a) Posterior estimate of σ

(b) Posterior estimate of γ

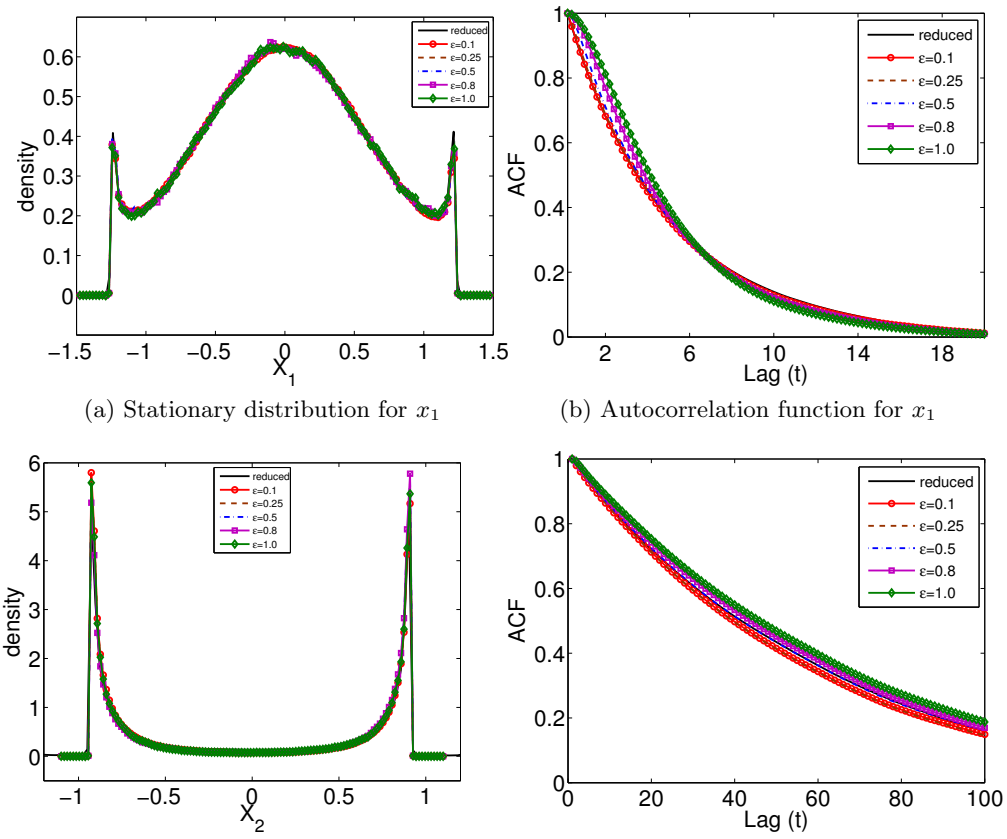
Figure 8.8: Posterior estimates of parameters σ and γ in Eq. 8.7 applied to data simulated from Eq. (8.6) with $\epsilon = 0.8$ for varying amounts of missing data m. These distributions were estimated using 3×10^5 samples from Algorithms 4.1 and 4.2 using the Modified Bridge proposal. The vertical black line is the mean of the posteriors estimated for the case $\epsilon = 0.01$ with $m = 16$.

Figure 8.9 shows the stationary probability densities and autocorrelation functions for Eq. (8.7), estimated for $\epsilon = \{0.1, 0.25, 0.5, 0.8, 1.0\}$. In each case the data is simulated from the full model Eq. (8.6), then the parameters are estimated using the reduced model Eq. (8.7) and this reduced model is simulated to calculate the predictive statistics. The posterior mean estimates were used for γ and σ with $m = 16$ missing data values (it was verified that the posteriors for $m = 16$ and $m < 16$ gave consistent estimates). Also plotted for comparison are the probability densities and autocorrelation functions for data simulated from Eq. (8.7) with parameters σ and γ estimated from data from the full model Eq. (8.6) with $\epsilon = 0.01$. This is referred to as the reduced model in Figure 8.9.

The autocorrelation functions have been collapsed onto the reduced model ($\epsilon = 0.01$) by rescaling the output interval of the prediction by their value of ϵ. The data collapse is very good for all ϵ. This implies that the parameter estimates for each case are compensating for the changing time scale separation.

8.2.2 Empirical Approach

An alternative to fitting a model to "climate" variables is a purely empirical approach where all the parameters of a generic model form are estimated from data. In this case one does not rely upon the reduced model Eq. (8.7) being appropriate



(a) Stationary distribution for x_1

(b) Autocorrelation function for x_1

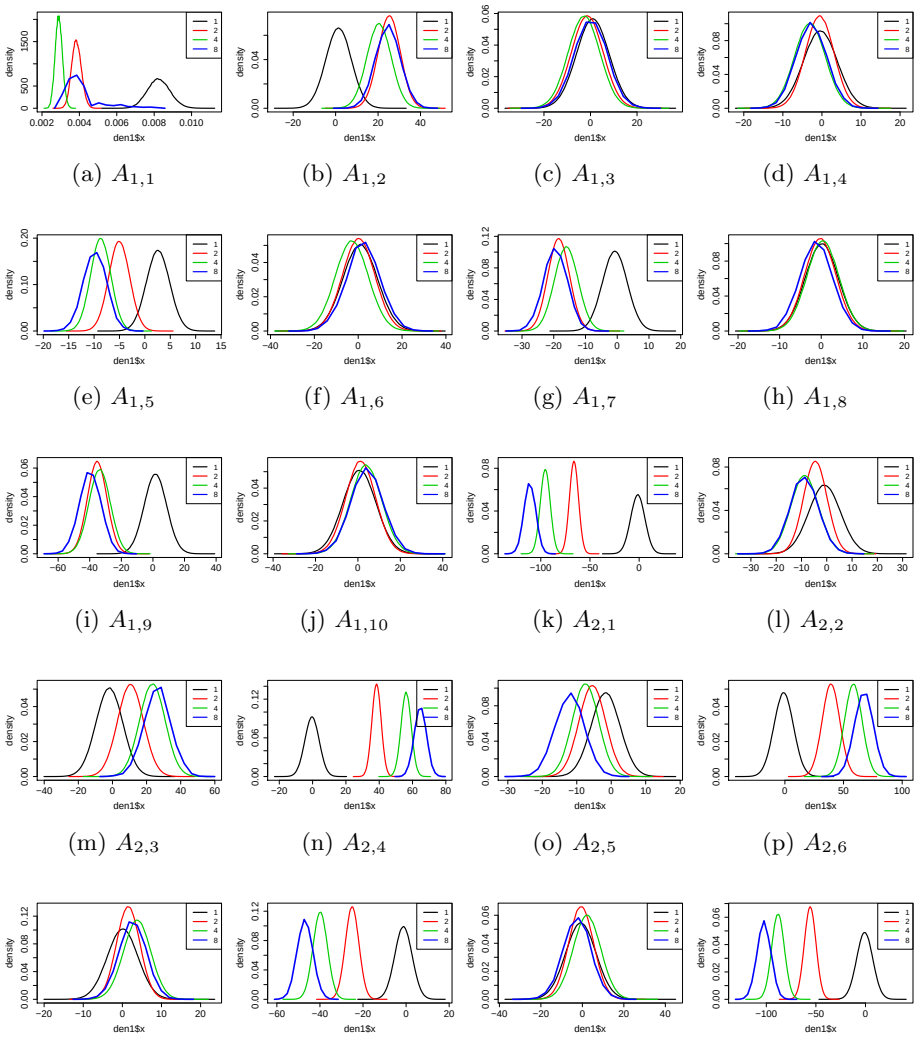
(c) Stationary distribution for x_2

(d) Autocorrelation function for x_2

Figure 8.9: Output statistics comparing the reduced model Eq. (8.7), with parameter estimates for σ and γ, for various ϵ with the full model.

when there is no complete time scale separation but a generic, flexible model form is proposed. For this we use the cubic model in Eq. (5.1) for which we have been developing inference methodology in Chapters 5-7. For a two dimensional system there are 28 unknown parameters. We estimated these from data simulated from the full model Eq. (8.6) with $\epsilon = 0.8$. Algorithm 4.1 was used to estimate the 8 diffusion parameters and 4.2 with the Modified Bridge proposal (Table 5.1 was used to impute the missing data. The Gibbs sampling algorithm in Chapter 5, Section 5.4 was used to sample the 20 parameters entering the drift function. To ensure stability of the resulting SDE we restricted the parameter space using the constraint in Chapter 7. This was sampled using the component-wise algorithm of Section 7.2.2.

We used three MCMC runs, each retaining 10^5 samples after discarding a burn in of 10^4. The posterior estimates for $\boldsymbol{A}$, the parameters in the drift function, are shown in Figure 8.10. Refer to Section 5.4 for details of how parameters in



(a) $A_{1,1}$ (b) $A_{1,2}$ (c) $A_{1,3}$ (d) $A_{1,4}$

(e) $A_{1,5}$ (f) $A_{1,6}$ (g) $A_{1,7}$ (h) $A_{1,8}$

(i) $A_{1,9}$ (j) $A_{1,10}$ (k) $A_{2,1}$ (l) $A_{2,2}$

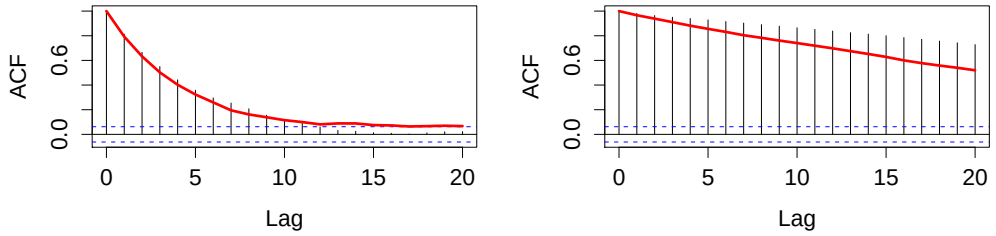
(m) $A_{2,3}$ (n) $A_{2,4}$ (o) $A_{2,5}$ (p) $A_{2,6}$

(q) $A_{2,7}$ (r) $A_{2,8}$ (s) $A_{2,9}$ (t) $A_{2,10}$

Figure 8.10: Posterior estimates of drift parameters for two dimensional cubic model Eq. (5.1) fitted to $N = 5000$ observations with interval $\Delta = 0.1$ of the triad-Burgers equation with $\epsilon = 0.8$. 3×10^5 MCMC samples were retained after discarding a burn in of 10^4, from Algorithms 4.1 and 4.2 with the Modified Bridge proposal. The Gibbs sampler of Section 5.4 was used to sample the matrix $\boldsymbol{A}$ shown here.

matrix $\boldsymbol{A}$ enter the SDE Eq. (5.1). To obtain these results took 2 days computing time on standard CPUs. The estimated posteriors begin to show consistency for $m = 8$ - larger values were taking an impractically long time to converge.

We use Eq. (5.1) in two dimensions, with the posterior estimates from $m = 8$, to form a predictive model for Eq. (8.6). Figure 8.11 compares the autocorrelation



(a) x_1 (b) X_2

Figure 8.11: Autocorrelation plots of the full system in Eq. (8.6) with $\epsilon = 0.8$ (vertical bars) and the empirical model Eq. 5.1) with parameters estimated as described in the text (red).

plots of the full model Eq. (8.6) with those computed from the predictive model Eq. (5.1). These predictions appear good but the predicted stationary probability distributions (not shown) are poor. In this case the empirical mode reduction strategy implemented here struggles to achieve the same predictive skill as the homogenisation procedure followed by parameter estimation demonstrated in Section 8.2.1. However, in this model recall that the problem has been specifically designed for homogenisation with clear differences in time scales. In more realistic systems, such as the one in the next section, it is much harder to identify these time scales and they have to be introduced as a working assumption. Also, the problem with the empirical method is that there are so many parameters to infer and perhaps our means of constraining the parameter space as in Chapter 7 is overly restrictive.

8.3 Model Reduction for the Quasi-Geostrophic Model with Mean Flow

In this section we study the two approaches to stochastic modelling applied to the Quasi-Geostrophic Model with Mean Flow that we derived in Chapter 3, Section 3.5.4. We first consider using the homogenisation procedure followed by inference for the few unknown parameters. Then we infer a general cubic model, estimating all of the parameters using methods developed in Chapters 5-7.

8.3.1 Stochastic Mode Reduction

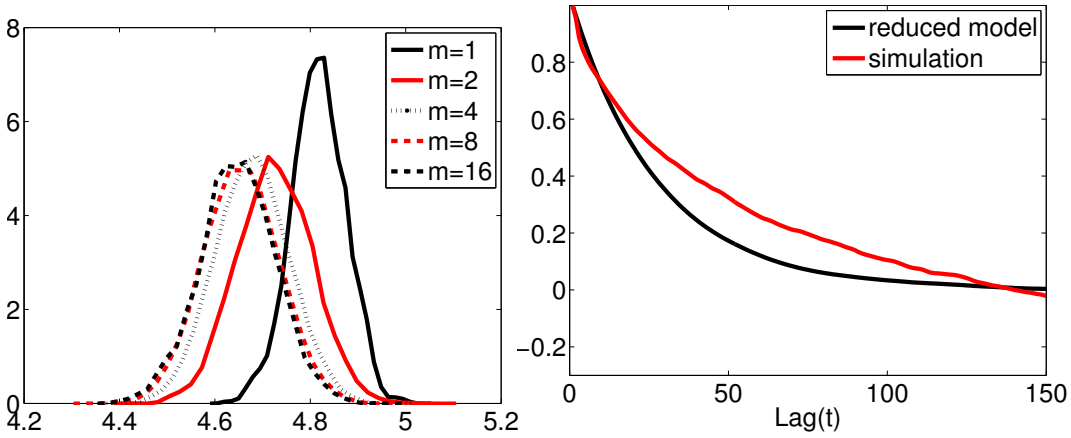
In Section 3.5.4 we derived a one dimensional diffusion model for the mean flow U of the quasi-geostrophic equation Eq. (3.40) on the β-plane by assuming complete time scale separation. This resulted in the SDE

$$dU = (-\tilde{\gamma}(U)U + \gamma'(U))dt + \sqrt{2\gamma(U)}dB_t\,, \tag{8.8}$$

where

$$\gamma(U) = 2\sum_k \frac{k_x^2 H_k^2 \sigma_k^2}{|\gamma_k(U)|^2}, \qquad \tilde{\gamma}(U) = 2\sum_k \frac{k_x^2 H_k^2 \gamma_k}{|\gamma_k(U)|^2}\,. \tag{8.9}$$

and $\gamma_k(U) = \gamma_k + i\Omega_k + ik_x/\sqrt{\alpha\mu}U$. The stochastic approximation introduces unknown parameters γ_k and σ_k. As in Majda et al. [2003], we assume $\sigma_k \approx \gamma_k$. We estimate γ_k from observations of U alone. In the case of real observations of the atmosphere it is more likely that only the large scale variables of interest will be available so estimating γ_k from the full system, as in Majda et al. [2003], would not be possible. We apply Algorithms 4.1 and 4.2, using the Modified Bridge proposal, to $N = 1000$ observations of U at interval $\Delta = 0.1$. We use the case with only one topographic mode such that $H_k \neq 0$ for $k = (1, 0)$, $H_k = 0$ otherwise. This means one only needs to consider a single term in the sums in Eq. (8.9) and that there is only one unknown parameter $\gamma_{(10)}$. We use an uninformative prior for $\gamma_{(10)}$. The estimated posteriors were calculated using 10^5 samples from three chains after discarding a burn in of 10^4. The results are shown in Figure 8.12a. The estimates converge rapidly for increasing imputed data. Figure 8.12b compares the autocorrelation functions of the one dimensional diffusion model, using the estimated value of $\gamma_{(10)}$, with that of the full system. The model works well for short predictions, although does not capture the negative autocorrelation at longer times. A major drawback of the homogenisation, followed by parameter estimation, approach is the need to estimate more parameters as the full system becomes more complex. In this case if there was more than a single topographic mode then this would introduce further γ_k to the point where there is not sufficient information in the data set to infer them all. For this reason the purely empirical approach is useful and the general form of cubic model with linear noise, as argued for in Chapter 3, is an appropriate and flexible model to fit to the data.



(a) Estimates of $\gamma_{(10)}$ (b) Autocorrelation functions

Figure 8.12: Results for inferring the one free parameter $\gamma_{(10)}$ in the one dimensional reduced model in Eq. (8.8) to $N = 1000$ observations at interval $\Delta = 0.1$ from the original system Eq. (3.40). On the left are the posterior estimates, for varying missing data m, obtained using 3×10^5 samples from Algorithms 4.1 and 4.2 with the Modified Bridge proposal. On the right are the estimated autocorrelation functions of the reduced model using the mean of the posterior estimate for $\gamma_{(10)}$ with $m = 16$ compared to the simulation of the original model Eq. (3.40).

8.3.2 Empirical Approach

We estimated the six parameters in the one dimensional cubic model

$$dU = (a_1 + a_2 U + a_3 U^2 + a_4 U^3)dt + (\sigma_1 + \sigma_2 U)dB_t \tag{8.10}$$

using the same observations of U as the previous section. We used Algorithms 4.1 and 4.2 with the Modified Bridge proposal. For this cubic model we are able to use the Gibbs sampler of Section 5.4 to infer the drift parameters. We ran 3 MCMC chains each discarding 10^4 samples as burn in before each retaining 10^5 samples for which to estimate the posterior distributions. Figure 8.13 shows that the posterior estimates are consistent for $m = 16$-32 Figure 8.14 shows the stationary probability densities and autocorrelation functions of the inferred model Eq. (8.10), using the mean posterior estimates for parameters, for different values of m. The plot highlights the importance of imputing data to infer a continuous time model from discrete observations. Figure 8.14a shows that a naive approach, with no imputed data, results in a model that misses the heavy tailed skew shape of the stationary distribution. The predicted stationary distribution improves rapidly with

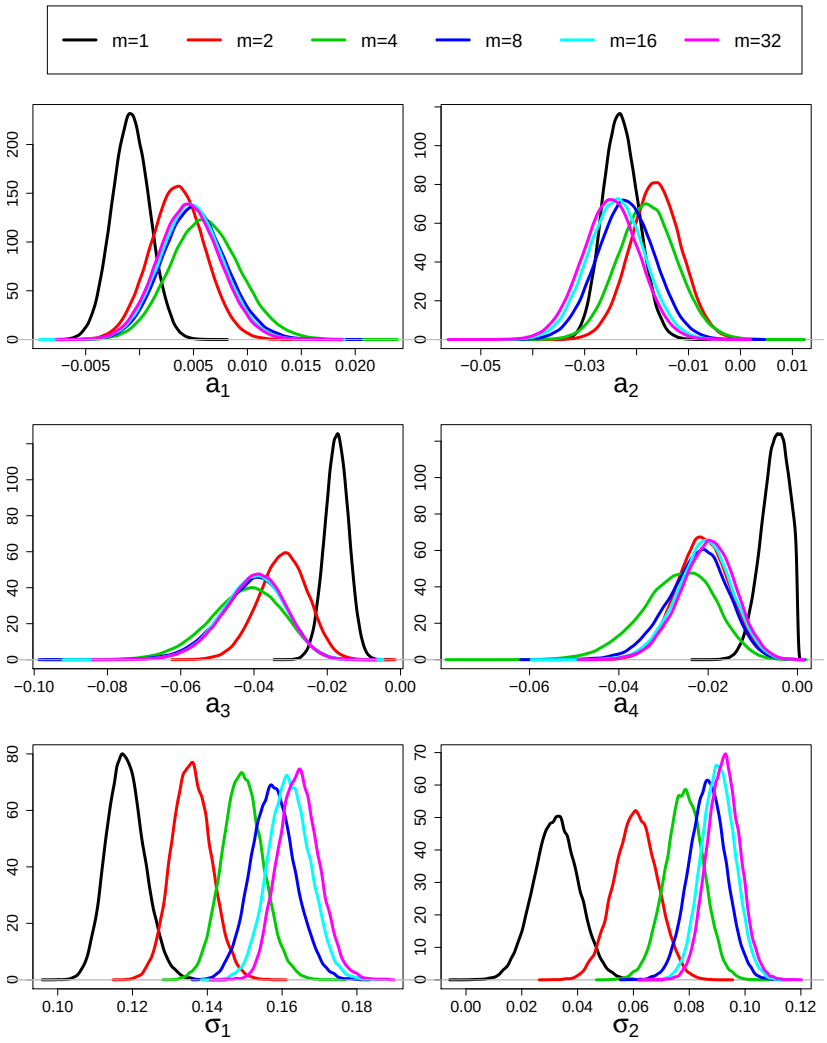
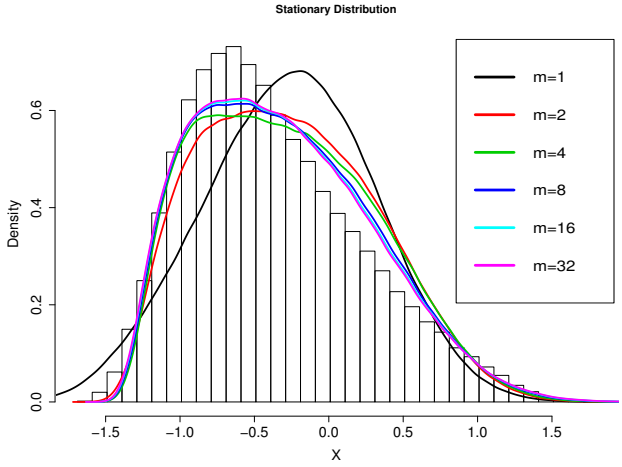
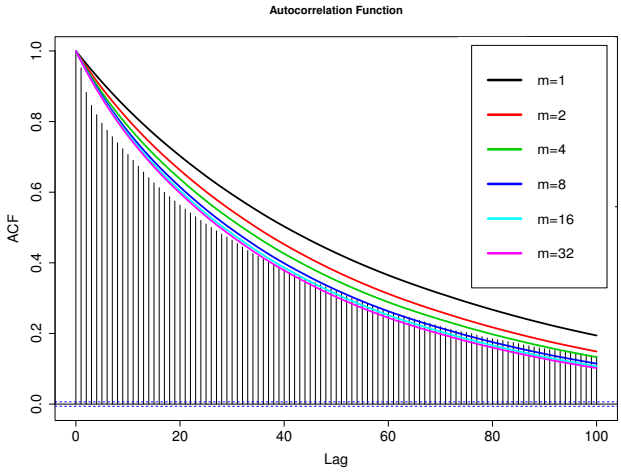


Figure 8.13: Inference results for varying amounts of missing data m for the cubic model Eq. (8.10) fitted to $N = 1000$ observations of the mean flow data with interval $\Delta = 0.1$. 3×10^5 samples were used to estimate these posterior distributions. Algorithm 4.1 was used to estimate the two diffusion parameters (shown bottom), Algorithm 4.2 with the Modified Bridge proposal was used to impute the missing data and the Gibbs sampler of Section 5.4 was used to infer the four drift parameters $a_i, i = 1 \ldots 4$.

increasing m. Figure 8.14b shows that the autocorrelation functions also improve with increasing m. Also, compared to the diffusion model in Eq. (8.8) the cubic model Eq. (8.10) better reproduces the autocorrelation function of the full system. A general, empirical model like Eq. (8.10) would also be more useful when there are several parameters entering into the functions in Eq. (8.9), due to their being



(a) Stationary distributions



(b) Autocorrelation functions

Figure 8.14: Predictive statistics for the mean flow using the inferred cubic model Eq. (8.10) for varying missing data m. Top: stationary distributions for various amounts of imputed data. The histogram is that of the full system. Bottom: autocorrelation plots of the inferred model compared to the full system (vertical bars).

several topographic modes. In some cases these parameters may not be identifiable.

In conclusion, for this section, the principled method of homogenisation followed by parameter estimation works well when the problem is designed to include only a few parameters. However, in a real atmospheric system it is likely that there would be a few parameters per unresolved mode of the system and, therefore, a lot of unknown parameters to be estimated in the reduced system. This was not a problem for Majda et al. [2003] and Franzke et al. [2008] as they estimated models

from observations of the full data. Here we have attempted to infer a reduced model from observations of the resolved modes only as this is more likely the case when considering real atmospheric data.

Chapter 9

Conclusions

The aim of this thesis was to apply an empirical modelling strategy to produce low dimensional diffusion models as approximations to high dimensional deterministic systems. We focused on models relating to atmospheric dynamics with the hope that the methods developed here might be useful to researchers developing stochastic models of large scale planetary variables such as the North Atlantic Oscillation. Stochastic modelling of the atmosphere could also be used to model longer time scale coupled atmosphere-ocean dynamics such as the El-Nino Southern Oscillation.

We compared the empirical approach with an analytical that assumes a significant separation of time scales between the variables we are interested in and those we want to ignore. We assumed that the empirical approach would be less affected when there was incomplete time scale separation and we see that this was the case for the chaotic Lorenz system in Section 8.1. In that section we also demonstrated that the models could be improved by including a latent noise process. This seems like a good direction for further study. Often in real systems, at short time scales where the non-linearities are significant, a diffusion model is not a good one: the paths are too smooth. Approximating the "weather" variables by an integrated stochastic process is a useful way of producing smooth functions and therefore models that will work at short time scales. At time scales at which a diffusion model works well for real climate data, the dynamics often seem linear. To model the short term non-linearities a smooth process is more suitable than a diffusion model. However, the statistical inference for such a process is much more difficult if one is imputing missing data between observations. In Section 8.1 we developed a method to infer a linear latent noise process which meant that the whole latent process could be sampled from the conditional posterior distribution. This Gibbs sampler enabled the inference to complete in a reasonable time. However, imputing non-linear la-

tent processes does not seem practical due to the enormous data sets needed and the difficulty in finding a good proposal distribution. Sampling the smooth, finite variation paths of the observed process is difficult and in fact it seems that the algorithm used in this thesis would degenerate in the continuous time limit as the chance of accepting a proposed path goes to zero. There are less rigorous ways of performing inference for such a model. For example, the Automatic Bayesian Computation (ABC) method, which is a simulation based approach that uses an approximation to the likelihood function. It would be useful to do further research in this area to develop theoretically well motivated algorithms that are based on the true likelihood.

In Chapter 5 we strove to improve the computational efficiency of inferring parameters for a general cubic SDE with linear diffusion function. We focussed upon designing more efficient proposal distributions for missing data. We developed the linear bridge proposal distribution and demonstrated that it is a useful improvement over the standard Modified Bridge in terms of MCMC efficiency although it is computationally expensive and difficult to implement. In general the algorithm of Golightly and Wilkinson [2008], that much of our work was based on, is much too inefficient in high dimensions. In practice it is not advisable to update all components of the missing data simultaneously but rather to sample component-wise. The method of updating one variable at a time, using an efficient proposal like the linear bridge, that we demonstrated in Section 5.2 is an improvement, although it also becomes inefficient as the dimension increases and is computationally demanding.

Most of the computations in this thesis took a long time to run. On the order of days at least for the cubic models. In many cases the predictive skill of the resulting model does not justify the amount of effort taken to obtain it. In some cases a linear model might be just as good and much easier to infer. Although the Gibbs sampling method of Section 5.4 makes sampling the drift parameters easier, we still have to use algorithms based on a Random Walk for the diffusion parameters. This is not very effective, especially considering that all of the missing data is also updated in this step when using the Golightly and Wilkinson [2008] algorithm. This means that only a small Random Walk step size is likely to be accepted and the mixing of the algorithm is poor. It would be worth considering how this proposal distribution could be improved but it is unlikely to solve the problem completely. There will remain the problem of having to update a large number of random variables simultaneously when using this algorithm and so this does not really solve the problem.

Although we have explored a general framework for inferring cubic SDEs,

these models are very complex and perhaps too general resulting in an unwieldy number of parameters to infer. In Section 8.1 we have seen that it can be useful to use the model predicted by homogenisation theory where there are not many parameters to infer. This approach, along with including a red noise model for short time scales, is probably the best method. One negative of this is that the parameters do not enter into a form that is amenable to Gibbs sampling and in some cases there can also be far too many parameters. This happens when there is at least one parameter associated with every "weather" variable as was the case for the quasi-geostrophic model in Section 8.3. It may be worth studying methods that make a further approximation to reduce the number of parameters in these models. As can be seen in the derivation in Section 3.3, in the final diffusion model Eq. (3.26), there are sums which range over all of the weather variables in the original system. These sums could be approximated by a single random variable, perhaps through the existence of a central limit theorem.

In early stages of this research we experienced a lot of problems with producing stable models that could be used for prediction: when simulating an ensemble from the estimated posterior distribution some solutions would explode to infinity leaving only a small number that could be used for producing average autocorrelation functions or other measures of predictive skill. One way of overcoming this would be to simply simulate lots of times with parameters sampled from the posterior and discard those that are unstable. The problem with this is that the size of the parameter space is very large and, as the dimension of the system increases, only a small subset of this space corresponds to diffusion models with stationary probability measures. In Chapter 7 we attempted to restrict the parameter space using a condition of negative definiteness of a matrix whose components are the parameters governing the magnitude of the cubic terms. This equates to enforcing a non-increase of energy associated with these terms. This was successful in producing SDEs with stable solutions. However, it is likely that this restricts the parameter space more than is necessary and future work would benefit from a more thorough study of this problem.

The negative definiteness constraint on the cubic parameters creates a novel state space from which to sample. In particular the Gibbs sampler is no longer applicable to these terms. Instead, in Section 7.2 we constructed MCMC algorithms to sample densities constructed on spaces of negative definite matrices. We attempted a component wise approach which is guaranteed to work but is complicated to implement. We then attempted to use the central and non-central Wishart distributions in a novel MCMC algorithm. We studied the efficiency of the central Wishart al-

gorithm and showed that it has similar properties to standard random walk type algorithms; the maximum efficiency occurs at the same optimal acceptance rate of 0.234. Further work could investigate the efficiency properties of the non-central Wishart proposal. It may be possible, through suitable choice of parameters, to adapt this algorithm to be equivalent to a second order algorithm similar to the Langevin algorithm, which involves the derivative of the target density.

In Section 8.2 we applied both the theoretical and empirical approach to model reduction for the Burgers system coupled to a triad model. We found that in this case the theoretical approach works well, although the set up is slightly artificial because the different time scales in the system are controlled. The empirical method reproduces well the autocorrelation functions for the case without significant time scale separation, though it did not produce good predictions of stationary probability distributions. It took a lot of computational time to infer these models. It seems more efficient to first construct a model with the right form through assuming different time scales, then estimate parameters. This even worked on the more difficult problem of the quasi-geostrophic model on the β-plane in Section 8.3, where time scales were not explicit although the empirical method was more skillful in this case. Improvements could be made to the theoretical method by including a red noise process as in Section 8.1. This still leaves the problem of the number of parameters increasing with the number of topographic modes in the system.

In summary, the conclusions of the thesis is that the homogenisation procedure, although not rigorous, is a good method of producing a model that has the right structure. The resulting parameters can then be estimated using the inference algorithm in this thesis, including the improved proposal methods for missing data of Chapter 5. Stability issues can be investigated on a per model basis and similar restrictions to the parameter space, such as those studied for cubic models in Chapter 7, can be implemented. The problem of lack of time scale separation can sometimes be overcome by introducing a latent red noise process, and this can be inferred from data using the method in Section 6.1. The problem remains of how to stop the number of parameters in the reduced system from scaling with the size of the original system. This is a topic for further research.

Appendix A

Example code for Empirical Climate Modelling

Note that the codes here are extracts from the various programs written in the thesis. As it is, it will not compile. Further code can be obtained by email. All code relies upon a matrix class and function library, written by myself, along with the GNU Scientific Library.

A.1 Main Program

This code shows the important parts of the input/output of the program and includes the main MCMC loop.

```
/* Program to infer parameters for cubic model for
discretely observed diffusions.
Drift parameters are sampled using
Gibbs sampler, diffusion parameters via the Innovation scheme of
Chib Et. al 04 + Golightly Wilkinson 08 and Missing data is imputed
using the modified bridge sampler of Durham Gallant 02.

Two input arguments: the first is the random number generator seed,
the second is the amount of missing data per observation interval.
*/
int main (int argc, const char* argv[]){
int seed = atoi(*(argv+1));
const gsl_rng_type * R;
gsl_rng * r;
```

```
gsl_rng_env_setup();
R = gsl_rng_mt19937;
r = gsl_rng_alloc(R);
gsl_rng_set (r,seed);

gsl_error_handler_t *err_off = gsl_set_error_handler_off ();

int mi = atoi(argv[2]);
int tin = atoi(argv[2]);
const char *I = argv[2];
const char *J = argv[1];
/* ms is an array of options for missing data.
The second input argument (1-10) indexes this array.
This of course can be expanded/modified.*/
int ms[10] = {1,2,4,8,16,32,64,128,256,512};
// m is the amount of missing data per interval.
// For example, if m=1 then there is no missing data.
int m = ms[mi-1];
double Ts[3] = {1,10,100};
double T = Ts[tin-1];
T = 1000;
// N is the number of observations
// n_samples is the number of MCMC samples
int i,j,k,l,s,n_samples=1e5,n_burnin=1e5,n_trial=1e5,n_iter=10,n_test=1e5;
double accep;
// X is the data and t is the observation times
double **X,*t;
// For more flexibility, M is an array containing a possibly differing amount
of missing data for each interval. This can be read in from a file
int *M;

double post1,post2,alpha,like,v;
double Mu[D],Mn[D];
// Dt is the standard observation interval
double Dt=2e-1;
int Dtint = 5;
int N = (int)T*Dtint;
```

```
//if a common m is used dt is the interval between (missing) data
double dt = Dt/(double)m;
// vector of parameters
// num is the number of paramters defined in global.h
double params[num];
double **Params;
double accep_rate;
bool b;

// Covs is the covariance matrix for MH algorithm to
update diffusion parameters.
// It is important to base this on a trial run of the posterior
// to improve efficiency
// nums is the number of diffusion parameters
myMatrix Covs(nums,nums);
// Prior mean and precision matrices for the drift parameters.
// There are P drift parameters per component and therefore D*P altogether
myVector prior_mean(D*P);
myMatrix prior_prec(D*P,D*P);
prior_prec.setIdentity();
prior_prec = 1e-2*prior_prec;

myMatrix Sig(D,D);
myMatrix Sig2(D,D);

// set root directory for file in/out
string root_dir = "/home/audrey/Phd/SDEs/triads/simulate/";

// data in
string file_in = root_dir;
file_in += "solution_sparse_0.8.dat";
ifstream data(file_in.c_str());

// covariance matrix for updating diffusion parameters
file_in = root_dir;
file_in += "mcmcvars.dat";
ifstream vars(file_in.c_str());
```

```
// starting parameter values
file_in = root_dir;
file_in += "param_means.dat";
ifstream inits(file_in.c_str());

// used if a variable observation time
file_in = root_dir;
file_in += "observation_times.dat";
ifstream times(file_in.c_str());

// used if different amount of missing data per observation
file_in = root_dir;
file_in += "missing.dat";
ifstream missing(file_in.c_str());

// MCMC samples output
string file_out = root_dir;
file_out += "samples_0.8_";
file_out += I;
file_out += "_";
file_out += J;
file_out += ".dat";
ofstream samples(file_out.c_str());

// optionally include output for missing data
string file_out2 = root_dir;
file_out2 += "process_1e-1_";
file_out2 += I;
file_out2 += "_";
file_out2 += J;
file_out2 += ".dat";
ofstream process(file_out2.c_str());

// allocate memory for arrays
X = new double*[(N-1)*m+1];
t = new double[(N-1)*m+1];
```

```
M = new int[N];

for(i=0;i<(N-1)*m+1;i++){
X[i] = new double[D];
}
// using equal m and dt set from argv[2]
// M[i] is the cumulative amount of data up to observation i
M[0]=0;
for(i=0;i<N-1;i++){
  M[i+1] = M[i]+m;
}
for(i=0;i<(N-1)*m+1;i++){
  t[i] = dt*i;
}

// using variable M and dt set from input files
/*for(i=0;i<(N-1)*m+1;i++){
  times >> t[i];
  }

for(i=0;i<N;i++){
  missing >> M[i];
}*/

// input data
k = 0;
for(i=0;i<M[N-1]+1;){
  for(j=0;j<D;j++){
            data >> X[i][j];
  }
  i += m;
}

// input covariance matrix for diffusion parameter sampling
for(i=0;i<nums;i++){
  for(j=0;j<nums;j++){
    vars >> v;
```

```
    Covs(i,j) = v;
  }
}
// or set to the identity
Covs.setIdentity();
Covs = Covs*1e-5/(double)(m*N);

Params = new double*[n_trial];
for(i=0;i<n_trial;i++){
  Params[i] = new double[nums];
}

// input initial parameter values
for(i=0;i<num;i++){
  inits >> params[i];
}

/*for(i=0;i<D*P;i++){
  params[i] = gsl_ran_gaussian(r,1);
}
for(i=0;i<(D*(D+1))/2;i++){
  params[i+D*P] = gsl_ran_gamma(r,1,1);
}
for(i=D*P+(D*(D+1))/2;i<(D*D*(D+1))/2;i++){
  params[i+D*P+(D*(D+1))/2] = gsl_ran_gaussian(r,1);
}*/

for(k=0;k<num;k++){
  samples << params[k] << ' ';
}
samples << endl;

// initialise(params,X,t,N,samples,r);

// initialise missing data between observations
for(i=0;i<N-1;i++){
  for(j=M[i];j<M[i+1]-1;j++){
```

```
      // set drift and diffusion functions
      set_coefs(*(X+j),params,Mu,Sig);
      // construct Durham Gallant 02 bridge distribution
      conditioned_prob_simple(*(X+j),*(X+M[i+1]),
      t[j],t[j+1],t[M[i+1]],Mn,Sig,Sig2);

      Sig2.choleskyDecomp();
      // sample missing data
      multivariate_normal(*(X+j+1),D,r,Mn,Sig2.getMat());
  }
}

// burn in
accep = 0;
j = 0;
for(i=0;i<n_burnin;i++){

            // update diffusion parameters
            sample_diffusion_parameters(X,t,params,N-1,M,Covs,&accep,r);

            // update all missing data
    sample_missing_data(X,t,params,N-1,M,r);

            // update drift parameters
            sample_drift_parameters(params,X,t,prior_mean,prior_prec,M[N-1]+1,r);
    j += 1;
    if(i%1000==0){
accep_rate = accep/(double)j;
                if(accep_rate>0.3){
                    Covs = 2*Covs;
}
                if(accep_rate<0.1){
                    Covs = 0.5*Covs;
                }

                accep = 0;
                j = 0;
```

```
                cout << accep_rate << endl;
            }
            for(k=0;k<num;k++){
                samples << params[k] << ' ';
            }
            samples << endl;
        }

        // estimate covariance of diffusion parameters by trial run
        for(i=0;i<n_trial;i++){

        }
        // main loop
        for(i=0;i<n_samples;i++){

            // update diffusion parameters
            sample_diffusion_parameters(X,t,params,N-1,M,Covs,&accep,r);

            // update all missing data
    sample_missing_data(X,t,params,N-1,M,r);

            // update drift parameters
            sample_drift_parameters(params,X,t,prior_mean,prior_prec,M[N-1]+1,r);

            // output parameters to file
            for(j=0;j<num;j++){
                samples << params[j] << ' ';
            }
            samples << endl;
        }

        return 0;
}
```

A.2 Sample Missing Data

This section includes some example code for updating missing paths between observations.

```
/* Update of missing data. Simplest method, which assumes using
modified bridge of Durham-Gallant and updating one block at a time
*/

void sample_missing_data(double **X, double *t, double *params, int N,
int *m, gsl_rng *r){
  int i,j,k;
  double Mu[D],Mn[D],post1,post2,prop1,prop2,alpha;

  // allocate space for proposal data
  double **Y;
  Y = new double*[m[N]+1];
  for(i=0;i<m[N]+1;i++){
      Y[i] = new double[D];
  }

  myMatrix Sig(D,D),Sig2(D,D);

  for(i=0;i<N;i++){
    for(j=0;j<D;j++){
      Y[m[i]][j] = X[m[i]][j];
      Y[m[i+1]][j] = X[m[i+1]][j];
    }
    prop1 = 0;
    prop2 = 0;
    for(j=m[i];j<m[i+1]-1;j++){
      // set drift and diffusion functions for proposal
      set_coefs(*(Y+j),params,Mu,Sig);
      // Gaussian density conditioned on the end points
      conditioned_prob_simple(*(Y+j),*(Y+m[i+1]),t[j],t[j+1],
t[m[i+1]],Mn,Sig,Sig2);
```

```
      Sig2.choleskyDecomp();
      // propose new data
      multivariate_normal(*(Y+j+1),D,r,Mn,Sig2.getMat());
      // calculate forward proposal density
      prop2 +=  multivariate_normal_pdf(*(Y+j+1),D,Mn,Sig2.getMat());
      // set drift and diffusion functions for data and bridge density
      set_coefs(*(X+j),params,Mu,Sig);
      conditioned_prob_simple(*(X+j),*(X+m[i+1]),t[j],t[j+1],
t[m[i+1]],Mn,Sig,Sig2);

      Sig2.choleskyDecomp();
      // calculate backward proposal density
      prop1 +=  multivariate_normal_pdf(*(X+j+1),D,Mn,Sig2.getMat());
    }
    // calculate posteriors
    post1 = likelihood(X+m[i],params,t+m[i],m[i+1]-m[i]) - prop1;
    post2 = likelihood(Y+m[i],params,t+m[i],m[i+1]-m[i]) - prop2;
    alpha = post2 - post1;

    // metropolis-hastings accept or reject of new data
    if(alpha>log(gsl_rng_uniform(r))){
      for(j=m[i];j<m[i+1];j++){
for(k=0;k<D;k++){
  X[j][k] = Y[j][k];
}
      }
    }

  }

  for(i=0;i<m[N];i++){
     delete[] Y[i];
  }
  delete[] Y;
}
```

A.3 Sampling Positive Definite Matrices

This section includes code for sampling positive definite matrices. The following uses the non-central Wishart distribution, which uses code for computing the hypergeometric function with a matrix argument.

```
 double noncentralWishart(myMatrix& S, myMatrix& Mu,
  myMatrix& Sigma, myMatrix& Theta){
int d = S.x();
double r,val,*p,q[1],E[d];
myMatrix Omega(d,d),iSigma(d,d),iTheta(d,d),Mut(d,d),A(d,d),B(d,d),Sc(d,d);
iSigma=Sigma;
iSigma.choleskyDecomp();
iSigma.choleskyInvert();
iTheta=Theta;
iTheta.choleskyDecomp();
iTheta.choleskyInvert();
Mut = Mu;
Mut.trans();

Omega = iSigma*Mut*iTheta*Mu;

A = -0.5*iSigma*S-0.5*Omega;
B = 0.25*Omega*iSigma*S;

B.calcEigs();
q[0] = d/2.0;

r = matrixHyper(20,2.0,p,0,q,1,B.getEigs(),d,NULL);

Sc = S;
Sc.choleskyDecomp();

val = A.calcTrace()+log(r)-0.5*log(Sc.calcDet());

return val;
}
```

Bibliography

Y Abe, S Ayik, PG Reinhard, and E Suraud. On stochastic approaches of nuclear dynamics. *Physics Reports - Review Section of Physics Letters*, 275(2-3):49–196, OCT 1996.

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

Yacine Ait-Sahalia. Closed-form likelihood expansions for multivariate diffusions. *Annals of Statistics*, 36(2):906–937, APR 2008.

Arnold. *Hasselmann's program revisted: the analysis of stochasticity in deterministic climate models*, chapter 2, pages 141–157. Birkhauser, 2001.

Matyas Barczy and Peter Kern. Representations of multidimensional linear process bridges. *arXiv:1011.0067v1 [math.PR]*, 2010.

Anthony G. Barnston and Robert E. Livezey. Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review*, 115:1083–1126, 1987.

J Berner. Linking nonlinearity and non-Gaussianity of planetary wave behavior by the Fokker-Planck equation. *Journal of the Atmospheric Sciences*, 62(7, Part 1): 2098–2117, JUL 2005.

A Beskos and GO Roberts. Exact simulation of diffusions. *Annals of Applied Probability*, 15(4):2422–2444, NOV 2005.

Alexandros Beskos, Omiris Papaspiliopoulos, Gareth O. Roberts, and Paul Fearnhead. Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes. *Journal of the Royal Statistical Society B*, 68:333–382, 2006.

Alexandros Beskos, Omiros Papaspiliopoulos, and Gareth O. Roberts. Retrospective exact simulation of diffusion sample paths with applications. *Bernoulli*, 12(6): 1077–1098, DEC 2006.

Alexandros Beskos, Orniros Papaspiliopoulos, and Gareth O. Roberts. A factorisation of diffusion measure and finite sample path constructions. *Methodology and Computing in Applied Probability*, 10(1):85–104, MAR 2008.

O Beskos and G Roberts. Retrospective markov chain monte carlo methods for dirichlet process hierarchical model. arXiv:0710.4228.

B. M. Bibby and M. Sorensen. Martingale estimation functions for discretely observed diffusion processes. *Bernoulli*, 1:17–39, 1995.

B. M. Bibby and M. Sorensen. On estimation for discretely observed diffusions: A review. *Theory of Stochastic Processes*, 2:49–56, 1996.

BM Bibby and M Sorensen. Simplified estimating functions for diffusion models with a high-dimensional parameter. *Scandinavian Journal of Statistics*, 28(1): 99–112, MAR 2001.

Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The Journal of Political Economy*, 81:637–654, 1973.

L Broze, O Scaillet, and JM Zakoian. Quasi-indirect inference for diffusion processes. *Econometric Theory*, 14(2):161–186, APR 1998.

Vincenzo Capasso and Daniela Morale. Stochastic modelling of tumour-induced angiogenesis. *Journal of Mathematical Biology*, 58(1-2):219–233, JAN 2009.

Jule G. Charney and John G. De Vore. Multiple flow equilibria in the atmosphere and blocking. *Journal of the Atmospheric Sciences*, 36:1205–1216, 1979.

S Chaumont, P Imkeller, and M Muller. Equilibrium trading of climate and weather risk and numerical simulation in a Markovian framework. *Stochastic Environmental Research and Risk Assessment*, 20(3):184–205, APR 2006.

WY Chen and S Bokka. Stochastic modeling of nonlinear epidemiology. *Journal of Theoretical Biology*, 234(4):455–470, JUN 21 2005.

S Chib, M Pitt, and N. Shephard. Likelihood based inference for diffusion driven state space models. *Working Paper, Nuffield College, Oxford University*, 2004.

JC Cox, JE Ingersoll, and SA Ross. An intertemporal general equilibrium-model of asset prices. *Econometrica*, 53(2):363–384, 1985.

Daan Crommelin and Eric Vanden-Eijnden. Reconstruction of diffusions using spectral data from timeseries. *Communications in Mathematical Sciences*, 4(3):651–668, SEP 2006.

DT Crommelin and AJ Majda. Strategies for model reduction: Comparing different optimal bases. *Journal of the Atmospheric Sciences*, 61(17):2206–2217, SEP 2004.

D. Dacunha-Castelle and D. Florens-Zmirou. Estimation of the coefficients of a diffusion from discrete observations. *Stochastics*, 19(4):263–284, 1986.

Fabio D'andrea. Extratropical low-frequency variability as a low-dimensional problem. ii: Stationarity and stability of large-scale equilibria. *Quarterly Journal of the Royal Meteorological Society*, 128(582):1059–1073, 2002.

Fabio D'Andrea and Robert Vautard. Extratropical low-frequency variability as a low-dimensional problem i: A simplified model. *Quarterly Journal of the Royal Meteorological Society*, 127(574):1357–1374, 2001.

Christiane Dargatz. *Bayesian Inference for Diffusion Processes with Applications in Life Sciences*. PhD thesis, Mathematik, Informatik und Statistik der Ludwig-MaximiliansUniversitt Mnchen, 2010.

T DelSole. Optimally persistent patterns in time-varying fields. *Journal of the Atmospheric Sciences*, 58(11):1341–1356, 2001.

PD Ditlevsen. Observation of alpha-stable noise induced millennial climate changes from an ice-core record. *Geophysical Research Letters*, 26(10):1441–1444, MAY 15 1999.

Randall M. Dole and Neil D. Gordan. Persistent anomalies of the extratropical northern hemisphere wintertime circulation: Geographical distribution and regional persistence characteristics. *Monthly Weather Review*, 11:1567–1586, 1983.

GB Durham and AR Gallant. Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. *Journal of Business and Economic Statistics*, 20(3):297–316, JUL 2002.

Morris L. Eaton. *Multivariate Statistics: A Vector Space Approach. Lecture Notes–Monograph Series, Volume 53*. Institute of Mathematical Statistics, 2007.

O. S. Elerian. *Simulation estimation of continuous-time models with applications to finance.* PhD thesis, Nuffield College, Oxford, 1999.

Ola Elerian. A note on the existence of a closed form conditional transition density for the milstein scheme. *Working Paper, Nuffield College, Oxford University*, 98.

Ola Elerian, Siddhartha Chib, and Neil Shephard. Likelihood inference for discretely observed nonlinear diffusions. *Econometrica*, 69(4):959–93, July 2001.

Bjorn Eraker. MCMC analysis of diffusion models with application to finance. *Journal of Business and Economic Statistics*, 19:177–191, 2001.

Paul Fearnhead. The stationary distribution of allele frequencies when selection acts at unlinked loci. *Theoretical Population Biology*, 70(3):376–386, NOV 2006.

SE Feller, YH Zhang, RW Pastor, and BR Brooks. Constant-pressure molecular-dynamics simulation - the Langevin piston method. *Journal of Chemical Physics*, 103(11):4613–4621, SEP 15 1995.

Danielle Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20(4):547–557, 1989.

GW Ford, JT Lewis, and RF Oconnell. Quantum Langevin equation. *Physical Review A*, 37(11):4419–4428, JUN 1 1988.

Claude Frankignoul and Klaus Hasselmann. Stochastic climate models. part2: Application to sea-surface temperature anomalies and thermocline variability. *Tellus*, 29:289–305, 1977.

Christian Franzke, Andrew J. Majda, and Eric Vanden-Eijnden. Low-order stochastic mode reduction for a realistic barotropic model climate. *Journal of the Atmospheric Sciences*, 62:1722–1745, 2005.

Christian Franzke, Daan Crommelin, Alexander Fischer, and Andrew Majda. A Hidden Markov Model perspective on regimes and metastability in atmospheric flows. *Journal of Climate*, 21:1740–1757, 2008.

Christian Franzke, Illia Horenko, Andrew J. Majda, and Rupert Klein. Systematic Metastable Atmospheric Regime Identification in an AGCM. *Journal of the Atmospheric Sciences*, 66(7):1997–2012, JUL 2009.

AR Gallant and G Tauchen. Which moments to match? *Econometric Theory*, 12 (4):657–681, OCT 1996.

C. W. Gardiner. *Handbook of Stochastic Methods*. Springer, 2004.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

W.R. Gilks and DJ Spiegelhalter. *Markov chain Monte Carlo in practice*. Chapman & Hall/CRC, 1996.

DT Gillespie. The chemical Langevin equation. *Journal of Chemical Physics*, 113 (1):297–306, JUL 1 2000.

A. Golightly and D. J. Wilkinson. Bayesian inference for stochastic kinetic models using a diffusion approximation. *Biometrics*, 61:781–788, 2005.

A Golightly and DJ Wilkinson. Bayesian sequential inference for stochastic kinetic biochemical network models. *Journal of Computational Biology*, 13(3):838–851, APR 2006a.

A. Golightly and D.J. Wilkinson. Bayesian inference for nonlinear multivariate diffusion models observed with error. *Computational Statistics & Data Analysis*, 52:1674 – 1693, 2008.

Andrew Golightly and Darren J. Wilkinson. Bayesian sequential inference for nonlinear multivariate diffusions. *Statistics and Computing*, 16(4):323–338, DEC 2006b.

Dan Gordon, Vikram Krishnamurthy, and Shin-Ho Chung. Generalized Langevin models of molecular dynamics simulations with applications to ion channels. *Journal of Chemical Physics*, 131(13), OCT 7 2009.

C. Gourieroux, A. Montfort, and E. Renault. Indirect inference. *Journal of Applied Econometrics*, 8(S):S85–S118, DEC 1993.

L. P. Hansen. Large sampler properties of generalized-method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

L. P. Hansen and J. A. Scheinkman. Back to the future - generating moment implication for continuous time markov processes. *Econometrica*, 63(4):767–804, JUL 1995.

K. Hasselmann. Stochastic climate models. part 1: Theory. *Tellus*, 28:473–484, 1976.

Rainer Hegger and Gerhard Stock. Multidimensional Langevin modeling of biomolecular dynamics. *Journal of Chemical Physics*, 130(3), JAN 21 2009.

Illia Horenko. On Robust Estimation of Low-Frequency Variability Trends in Discrete Markovian Sequences of Atmospheric Circulation Patterns. *Journal of the Atmospheric Sciences*, 66(7):2059–2072, JUL 2009.

Illia Horenko, Stamen I. Dolaptchiev, Alexey V. Eliseev, Igor I. Mokhov, and Rupert Klein. Metastable Decomposition of High-Dimensional Meteorological Data with Gaps. *Journal of the Atmospheric Sciences*, 65(11):3479–3496, NOV 2008.

A. S. Hurn, J. I. Jeisman, and K. A. Lindsay. Seeing the wood for the trees: A critical evaluation of methods to estimate the parameters of stochastic differential equations. *Journal of Financial Econometrics*, 5(3):390–455, SUM 2007.

AS Hurn and KA Lindsay. Estimating the parameters of stochastic differential equations. *Mathematics and Computers in Simulation*, 48(4-6):373–384, JUN 1999.

AS Hurn, KA Lindsay, and VL Martin. On the efficacy of simulated maximum likelihood for estimating the parameters of stochastic differential equations. *Journal of Time Series Analysis*, 24(1):45–63, JAN 2003.

Bjarke Jensen and Rolf Poulsen. Transtion densities of diffusion processes: numerical comparison of approximation techniques. *Journal of Derivatives*, 9:18–32, 2002.

Scot D. Johnson, David S. Battisti, and E. S. Sarachik. Empirically derived markov models and prediction of tropical pacic sea surface temperature anomalies. *Journal of Climate*, 13:3–17, 2000.

Konstantinos Kalogeropoulos, Gareth O. Roberts, and Petros Dellaportas. Inference for stochastic volatility models using time change transformations. *Annals of Statistics*, 38(2):784–807, APR 2010.

Konstantinos Kalogeropoulos, Petros Dellaportas, and Gareth O. Roberts. Likelihood-based inference for correlated diffusions. *Canadian Journal of Statistics*, 39(1):52–72, MAR 2011.

I. Karatzas and S.E. Shreve. *Methods of Mathematical Finance*. Springer-Verlag, 1997.

M Kessler. Simple and explicit estimating functions for a discretely observed diffusion process. *Scandinavian Journal of Statistics*, 27(1):65–82, MAR 2000.

R.Z. Khasminskii. A limit theorem for solutions of differential equations with random right-hand side. *Theory of Probability and its Applications*, 11:390–406, 1966.

Y. Kifer. Averaging and climate models.

Yuri Kifer. Limit theorems in averaging for dynamical systems. *Ergodic Theory and Dynamical Systems*, 15:1143–1172, 1995.

Yuri Kifer. *Averaging and climate models*, chapter 2.3, pages 171–188. Birkhauser, 2001.

Masahide Kimoto and Michael Ghil. Multiple flow regimes in the northern hemisphere winter. part 1: methodology and hemispheric regimes. *Journal of the Atmospheric Sciences*, 50:2625–2643, 1993.

Peter E. Kloeden and Eckhard Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, 1992.

P. Koev and A. Edelman. The efficient evaluation of the hypergeometric function of a matrix argument, 2006.

S. Kravtsov, D. Kondrashov, and M. Ghil. Multilevel regression modelling of nonlinear processes: derivation and application to climate variability. *Journal of Climate*, 18:4404–4424, 2005.

F Kwasniok. The reduction of complex dynamical systems using principal interaction patterns. *Physica D*, 92(1-2):28–60, APR 15 1996.

F Kwasniok. Optimal Galerkin approximations of partial differential equations using principal interaction patterns. *Physical Review E*, 55(5, Part a):5365–5375, MAY 1997.

F Kwasniok. Empirical low-order models of barotropic flow. *Journal of the Atmospheric Sciences*, 61(2):235–245, JAN 2004.

Anthony Lee, Christopher Yau, Michael B. Giles, Arnaud Doucet, and Christopher C. Holmes. On the utility of graphics cards to perform massively parallel simulation of advanced monte carlo methods. *Journal of Computational and Graphical Statistics*, 19(4):769–789, 2010.

AW LO. Maximum-likelihood estimation of generalized ito processes with discretely sampled data. *Econometric Theory*, 4(2):231–247, AUG 1988.

A. Majda, I. Timofeyev, and E. Vanden-Eijnden. A mathematical framework for stochastic climate models. *Communications on Pure and Applied Mathematics*, 54:891974, 2001.

A. Majda, I. Timofeyev, and E. Vanden-Eijnden. A priori tests of a stochastic mode reduction strategy. *Physica D: Nonlinear Phenomena*, 170(3-4):206 – 252, 2002.

Andrew J. Majda and Xiaoming Wang. *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*. Cambridge University Press, 2006.

Andrew J. Majda and Yuan Yuan. Fundamental limitations of ad hoc linear and quadratic multi-level regression models for physical systems. *Discrete and Continuous Dynamical Systems - Series B*, 17(4, SI):1333–1363, JUN 2012.

Andrew J. Majda, Ilya Timofeyev, and Eric Vanden Eijnden. Models for stochastic climate prediction. *Proceedings of the National Academy of Sciences*, 96:14687–14691, 1999.

Andrew J. Majda, Ilya Timofeyev, and Eric Vanden-Eijnden. Systematic strategies for stochastic mode reduction in climate. *Journal of the Atmospheric Sciences*, 60:1705–1721, 2003.

Andrew J. Majda, Christian Franzke, and Daan Crommelin. Normal forms for reduced stochastic climate models. *Proceedings of the National Academy of Sciences*, 106:36493653, 2009.

J Marshall and F Molteni. Toward a dynamical understanding of planetary-scale flow regimes. *Journal of the Atmospheric Sciences*, 50(12):1792–1818, JUN 15 1993.

XL Meng and D vanDyk. The EM algorithm - An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society Series B- Methodology*, 59(3): 511–540, 1997.

Lewis Mitchell and Georg A. Gottwald. Data Assimilation in Slow-Fast Systems Using Homogenized Climate Models. *Journal of the Atmospheric Sciences*, 69(4): 1359–1377, APR 2012.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley, 1982.

B. Øksendal and A. Sulem. *Applied Stochastic Control of Jump Diffusions*. Springer, 2007.

Bernt Øksendal. *Stochastic Differential Equations*. Springer, 2007.

L. S. F. Olavo, L. C. Lapas, and A. Figueiredo. Foundations of quantum mechanics: The Langevin equations for QM. *Annals of Physics*, 327(5):1391–1407, MAY 2012.

Keisuke Ota, Takamasa Tsunoda, Toshiaki Omori, Shigeo Watanabe, Hiroyoshi Miyakawa, Masato Okada, and Toru Aonishi. Is the Langevin phase equation an efficient model for oscillating neurons? 197, 2009.

T Ozaki. A bridge between nonlinear time-series models and nonlinear stochastic dynamic-systems - a local linearization approach. *Statistica Sinica*, 2(1):113–135, JAN 1992.

Lionel Pandolfo. Observational aspects of the low-frequency intraseasonal variability of the atmosphere in middle latitudes. 34:93 – 174, 1993.

Grigorios A. Pavliotis and Andrew M. Stuart. *Multiscale Methods: Averaging and Homogenization*. Springer, 2008.

A. R. Pedersen. A new approach to maximum-likelihood-estimation for stochastic differential-equations based on discrete observations. *Scandinavian Journal of Statistics*, 22(1):55–71, MAR 1995.

Joseph Pedlosky. *Geophysical Fluid Dynamics*. Springer, 1987.

Cecile Penland. A stochastic model of indopacific sea-surface temperature anomalies. *Physica D: Nonlinear Phenomena*, 98:534–558, 1996.

Cecile Penland and Michael Ghil. Forecasting northern hemisphere 700-mb geopotential height anomalies using empirical normal modes. *Monthly Weather Review*, 121:2355–2372, 1993.

Cecile Penland and Ludmila Matrosova. Prediction of tropical atlantic sea surface temperatures using linear inverse modeling. *Journal of Climate*, 11:483–496, 1998.

Roger Peyret. *Spectral Methods for Incompressible Viscous Flow*. Springer, 2002.

Y. Pokern. *Fitting stochastic differential equations to molecular dynamics data*. PhD thesis, University of Warwick, Coventry, 2007.

Yvo Pokern, Andrew M. Stuart, and Petter Wiberg. Parameter estimation for partially observed hypoelliptic diffusions. *Journal Of The Royal Statistical Society Series B*, 71(1):49–73, 2009.

N. G. Polson and G. O. Roberts. Bayes factors for discrete observations from diffusion processes. *Biometrika*, 81:11–26, 1994.

SB POPE. Lagrangian pdf methods for turbulent flows. *Annual Review of Fluid Mechanics*, 26:23–63, 1994.

B. L. S. Prakasa-Rao. Asymptotic theory for non-linear least square estimator for diffusion processes. *Math. Operationsforch. Statist. Ser. Stat*, 14:195–209, 1983.

R. W. Preisendorfer. *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 1988.

John D. Ramshaw. Augmented langevin approach to fluctuations in nonlinear irreversible processes. *Journal of Statistical Physcis*, 38:669–680, 1985.

Daniel Rex. Blocking action in the middle troposphere and its effect upon regional climate. *Tellus*, 2:196–211, 1950.

Christian Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125, 1995.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

G. O. Roberts and O. Stramer. On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. *Biometrika*, 88:603–621, 2001.

G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *Annals of Applied Probability*, 7:110–120, 1997.

G.O. Roberts and J. S. Rosenthal. General state space markov chains and mcmc algorithms. *Probability Surveys*, 1:20–71, 2004.

J. S. Rosenthal. Minorization conditions and convergence-rates for Markov-Chain Monte-Carlo. *Journal of the American Statistical Association*, 90(430):558–566, JUN 1995.

SK Sahu and GO Roberts. On convergence of the EM algorithm and the Gibbs sampler. *Statistics and Computing*, 9(1):55–64, JAN 1999.

F. M. Selten. An efficient description of the dynamics of barotropic flow. *Journal of the Atmospheric Sciences*, 52(7):915–936, APR 1 1995.

N Shephard and MK Pitt. Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667, SEP 1997.

I Shoji. Approximation of continuous time stochastic processes by a local linearization method. *Mathematics of Computation*, 67(221):287–298, JAN 1998.

I Shoji and T Ozaki. Estimation for nonlinear stochastic differential equations by a local linearization method. *Stochastic Analysis and Applications*, 16(4):733–752, 1998.

Isao Shoji and Tohru Ozaki. A statistical method of estimation and simulation for systems of stochastic differential equations. *Biometrika*, 85(1):240–243, 1998.

Paul Sjberg, Per Ltstedt, and Johan Elf. Fokker-planck approximation of the master equation in molecular biology. *Computing and Visualization in Science*, 12:37–50, 2005.

H Sorensen. Discretely observed diffusions: Approximation of the continuous-time score function. *Scandinavian Journal of Statistics*, 28(1):113–121, MAR 2001.

Halle Sorensen. Parametric inference for diffusion processes observed at discrete points in time: a survey. *International Statistical Reviews*, 72:337–354, 2004.

M Sorensen. *Selected Proceedings of the Symposium on Estimating Functions*, volume 32, Proceedings Paper Estimating functions for discretely observed diffusions: A review, pages 305–325. C/O B E TRUMBO, 3401 INVESTMENT BLVD 6, HAYWARD, CA 94545 USA, 1997. Symposium on Estimating Functions, UNIV GEORGIA, ATHENS, GA, MAR 21-23, 1996.

M. Sorensen. Prediction-based estimating functions. *Econometrics Journal*, 3:123–147, 2000.

Michael Sorensen. On asymptotics of estimating functions. *Brazilian Journal of Probability and Statistics*, 13:111–136, 1999.

Michael Sorensen. Prediction-based estimating functions: Review and new developments. *Brazilian Journal of Probability and Statistics*, 25(3):362–391, NOV 2011.

Thomas Stemler, Johannes P. Werner, Hartmut Benner, and Wolfram Just. Stochastic modeling of experimental chaotic time series. *Phys. Rev. Lett.*, 98:044102, Jan 2007.

Marc A. Suchard, Quanli Wang, Cliburn Chan, Jacob Frelinger, Andrew Cron, and Mike West. Understanding gpu programming for statistical computation: Studies in massively parallel massive mixtures. *Journal of Computational and Graphical Statistics*, 19(2):419–438, 2010.

P Sura. Stochastic analysis of Southern and Pacific Ocean sea surface winds. *Journal of the Atmospheric Sciences*, 60(4):654–666, FEB 2003.

P Sura and J Barsugli. A note on estimating drift and diffusion parameters from timeseries. *Physics Letters A*, 305(5):304–311, DEC 9 2002.

S. Suweis, A. Rinaldo, S. E. A. T. M. Van der Zee, E. Daly, A. Maritan, and A. Porporato. Stochastic modeling of soil salinity. *Geophysical Research Letters*, 37, APR 7 2010.

M. A. Tanner and H. W. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, JUN 1987.

Uffe H. Thygesen. A survey of lyapunov techniques for stochastic differential equations, 1997.

G. Uhlenbeck and L. Ornstein. On the theory of brownian motion. *Physical Review*, 36:823–841, 1930.

Koichi Unami, Felix Kofi Abagale, Macarius Yangyuoru, Abul Hasan M. Badiul Alam, and Gordana Kranjac-Berisavljevic. A stochastic differential equation model for assessing drought and flood risks. *Stochastic Environmental Research and Risk Assessment*, 24(5):725–733, JUL 2010.

N. van Kampen. The validity of nonlinear langevin equations. *Journal of Statistical Physics*, 25:431–442, 1981.

N.G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, 1997.

H Von Storch, G Burger, R Schnur, and JS Von Storch. Principal Oscillation Patterns - a review. *Journal of Climate*, 8(3):377–400, MAR 1995.

Hans von Storch and David P. Baumhefner. Principal oscillation pattern analysis of the tropical 30 to 60 day oscillation. *Climate Dynamics*, 6:1–12, 1991.

John M. Wallace and David S. Gutzler. Teleconnections in the geopotential height field during the northern hemisphere winter. *Monthly Weather Review*, 109:784–109, 1981.

N. Wiener, A. Siegel, B. Rankin, and W. T. Martin. *Differential Space, Quantum Systems and Prediction*. The M.I.T Press, 1966.

Daniel S. Wilks. Effects of stochastic parametrization on conceptual climate models. *Philosophical Transactions of the Royal Society A - Mathematical, Physical and Engineering Sciences*, 366(1875):2477–2490, JUL 28 2008.

Christopher R. Winkler, Matthew Newman, and Prashant D. Sardeshmukh. A linear model of wintertime low-frequency variability. part i: Formulation and forecast skill. *Journal of Climate*, 14:4474–4494, 2001.

Jin-Song Xu and Hans von Storch. Predictiing the state of the southern oscillation using principal oscillation pattern analysis. *Journal of Climate*, 3:1316–1329, 1990.