Bayesian inference and model selection for multi-dimensional diffusion process models with non-parametric drift and constant diffusivity

Tjun Yee Hoh

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

of

University College London.

Department of Statistical Science
University College London

Tuesday 25th June, 2019
I, Tjun Yee Hoh, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.
Abstract

For a multi-dimensional, partially observed diffusion process model with unknown drift and variable-independent diffusivity, we construct a composite methodology to perform Bayesian inference for the coefficients. Recent development of non-parametric Bayesian estimation of the drift has been restricted to dimension one, since the local time process is unavailable in the multi-dimensional case. We involve the empirical measure instead and show that the drift likelihood has a quadratic form, which allows a conjugate Gaussian measure prior whose precision operator is chosen to be a high order differential operator. We detail a computationally efficient pseudo-spectral method for solving the posterior mean, and describe how inference for the drift can be constrained to allow only conservative drifts.

We also adapt a Langevin MCMC approach to sampling from diffusion bridges as a data augmentation scheme. To sample from the diffusivity, we specify an Inverse Wishart prior and implement a random walk Metropolis-Hastings algorithm.

Evaluation of model fit for diffusion processes historically involved frequentist goodness-of-fit testing for fully parametric null models. We extend an existing transition density-based omnibus test to the null model case with non-parametric drift. We study the finite-sample behaviour of the test statistic and show that existing asymptotic results are inappropriate for settings involving real data. We implement the Bayesian discrepancy $p$-value to complement our inference methodology. With the goal of model improvement in mind, we describe how outlier removal and systematic sub-sampling of the data can be beneficial.
Impact Statement

The study of diffusion processes is becoming ever more widespread since the introduction of the Black-Scholes model. Over the past decade, there have been increasing applications of diffusion process models outside the usual subjects areas of mathematical finance, physics and molecular chemistry, since the drift and diffusivity coefficients of a diffusion process allow for a convenient interpretation of the behaviour of observed data.

However, these models are often restricted to the 1-dimensional case, and the construction of a suitable drift coefficient can be a challenge. In this thesis, we seek to provide a comprehensive Bayesian inference methodology for multi-dimensional diffusion process models where the drift is specified by a Gaussian measure, which allows for powerful data-driven estimation of the drift from a large class of continuous functions. This hands-off approach for the choice of drift can be beneficial for specialists and non-specialists alike by simplifying the task of experimental design. We also include as an alternative model a diffusion process with drift conditioned to be conservative as this choice of drift is also becoming increasingly common.

There is also a surprising lack of model fit evaluation in the literature for diffusion processes, which we hope to tackle by taking a holistic approach. In this thesis, we prescribe a Bayesian approach to goodness-of-fit hypothesis testing, with the goal of improving model fit through the use of sub-sampling and outlier detection. This can help guide researchers when considering diffusion process models as a tool for performing statistical analysis and should help pave the way for more wide-spread testing of diffusion process models.
Acknowledgements

First and foremost, I would like to express my sincere gratitude to Dr Yvo Pokern for his guidance throughout my journey at UCL ever since taking me in as a final year undergraduate project student. Without his encouragement and guidance I would never have made it this far.

I also give my thanks to Dr Ioanna Manolopoulou for her helpful insights on the applied side of my research. I am grateful to Dr Alexandros Beskos and Professor Richard Chandler for their constructive advice especially during the PhD upgrade process.

To the unending patience and love shown by my parents and sisters, I will be eternally grateful. Finally, thank you Carmen for being my pillar of support.
Contents

1 Introduction 20
   1.1 Introduction ............................................. 20
   1.2 Thesis Outline ........................................... 24

2 Preliminary Material 28
   2.1 Diffusion processes ..................................... 28
   2.2 Gaussian measures ....................................... 40
   2.3 Partial differential operators ............................ 41
   2.4 Krylov subspace methods ................................ 47
   2.5 Markov chain Monte Carlo methods ..................... 49

3 Inference for diffusion processes 51
   3.1 Introduction .............................................. 51
   3.2 Partially observed diffusion processes ................. 56
     3.2.1 Diffusion bridges .................................. 57
     3.2.2 Random walk Metropolis-Hastings ................ 58
     3.2.3 Alternative diffusion bridge samplers ............ 63
   3.3 Reparameterisation of diffusion processes ............ 67
     3.3.1 Reparameterisation .................................. 67
   3.4 Inference for constant diffusivity ...................... 72
     3.4.1 Diffusivity inference ................................ 72
   3.5 Non-parametric drift inference: dimension one ........ 75
     3.5.1 The likelihood and local time ....................... 76
## List of Figures

1.1 A sample path from a 1-dimensional Ornstein-Uhlenbeck diffusion process (in blue) and its deterministic counterpart (in red). ... 20

1.2 Left: Movement data of a capuchin monkey over the course of 46 non-consecutive days. Right: Movement data of the same capuchin monkey over the course of a day. The $*$ and $+$ symbols respectively denote the start and end-points of each trajectory ... 22

2.1 Sample paths of the 1-dimensional (left) and 2-dimensional (right) Wiener process. ... 29

2.2 A sample path from the 1-dimensional Ornstein-Uhlenbeck diffusion process, with a red line marking the “mean” term $\beta = 3$. ... 32

2.3 A visualisation of the drift $b$ (left) and a sample path (right) from the toy diffusion process. ... 34

2.4 Further visualisations of the drift $b$, where the $x$- and $y$-components of the drift are displayed as surface plots (a)-(c) with corresponding bird’s eye view plots (b)-(d). ... 34

2.5 A sample path from the 1-dimensional standard Brownian bridge process, for $t \in [0,1]$, with a red line marking the start and end-points $X_0 = 0$ and $X_1 = 0$. ... 35

3.1 Left: a simulated path from the toy diffusion process with the collection of red points preserved as a low frequency path $\{X_t\}_{t=0}^n$. Right: a path $X_{\text{aug}}$ obtained by augmenting the low frequency path $\{X_t\}_{t=0}^n$ with diffusion bridges. ... 62
### 3.2 Computational times of our implemented Bayesian inference components for the capuchin monkeys data in Section 3.9.1. These components involve the diffusion bridge sampler, estimation of the constant diffusivity and non-parametric estimation of the drift. ... 63

### 3.3 The $x$-component of samples from the mean zero Gaussian measure with precision operator $A_0 = \eta (-\Delta)^k + \varepsilon$, for fixed $\eta = 5 \times 10^{-5}$, $\varepsilon = 5 \times 10^{-3}$, and for varying $k = 1, 2, 3$ and 4 (a,b,c,d). ... 83

### 3.4 The $x$-component of samples from the mean zero Gaussian measure with precision operator $A_0 = \eta (-\Delta)^k + \varepsilon$, for fixed $k = 3$, $\varepsilon = 5 \times 10^{-3}$, and for varying $\eta = 5 \times 10^{-l}$, where $l = 3, 5, 7$ and 9 (a,b,c,d). ... 84

### 3.5 The $x$-component of samples from the mean zero Gaussian measure with precision operator $A_0 = \eta (-\Delta)^k + \varepsilon$, for fixed $k = 3$, $\eta = 5 \times 10^{-7}$ and for varying $\varepsilon = 5 \times 10^{-l}$, where $l = 1, 3, 5$ and 7 (a,b,c,d). ... 85

### 3.6 The full (left column) and $x$-component (right column) of the posterior mean of the drift, with hyper-parameter $\eta$ set to $\eta = 5 \times 10^{-l}$, for $l = 3, 5, 7$, and 9 (top to bottom), for fixed $\varepsilon = 5 \times 10^{-3}$. The colour bar denotes the posterior variance. ... 89

### 3.7 The full (left column) and $x$-component (right column) of the posterior mean of the drift, with hyper-parameter $\varepsilon$ set to $\varepsilon = 5 \times 10^{-l}$, for $l = 2, 3$ and 4 (top to bottom), for fixed $\varepsilon = 5 \times 10^{-7}$. The colour bar denotes the posterior variance. ... 90

### 3.8 The toy diffusion process with diffusivity as a multiple of the identity $\sigma = (1.4, 0; 0, 1.4)$ was used to generate the input trajectory used for drift estimation. We have here (a) the true drift, (b) an estimate of the unconditioned drift $\hat{b}_{\text{full}}$, (c) the projection of $\hat{b}_{\text{full}}$, and (d) an estimate of the drift conditioned to be conservative $\hat{b}_c$. ... 118
3.9 The toy diffusion process with non-symmetric diffusivity $\sigma = (1.4, -0.25; -0.25, 1.4)$ was used to generate the input trajectory used for drift estimation. We have here (a) the true drift, (b) an estimate of the unconditioned drift $\hat{b}_{\text{full}}$, (c) the projection of $\hat{b}_{\text{full}}$, and (d) an estimate of the drift conditioned to be conservative $\hat{b}_{c}$. . . . . 119

3.10 A sample path simulated from the toy diffusion process with sample size $n = 801$ with total time $T = 20$. . . . . . . . . . . . . . . . . . . 122

3.11 These are the results of the composite Bayesian inference methodology for the models where the drift is unconditioned (left column) and where the drift is conditioned to be conservative (right column). (a)-(b) The computational times of the individual Metropolis-within-Gibbs sampler components, which are respectively the diffusion bridge, diffusivity and drift samplers. (c)-(d) The posterior mean of the drift. (e)-(f) The posterior densities of the diffusivity. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 123

3.12 The trace plots (left) of the x- (top) and y-components (bottom) of the mid-point of diffusion bridges with 10, 100 and 1000 observed points. The autocorrelation plots (right) correspond to the diffusion bridge with the worst autocorrelation, which in this case is the one with 10 observed points. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 125

3.13 The trace plots (left) of the x- (top) and y-components (bottom) of the drift evaluated at $(5.7656, 5.7656)$, with traces for the different initial states. The autocorrelation plots (right) correspond to the trace with the worst autocorrelation. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 126

3.14 The x- (left) and y-components (right) of a sample from the prior (top) and posterior (bottom) of the drift. . . . . . . . . . . . . . . . 128
3.15 Comparison of the prior and posterior distributions for the x- (left) and y-components (right) of the drift using scatter plots (top) and QQ-plots (bottom) for one random point on the grid, and histograms (bottom) of p-values of KS2-tests for 200 randomly sampled points on the grid. 10000 samples were drawn from the prior for the KS2-tests, 500 of which were coupled with the 500 posterior samples to generate the scatter plots. ................................. 129

3.16 The trace plots (left) of the $\sigma_{11}$ (top), $\sigma_{12}$ (middle) and $\sigma_{22}$ components of the diffusivity, with traces for the different initial states. The autocorrelation plots (right) correspond to the trace with the worst autocorrelation. ........................................... 131

3.17 Prior (left) and posterior (right) densities of the diffusivity components $\sigma_{11}$, $\sigma_{12}$ and $\sigma_{22}$. .................................................. 132

3.18 Comparison of the prior and posterior distributions for the $\sigma_{11}$ (top), $\sigma_{12}$ (middle) and $\sigma_{22}$ components of the diffusivity using scatter plots (left) and QQ-plots (right). 10000 samples were drawn from the prior for the KS2-tests, 500 of which were coupled with the 500 posterior samples to generate the scatter plots. ...................... 133

3.19 Trace (left) and autocorrelation (right) plots for the middle point of a randomly selected diffusion bridge. ............................... 134

3.20 Trace (left) and autocorrelation (right) plots for the drift evaluated at a randomly selected point on the grid. ......................... 135

3.21 Trace (left) and autocorrelation (right) plots for the diffusivity. . . . 135
### 3.22 (a) Movement data of the capuchin monkey over the course of 46 non-consecutive days. (b) Movement data of the capuchin monkey over the course of a day. The * and + symbols respectively denote the start and end-points of each trajectory. (c) Movement data of the capuchin monkey with colour-coded data points to represent their time-indexing, with the start points of the trajectory in yellow and the end points in blue. (d) A histogram of the time differences between any two time-wise adjacent points.

### 3.23 (a) Movement data of the capuchin monkey. (b) Posterior mean of the drift. (c), (d): $x$ and $y$-components of the posterior mean. (e): Posterior densities of the components of the diffusivity. (f) Computational times of the Gibbs sampler components.

### 3.24 (a) Movement data of the capuchin monkey. (b) Posterior mean of the drift conditioned to be conservative. (c), (d): $x$ and $y$-components of the posterior mean. (e): Posterior densities of the components of the diffusivity. (f) Computational times of the Gibbs sampler components.

### 3.25 Trace (left) and autocorrelation (right) plots for the middle point of a randomly selected diffusion bridge.

### 3.26 Trace (left) and autocorrelation (right) plots for the drift evaluated at a randomly selected point on the grid.

### 3.27 Trace (left) and autocorrelation (right) plots for the diffusivity.

### 4.1 Finite sample mean (left) and variance (right) of the unscaled test statistic $\hat{Q}_V(1)$ versus uniform generalised residuals of size $n = 100 \times 2^l, l \in \{0, 1, \ldots, 25\}$. The uniform generalised residuals were simulated 10000 times for $l = 0, 1, \ldots, 19$, 1000 times for $l = 20$, 500 times for $l = 21, 22$ and 100 times for $l = 23, 24$ due to computational constraints. This is reflected in the size of their respective confidence intervals.
4.2 Q-Q plots comparing the distribution of \( \hat{Q}_C(1) \) against the standard normal distribution for data sample sizes \( n = 800, 6400, 51200 \) and 409600 (a, b, c, d), by simulating \( \hat{Q}_C(1) \) 1000 times each using draws from the uniform distribution as a substitute for generalised residuals. ........................................... 166

4.3 Left column: the kernel density estimates of \( \hat{Q}_C \) evaluated using generalised residuals arising from simulated data (orange) and its reference finite sample distribution (blue). Right column: the histograms and kernel density estimates of the plug-in \( p \)-values obtained by comparing \( \hat{Q}_C \) arising from simulated data against its reference finite sample distribution. These plots are arranged according to the (a)-(b) first component, (c)-(d) second component and (e)-(f) combined generalised residuals. ........................................... 168

4.4 Kernel density estimates of \( \hat{Q}_V \) for varying bandwidth scaling factors 0.5, 1, 1.5, 2, 2.5, 2.6458, 3, with the standard normal density function as reference. \( \hat{Q}_V \) was simulated 10000 times each. Data sizes are \( n = 800, 6400, 51200 \) and 409600. ........................................... 170

4.5 The empirical power, \( P(\text{reject } H_0|H_1 \text{ is true}) \), for the null models (a, b, c) with drift and diffusivity \((\gamma b_0, \sigma_0), (\gamma b, \sigma_0)\) and \((\gamma b, \sigma)\) versus the alternative \((b_0, \sigma_0)\), and the null model (d) with drift and diffusivity \((b, \sigma)\) versus the alternative \((b_0, \sigma_0(\zeta + (1 - \zeta)\frac{b_0}{|b_0, \zeta|}))\). The simulated data consists of trajectories with sample size \( n = 400 \). The three superimposed lines correspond to the first (blue) and second (red) component generalised residuals, and the combined (orange) generalised residuals. ........................................... 173

4.6 Kernel density estimates of the portmanteau statistic \( \hat{W}(p) \) for the first component (left) and combined (right) generalised residuals at cumulative lag \( p = 10 \). Here 10000 samples of \( \hat{W}(p) \) was used to generate the kernel density estimates. ........................................... 175
4.7 The covariance matrix of $\text{cov}[^\hat{Q}(i), ^\hat{Q}(j)]$ represented as a 2-dimensional coloured grid with each cell corresponding to a covariance matrix term, across lags $i, j \in \{1, \ldots, 10\}$, where $i \neq j$ for the first component (left) and combined (right) generalised residuals.

4.8 The covariance terms $\text{cov}[^\hat{Q}(i), ^\hat{Q}(j)]$, for lags 1 and 2, evaluated against sample size (or equivalently, number of uniform samples used as a substitute for the generalised residuals under correct model specification).

4.9 Drift and diffusivity inference results after 40 MCMC iterations excluding burn-in, from one iteration of Algorithm 14. Left: The posterior mean drift. Right: Posterior densities of the diffusivity. The discrepancy $p$-values $\hat{p}_{\text{disc}}$ for this data are 0.375717 (first component), 0.237132 (second component) and 0.255960 (combined).

4.10 The empirical distributions of the discrepancy $p$-values $\hat{p}_{\text{disc}}$ based on the (a) first, (b) second component and (c) combined generalised residuals of the data. The discrepancy $p$-values $\hat{p}_{\text{disc}}$ were sampled 200 times and used to construct the histograms and kernel density estimates here.

4.11 Drift (conditioned to be conservative) and diffusivity inference results after 40 MCMC iterations excluding burn-in from one iteration of Algorithm 14. Left: The posterior mean drift. Right: Posterior densities of the diffusivity. The discrepancy $p$-values $\hat{p}_{\text{disc}}$ for this data are 0.000000 (first component), 0.105438 (second component) and 0.045869 (combined).

4.12 The empirical distributions of the discrepancy $p$-values $\hat{p}_{\text{disc}}$ based on the (a) first, (b) second component and (c) combined generalised residuals (for drift conditioned to be conservative) of the data. The discrepancy $p$-values $\hat{p}_{\text{disc}}$ were sampled 200 times and used to construct the histograms and kernel density estimates here.
4.13 The $x$- (left column) and $y$-components (right column) of (a)-(b) the true drift, (c)-(d) the posterior mean of the unconditioned drift, and (e)-(f) the posterior mean of the drift conditioned to be conservative. 188

4.14 The posterior mean drifts of the model with unconditioned drift (left column) and the model with drift conditioned to be conservative (right column), given as (a)-(b) the full posterior mean, and (c)-(d) the $x$- and (e)-(f) $y$-components of the full posterior mean. 194

4.15 Empirical quadratic variations of the capuchin monkey movement data when sub-sampled along the time index. 198

4.16 The piece-wise constant estimates of the $\hat{\sigma}_{1,1}$, $\hat{\sigma}_{1,2}$ and $\hat{\sigma}_{2,2}$ terms (a,b,c) of the diffusivity coefficient for the capuchin monkey movement data. 201

4.17 A non-parametric estimate (left) and naive piece-wise constant estimate (right) of the drift coefficient for the capuchin monkey movement data. 201

4.18 Trajectories of the observed capuchin monkey movement data (left), and simulated data (right) from our diffusion process model using the posterior means for its drift and diffusivity. 202

4.19 Left: the transformed generalised residuals $\{Z_{i-1}^N, Z_i^N\}_{i=2}^n$ for the $x$-component of the simulated trajectory, with a red eclipse identifying any points lying more than three standard deviations away from the mean. Right: simulated trajectory from the toy diffusion process where the incremented point has been identified and circled in red, with adjacent suspect points circled in magenta. 205

4.20 Left: capuchin monkey movement trajectories which contain outliers. Right: a closer look at one of the trajectories. 205
4.21 Left column: movement data of the capuchin monkey with colour-coded data points to represent their time-indexing, with the start-points of the trajectory in yellow and the end-points in blue. Right column: histogram of minimum time differences $\Delta t$ after sub-sampling. These plots are obtained for the capuchin monkey movement data after sub-sampling so that there is a minimum time difference of (a) 10 minutes, (b) 20 minutes and (c) 30 minutes between data points.

4.22 The (a) $\hat{Q}_C(1)$ values, (b) plug-in $p$-values $\hat{p}_{\text{plug}}$ and (c) Bayesian discrepancy $p$-values $\hat{p}_{\text{disc}}$ for the capuchin monkey movement data at various levels of sub-sampling, for the diffusion process models with unconditioned drift and drift conditioned to be conservative.

4.23 The $\hat{Q}_C(1)$ values for the capuchin monkey movement data at various levels of sub-sampling, for the diffusion process models with unconditioned drift. The sub-sampling was performed in two ways: (1) data was removed sequentially to ensure a minimum time difference between any two points and (2) data was removed randomly while preserving the same number of points as obtained using the first sub-sampling method.

4.24 (a) Movement data of the olive baboon. (b) Posterior mean of the drift. Test statistic $\hat{Q}_C(1)$ values (c) with a close-up view (d) after systematic sub-sampling of the baboon movement data.

A.1 The trace plots (left) of the $x$- (top) and $y$-components (bottom) of the drift evaluated at $(0.28125, 1.6875)$, where the traces were started at different initial states of the drift. The autocorrelation plots (right) correspond to the trace with the worst autocorrelation.

A.2 The $x$- (left) and $y$-components (right) of a sample from the prior (top) and posterior (bottom) of the drift.
A.3 Comparison of the prior and posterior distributions for the x- (left) and y-components (right) of the drift using scatter plots (top) and QQ-plots (bottom) for one random point on the grid, and histograms (bottom) of p-values of KS2-tests for 200 randomly sampled points on the grid. 10000 samples were drawn from the prior for the KS2-tests, 500 of which were coupled with the 500 posterior samples to generate the scatter plots.

A.4 Trace (left) and autocorrelation (right) plots for the middle point of a randomly selected diffusion bridge.

A.5 Trace (left) and autocorrelation (right) plots for the drift evaluated at a randomly selected point on the grid.

A.6 Trace (left) and autocorrelation (right) plots for the diffusivity.

A.7 Trace (left) and autocorrelation (right) plots for the middle point of a randomly selected diffusion bridge.

A.8 Trace (left) and autocorrelation (right) plots for the drift evaluated at a randomly selected point on the grid.

A.9 Trace (left) and autocorrelation (right) plots for the diffusivity.
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>From left to right: the discrepancy statistic (top), plug-in $p$-value (middle) and discrepancy $p$-value (bottom) evaluated for the first, second components and the combined generalised residuals for the capuchin monkey movement data.</td>
<td>192</td>
</tr>
<tr>
<td>4.2</td>
<td>From left to right: the discrepancy statistic (top), plug-in $p$-value (middle) and discrepancy $p$-value (bottom) evaluated for the first, second components and the combined generalised residuals for the capuchin monkey movement data fitted with a diffusion process model with conservative drift</td>
<td>193</td>
</tr>
<tr>
<td>4.3</td>
<td>Number and percentage of data points remaining after systematic time-wise sub-sampling of the capuchin monkey movement data.</td>
<td>210</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Introduction

Diffusion processes are solutions to stochastic differential equations, which serve as a generalisation to ordinary differential equations by the addition of a “white noise” term. A wide range of phenomena have been modelled using diffusion processes since they provide a convenient way to interpret patterns which manifest within time series data. The ordinary differential equation provides a guiding force which describes the shape of the data trajectory, while the “white noise” term allows for greater versatility by accounting for deviations of the data from the ideal trajectory which may be attributed to the inherent randomness present in the phenomenon itself.

![Figure 1.1: A sample path from a 1-dimensional Ornstein-Uhlenbeck diffusion process (in blue) and its deterministic counterpart (in red).](image)

For example, Figure 1.1 shows the solution (in red) to the ordinary differential
equation
\[ \frac{dX_t}{dt} = 3 - X_t, \]
and a sampled trajectory from its diffusion process (the Ornstein-Uhlenbeck process) counterpart
\[ \frac{dX_t}{dt} = (3 - X_t) + \frac{dW_t}{dt}. \]

Simply put, diffusion processes consist of a deterministic part termed the drift which guides the trajectory and a stochastic part termed the diffusivity which dictates how “random” the deviations are. As such, diffusion processes are well suited for modelling data that are suspected to come from a framework which describes the time evolution of a trajectory subject to random deviations.

In particular, diffusion processes with drift coefficients of parametric form provide an intuitive way of understanding its effects on a trajectory. The Ornstein-Uhlenbeck diffusion process example in Figure 1.1 showcases how the \((3 - X_t)\) drift term allows for a “mean-reverting” behaviour where the trajectory is guided towards the mean 3 of the process. Unsurprisingly, Bayesian inference for diffusion models with parametric drift is a widely investigated subject. For examples of such methods we refer to the papers by Eraker [48], Beskos et al. [16] [17], and the overview by Bishwal [19] which deal exclusively with such parametric drifts, where inference for the drifts often reduces to making inferences about their respective finite-dimensional parameter vectors.

However, when fitting a diffusion process model to data, the construction of the drift of parametric form becomes a hard problem as the dimensionality of the data increases. A more general specification using a non-parametric drift potentially arising from a large family of function spaces would be preferable to the more limiting parametric form. For example in the 2-dimensional case, Brillinger [22] fitted to the trajectory of a group of elk the diffusion process model with the gradient of a potential as the drift, which allowed for the identification of multiple attraction points. However, as with drifts of parametric form, this leads to restrictions on the model since such a drift does not allow for itinerant, circular movement, which may
1.1. Introduction

be inherently present in the movement of animals, for example as brought up by Wall et al.’s [121] analysis of elephant movement.

![Figure 1.2: Left: Movement data of a capuchin monkey over the course of 46 non-consecutive days. Right: Movement data of the same capuchin monkey over the course of a day. The \( * \) and \( + \) symbols respectively denote the start and end-points of each trajectory.]

As such, for data where the shape of the drift is not immediately obvious, we can consider instead drifts without a parametric form. For example, Figure 1.2 displays the movement data of a capuchin monkey. The range and shape of the trajectories suggests a drift with strong boundary forces with underlying rotational components. The concentration of data points in the top right and bottom right of the figure also point to the possible existence of attraction points. It is therefore non-trivial to construct an appropriate drift for this data. To bypass this problem, one can turn to non-parametric Bayesian inference for the drift, where the shape of the drift is determined by the data. The main papers that cover practical Bayesian inference methodology for drift of non-parametric form are those by Papaspiliopoulos et al. [81], Pokern et al. [88] and van der Meulen et al. [116]. However, these methods are restricted to the case of 1-dimensional diffusion processes only, and as far as we are aware, no practical method is available yet for multi-dimensional diffusion processes in the literature.

In this thesis, we will focus on Papaspiliopoulos et al.’s [81] non-parametric Bayesian inference methodology for the drift of a 1-dimensional diffusion process model. They showed that through a rearrangement of the log-likelihood of the drift to achieve a quadratic form, it is natural to consider specifying a conjugate Gaussian
measure prior for the drift. This allows the construction of a tractable prior-posterior update for both the mean and precision operators of the Gaussian measure posterior conditioned on a high frequency trajectory and known diffusivity.

For one of the main contributions of this thesis, we extend this drift inference methodology to allow for multi-dimensional diffusion process models. Our innovation involves the use of the empirical measure, as opposed to the local time process in the 1-dimensional case, to factorise the log-likelihood of the drift into a quadratic form. We also extend our inference methodology to include the diffusion process model where the drift is conditioned to be conservative.

In conjunction with this drift inference methodology, we construct a composite Bayesian inference framework for the drift and diffusivity (which is assumed to be constant) of a diffusion process model, given a partially observed data trajectory. This framework involves alternating between augmenting the partially observed data with diffusion bridges, reparameterising the data, estimating the constant diffusivity, and estimating the drift.

Of course, once a diffusion process model has been fitted to the data, it would be ideal to be able to empirically determine whether it was an appropriate choice or not. Goodness-of-fit testing methods for diffusion processes were explored in the overview by Fan [50], most of which arose from financial and econometric applications. However, beyond these applications model evaluation has not been wide spread in the diffusion process literature.

In this thesis we propose a holistic framework for model evaluation for diffusion process models with the goal to improve model fit. We extend Hong and Li’s [66] goodness-of-fit hypothesis testing method to allow testing for our null diffusion process models, for both the unconditioned drift and the drift conditioned to be conservative, and investigate its performance under a finite sample setting. To complement our Bayesian inference methodology, and to transition from a frequentist hypothesis testing framework to a Bayesian setting, we implement Meng’s [76] Bayesian discrepancy $p$-value. The general problem with testing however, is that any model can be shown to be wrong given an abundance of data. We therefore
illustrate an alternative use for the above testing framework, and show how the generalised residuals used in the construction of the test statistic can be used for outlier detection. We can also use them to identify the time scales on which the diffusion process model fits best and to decide whether it is appropriate to model the drift as the gradient of a potential, i.e. whether the drift should be conservative.

1.2 Thesis Outline

To complement the main content of this thesis, we have included Chapter 2 to fix important notation as well as provide some mathematical background for the techniques used throughout this thesis.

Chapter 2: Preliminary Material

1. Section 2.1 reviews some fundamental details about diffusion processes, of which the Radon-Nikodym derivative, the local time process and the empirical measure are most important here.

2. Section 2.2 outlines the definition of Gaussian measures.

3. Section 2.3 covers some basics on partial differential equations and outlines the pseudo-spectral method which reduces a boundary value problem into a workable system of equations.

4. Section 2.4 describes the conjugate gradient algorithm which iteratively solves a system of equations, and the Lanczos algorithm which will be used to sample from Gaussian measures.

5. Section 2.5 reviews the Metropolis-Hastings algorithm and Gibbs sampler, both of which will be used extensively in our Bayesian inference methodology.

In this thesis, the main goal of Chapter 3 is to be able to perform non-parametric Bayesian inference for the unknown drift term of a multi-dimensional diffusion process with a constant but unknown diffusivity given a partially (low
frequency, discrete) observed trajectory. We can break down the composite methodology for such inference as follows:

**Chapter 3: Inference for diffusion processes**

1. Section 3.1 We briefly outline the history of parametric and non-parametric inference for the drift and diffusivity coefficients of 1- and multi-dimensional diffusion processes.

2. Section 3.2 The missing data problem posed by the partially observed trajectory is tackled by augmenting the space between any two points from the trajectory with diffusion bridges to obtain a high frequency trajectory. This requires sampling from diffusion bridges and we implement a diffusion bridge sampler in the form of a Langevin MCMC-based random walk Metropolis-Hastings algorithm.

3. Section 3.3 The high frequency augmented trajectory results in poor performance for likelihood-based diffusivity inference, so a reparameterisation technique is used to transform the trajectory into one with unit diffusivity. This transformation preserves information about the diffusivity of the underlying diffusion process while abating the high frequency-related performance issues.

4. Section 3.4 Given the reparameterised trajectory, we specify an Inverse-Wishart prior for the diffusivity coefficient and implement a random walk Metropolis-Hastings algorithm to sample from its posterior distribution.

5. Section 3.5 Given the high frequency trajectory, an existing non-parametric Bayesian inference methodology for the drift of a 1-dimensional diffusion process allows for the specification of a conjugate Gaussian measure prior. This is achieved by the use of the local time process which factorises the likelihood into a quadratic form in the drift.

6. Section 3.6 We extend the drift inference methodology for 1-dimensional diffusion processes from Section 3.5 to the multi-dimensional case, by using
1.2. Thesis Outline

the empirical measure instead of the local time process, and show that the likelihood can be factorised into a quadratic form as well to allow for a conjugate Gaussian measure prior. We implement the pseudo-spectral method alongside the conjugate gradient algorithm to solve the boundary-value problem posed by the posterior mean of the drift, and describe an amalgamate conjugate gradient and Lanczos algorithm to sample from the posterior Gaussian measure. We additionally describe an alternative diffusion process model where the prior of the drift, and thus its posterior, are conditioned to be conservative.

7. Section 3.7 We describe the composite Bayesian inference methodology using a Metropolis-within-Gibbs algorithm which alternates between augmenting the partially observed trajectory, performing the reparameterisation technique, and sampling from the posteriors of the drift and diffusivity coefficients. We illustrate this methodology for both the unconditioned drift prior and the drift prior conditioned to be conservative using a toy diffusion process as an example.

8. Section 3.8 We evaluate the performance of the Metropolis-within-Gibbs algorithm and its components.

9. Section 3.9 We repeat the composite methodology in Section 3.7 with the movement data of a capuchin monkey as a real data example.

In Chapter 4 we aim to provide a holistic model evaluation framework for diffusion processes.

Chapter 4: Model checking

1. Section 4.1 We briefly review the informal model checking and goodness-of-fit testing methods available for diffusion process models.

2. Section 4.2 We adopt the frequentist goodness-of-fit hypothesis testing methodology by implementing an existing transition density-based test statistic.
We investigate its finite sample properties and power, and determine that for practical data sizes a significant finite sample correction must be applied.

3. Section 4.3: To complement our Bayesian inference approach, we implement the Bayesian discrepancy $p$-value as an alternative to the plug-in $p$-value. We describe an efficient method to estimate the discrepancy $p$-value and examine its distribution under the null.

4. Section 4.4: We apply the above frequentist and Bayesian testing methodology to the capuchin monkey movement data example.

5. Section 4.5: The results from the testing methodology and components of the test statistic are used to identify sources of discrepancy between the diffusion process model and the data. We construct an outlier detection method and describe how the presence of multi-scale properties in the data can lead to an improved model fit. We also describe how informal model checking can be applied for diffusion processes.

Finally in Chapter 5 we present our conclusions and future work.
Chapter 2

Preliminary Material

The purpose of this chapter is to provide a brief overview of the mathematical background required for this thesis. We include several basic definitions and results from their respective source materials, with modifications only to provide some consistency to the notation used throughout this thesis. Wherever possible we will provide an explanation for their inclusion here or point forwards to wherever they are needed in this report.

2.1 Diffusion processes

As this thesis primarily deals with diffusion processes specified as solutions to stochastic differential equations, we state here some definitions from stochastic analysis and Itô calculus.

Wiener process

Most fundamental of all is the concept of a Wiener process (also known as Brownian motion), which can be thought of as the trajectory of a randomly moving particle.

Definition 1. [77, p. 7] A real-valued stochastic process $W : [0, \infty) \to \mathbb{R}$ is called a (linear) Wiener process with start in $x_0 \in \mathbb{R}$ if the following holds:

- $W_0 = x_0$,

- the process has independent increments, i.e. for all times $0 \leq t_1 \leq t_2 \leq \ldots \leq t_n$ the increments $W_{t_n} - W_{t_{n-1}}, W_{t_{n-1}} - W_{t_{n-2}}, \ldots, W_{t_2} - W_{t_1}$ are independent random variables,
• for all $t \geq 0$ and $h > 0$, the increments $W_{t+h} - W_t$ are normally distributed with expectation zero and variance $h$,

• almost surely, the function $t \mapsto W_t$ is continuous.

We say that $W$ is a standard Brownian motion if $x_0 = 0$.

Definition 2. [77, p. 36] If $W^{(1)}, \ldots, W^{(d)}$ are independent linear Wiener processes started in $x^{(1)}, \ldots, x^{(d)}$, then the stochastic process $W$ given by

$$W_t = (W_{t}^{(1)}, \ldots, W_{t}^{(d)})^T$$

is called the $d$-dimensional Wiener process started in $(x^{(1)}, \ldots, x^{(d)})^T$. The $d$-dimensional Wiener process started in the origin is also called a standard Wiener process.

\[\text{Figure 2.1: Sample paths of the 1-dimensional (left) and 2-dimensional (right) Wiener process.}\]

Figure 2.1 illustrates paths sampled from the 1- and 2-dimensional Wiener processes respectively. It can be seen that a Wiener process on its own has limited value as a statistical model since there is no guiding force acting on the trajectory of the particle: it is equally likely to move towards every direction. Deterministic models involving ordinary differential equations can provide such a guiding force, which with the addition of the Wiener process as a random perturbation, form the basis of diffusion process models.
Stochastic differential equations

The setup is as follows: consider an ordinary differential equation of the form

\[
\frac{dX_t}{dt} = b(X_t),
\]

with drift coefficient \(b\), which guides the trajectory of the particle \(X\). This equation can be generalised by the addition of the Wiener process which results in a stochastic differential equation of the form

\[
\frac{dX_t}{dt} = b(X_t) + \sigma \frac{dW_t}{dt},
\]

with diffusivity coefficient \(\sigma\), which regulates the perturbation arising from the Wiener process \(W\).

It should be noted that the \(\frac{dW_t}{dt}\) term cannot be interpreted as an ordinary derivative, as the paths of the Wiener process are nowhere differentiable (see Mörters [77, Theorem 1.30, p. 21]). Itô calculus (see for example Øksendal [80]) was invented to resolve this problem and the above equation is typically written in the following form

\[
dX_t = b(X_t)dt + \sigma dW_t.
\]

This form of the stochastic differential equation is in fact an informal way of expressing the integral equation

\[
X_t - X_s = \int_s^t b(X_u)du + \int_s^t \sigma dW_u, \quad s \leq t,
\]

which is made mathematically rigorous through the use of Itô calculus.

Diffusion processes

We refer to a diffusion process (also known as an Itô process) as a solution to a stochastic differential equation, by which we mean the following:
Definition 3. [80, p. 110] A (time-homogeneous) diffusion process is a stochastic process \( X : [0, \infty) \rightarrow \mathbb{R}^d \) satisfying a stochastic differential equation of the form

\[
dX_t = b(X_t)dt + \sigma dW_t, \quad X_0 = x_0,
\]

with \( d \)-dimensional Wiener process \( W \), drift \( b : \mathbb{R}^d \rightarrow \mathbb{R}^d \) and positive definite symmetric diffusivity \( \sigma \in \mathbb{R}^{d \times d} \) which satisfy the conditions in [80, Theorem 5.2.1, p. 66], which in this case simplify to the global Lipschitz condition

\[
|b(x) - b(y)| \leq D|x - y|, \quad x, y, \in \mathbb{R}^d,
\]

for some constant \( D \).

We note that while the diffusivity coefficient \( \sigma \) here is set to be constant, it is fairly common in the diffusion process literature for it to be spatially dependent and thus written as \( \sigma(x) \). We make this choice out of convenience, since throughout this thesis we will rely on techniques which require the assumption of a constant diffusivity coefficient \( \sigma \). In this thesis unless specified otherwise, whenever we refer to a diffusion process \( X \) we mean the diffusion process of the form given by Definition 3. In diffusion process literature the drift and diffusivity are also often written as \( b(x, t) \) and \( \sigma(x, t) \) to allow for an explicit dependence on the time index, but we since we will be working with time-homogeneous diffusion processes we do not use such notation.

As an example of a well-known diffusion process we include here the Ornstein-Uhlenbeck process.

Definition 4. [80, p. 75] The mean-reverting Ornstein-Uhlenbeck process is the solution \( X \) of the stochastic differential equation

\[
dX_t = \alpha(\beta - X_t)dt + \sigma dW_t, \quad X_0 = x_0,
\]

where \( \beta \in \mathbb{R}^d, \alpha, \sigma \in \mathbb{R}^{d \times d} \).

Figure 2.2 illustrates a path sampled from the 1-dimensional Ornstein-
2.1. Diffusion processes

Uhlenbeck process $X : [0, 10] \rightarrow \mathbb{R}$ specified as a solution of the stochastic differential equation

$$dX_t = (3 - X_t)dt + dW_t, \quad X_0 = 0,$$

where $W$ is the 1-dimensional Wiener process. As the “mean-reverting” description suggests, this trajectory converges around the “mean” term $\beta = 3$, while the $\alpha = 1$ term controls the speed at which it reverts towards the mean, which makes the Ornstein-Uhlenbeck a popular model choice, since it can for example be used to describe the price of a stock that fluctuates around a certain price.

Toy diffusion process

To ease the implementation of numerical methods in this thesis, and to satisfy the periodic boundary conditions required in Sections 2.3 and 3.5.2 we specify here the torus $\tau = \prod_{j=1}^d [a_j, c_j] \subset \mathbb{R}^d$, with periodic boundary conditions (e.g. $(a_1, x_2, \ldots, x_d)^T = (c_1, x_2, \ldots, x_d)^T \in \tau$).

We introduce here a novel 2-dimensional diffusion process which will be used throughout the thesis as a toy example and henceforth be referred to as the toy diffusion process. Let $X : [0, \infty) \rightarrow \tau$ be specified as a solution to the stochastic differential equation

$$dX_t = b(X_t)dt + \sigma dW_t, \quad X_0 = x_0,$$
where $\tau$ is a 2-dimensional torus, parameterised as $[a_1, c_1] \times [a_2, c_2]$ with $a_1 = a_2 = 0$ and $c_1 = c_2 = 9$. $W$ is the 2-dimensional Wiener process, $\sigma \in \mathbb{R}^{2 \times 2}$ is a constant diffusivity term and the drift term $b : \tau \to \mathbb{R}^2$ is given by

$$b\begin{pmatrix} x \\ y \end{pmatrix} = \frac{5}{2} \exp\left\{ 3 \left( \frac{2}{\pi} \arctan(8 - 2r) - 1 \right) \right\} (2 - r) r E_r + r E_\phi,$$

with $r, E_r$ and $E_\phi$ terms given by

$$r = \sqrt{\left( x - \frac{c_1}{2} \right)^2 + \left( y - \frac{c_2}{2} \right)^2},$$

$$E_r = r^{-1} \begin{pmatrix} x - \frac{c_1}{2} \\ y - \frac{c_2}{2} \end{pmatrix},$$

and

$$E_\phi = r^{-1} \begin{pmatrix} y - \frac{c_2}{2} \\ \frac{c_1}{2} - x \end{pmatrix},$$

for all $(x, y)^T \in \tau$.

This is an example of a drift with a rotational component and a confining force in the radial direction. Figure 2.3 (left) shows what the drift $b$ of this diffusion process looks like. The path in Figure 2.3 (right) is sampled from the toy diffusion process with diffusivity set to $\sigma = (1.4, -0.25; -0.25, 1.4)$, but we note that this diffusivity term will be set to different values depending on the context when the toy diffusion process is used throughout this thesis.

### Diffusion bridges

In practice, observed paths from diffusion processes will be discrete and sometimes of low frequency. This complicates inference for the drift and diffusivity of diffusion processes, which requires approximately continuous or high frequency observations. This is a missing data problem, which can be resolved by augmenting any two time-wise adjacent points with imputed data. We will discuss this further in Section 3.2 and note that this data augmentation method requires the use of the
2.1. Diffusion processes

Figure 2.3: A visualisation of the drift $b$ (left) and a sample path (right) from the toy diffusion process.

Figure 2.4: Further visualisations of the drift $b$, where the $x$- and $y$-components of the drift are displayed as surface plots (a)-(c) with corresponding bird’s eye view plots (b)-(d).

conditioned version of the Wiener process, the Brownian bridge.
Definition 5. [77] p. 30] Fix \( x_0, x_T \in \mathbb{R}^d \). The \( d \)-dimensional Brownian bridge with start in \( x_0 \) and end in \( x_T \) is the process \( B : [0, T] \to \mathbb{R}^d \) defined by

\[
B_t = W_t - \frac{t}{T}W_T + \left( 1 - \frac{t}{T} \right) x_0 + \frac{t}{T} x_T, \quad \forall t \in [0, T],
\]

where \( W \) is a \( d \)-dimensional Wiener process started in 0. The Brownian bridge is an almost surely continuous process such that \( B_0 = x_0 \) and \( B_1 = x_T \).

The Brownian bridge characterised by start and end-points \( B_0 = 0 \) and \( B_T = 0 \) is also known as a standard Brownian bridge. Figure 2.5 shows a sampled path of a standard Brownian bridge.

![Figure 2.5: A sample path from the 1-dimensional standard Brownian bridge process, for \( t \in [0, 1] \), with a red line marking the start and end-points \( X_0 = 0 \) and \( X_1 = 0 \).](image)

Lamperti transform

We will utilise a reparameterisation technique in Section 3.3 that involves transforming a diffusion process with diffusivity \( \sigma \) into one which has unit diffusivity. This transformation is given by the Lamperti transform as follows:

Definition 6. [68] p. 40] Suppose we have the diffusion process which solves the stochastic differential equation

\[
dX_t = b(X_t)dt + \sigma dW_t,
\]
as in Definition 3. Such a process can be transformed into one with a unitary diffusion coefficient by applying the Lamperti transform,

\[ V_t := \eta(X_t) = \int_z^X \sigma^{-1} du. \]

Here \( z \) is any arbitrary value in the state space of \( X \). Indeed, the process \( V \) solves the stochastic differential equation

\[ dV_t = \alpha(V_t) dt + dW_t, \]

where \( \alpha(v) = \sigma^{-1}b(\sigma v) \).

We note here that this definition of the Lamperti transform is for the special case where the diffusivity \( \sigma \) is a constant term, and is valid for \( d \)-dimensional diffusion processes. The usual Lamperti transform involves a spatially dependent \( \sigma(x) \), which results in a different transformed drift term \( \alpha \), and is generally unavailable for diffusion processes of dimensions greater than one, see for example Aït-Sahalia [5].

**Radon-Nikodym derivative**

We will require the density of the diffusion process \( X \) for performing inference. This density is available in the form of a Radon-Nikodym derivative as a consequence of the Girsanov theorem (see for example Øksendal [80] or Kutoyants [73]), which describes the transformation between absolutely continuous measures.

Let us define here some commonly used probability measures, namely the \( \sigma \)-Wiener measure, the measure of the diffusion process \( X \) and the \( \sigma \)-Brownian bridge measure.

**Definition 7.** The \( \sigma \)-Wiener measure \( \mathbb{W}_\sigma \) is the probability measure on \((C[0,\infty)^d,\mathcal{B}(C[0,\infty)^d))\) induced by the diffusion process \( X \) specified as the solution to

\[ dX_t = \sigma dW_t, \]
2.1. Diffusion processes

where \( W \) is the \( d \)-dimensional Wiener process, \( \sigma \in \mathbb{R}^{d \times d} \) is constant, \( C[0,\infty) \) is the space of continuous functions on \( [0,\infty) \) and \( \mathcal{B}(C[0,\infty)) \) is the Borel \( \sigma \)-algebra of \( C[0,\infty) \).

**Definition 8.** The probability measure \( \mathbb{P}_\sigma \) of \( X \) is the probability measure on \( (C[0,\infty)^d, \mathcal{B}(C[0,\infty)^d)) \) induced by the diffusion process \( X \) specified as the solution to the stochastic differential equation

\[
\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma \mathrm{d}W_t,
\]

as in Definition 3.

**Definition 9.** The \( \sigma \)-Brownian bridge measure \( \mathbb{B}_\sigma(0,x_0;T,x_T) \) is the probability measure on \( (C[0,\infty)^d, \mathcal{B}(C[0,\infty)^d)) \) induced by the Brownian bridge with start in \( x_0 \) and end in \( x_T \) defined by

\[
B_t = \sigma W_t - \frac{t}{T} \sigma W_T + \left( 1 - \frac{t}{T} \right)x_0 + \frac{t}{T} x_T, \quad \forall t \in [0,T],
\]

where \( W \) is a \( d \)-dimensional Wiener process started in 0.

We can now define the density of the diffusion process \( X \) as a Radon-Nikodym derivative as follows:

**Theorem 1.** [73, p. 34] The probability measure \( \mathbb{P}_\sigma \) of \( X_t \) and the \( \sigma \)-Wiener measure \( \mathbb{W}_\sigma \) are equivalent and the corresponding Radon-Nikodym derivative is

\[
\frac{\mathrm{d}\mathbb{P}_\sigma}{\mathrm{d}\mathbb{W}_\sigma} = \exp \left( \int_0^T S^{-1} b(x) \cdot \mathrm{d}X_t - \frac{1}{2} \int_0^T b(x) \cdot S^{-1} b(x) \mathrm{d}t \right),
\]

(2.1)

where \( S = \sigma \sigma^\dagger \).

**Euler method**

To numerically approximate the Radon-Nikodym derivative given above, we can use a discretisation method known as the Euler method.
2.1. Diffusion processes

**Definition 10.** [68, p. 62, 122] Let $X : [0,T] \rightarrow \mathbb{R}^d$ be a diffusion process specified as the solution to the stochastic differential equation

$$dX_t = b(X_t)dt + \sigma dW_t,$$

as in Definition [3] and let $\pi_N([0,T])$ be a discretisation of the interval $[0,T]$, where $0 = t_0 < t_1 < \ldots < t_N = T$ with time increment $\Delta t = t_{i+1} - t_i$ for $i = 1, \ldots, N - 1$. The Euler method approximation of $X$ is the stochastic process satisfying the iterative scheme

$$Y_{i+1} = Y_i + b(Y_i)\Delta t + \sigma (W_{i+1} - W_i),$$

where the increments $Y_{i+1} - Y_i$ are independent Gaussian random variables with mean $b(Y_i)\Delta t$ and variance $S\Delta t$.

**Lemma 1.** The transition density of $Y_i$ can be written as

$$p(y_{i+1},t_{i+1}|y_i,t_i) = (2\pi\Delta t|S|)^{-1/2} \exp \left\{ -\frac{1}{2} (\Delta t)^{-1/2} \left| S^{-1/2}(y_{i+1} - y_i - b(y_i)\Delta t) \right|^2 \right\},$$

with joint density

$$p(y_i^{N-1}) = (2\pi\Delta t|S|)^{-N/2} \exp \left\{ -\frac{1}{2} (\Delta t)^{-1/2} \sum_{i=0}^{N} \left| S^{-1/2}(y_{i+1} - y_i - b(y_i)\Delta t) \right|^2 \right\}. \quad (2.2)$$

This Euler method approximation works well for small time increments $\Delta t$, and the Radon-Nikodym derivative (2.1) can then be approximated using the joint density (2.2).

**Local time**

The local time of a process at a point $x$ essentially measures the amount of time the process spends in the neighborhood of $x$. In Section [3.5.1] we will use the local time to perform a change of variables for the density of a diffusion process from the time domain to the spatial domain.
2.1. Diffusion processes

**Definition 11.** [30 p. 141] [73 p. 27] The local time of a diffusion process $X$ at $x$ may be expressed as follows:

\[
L_t(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_0^t \mathbb{1}\{X_s \in (x - \varepsilon, x + \varepsilon)\} ds, \quad \forall t \in [0, T],
\]

\[
= \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \text{Leb}\{s \in [0, t] : X_s \in (x - \varepsilon, x + \varepsilon)\},
\]

where $\mathbb{1}$ is the indicator function and Leb is the Lebesgue measure. Additionally, for any Borel measurable and bounded function $f$

\[
\int_0^t f(X_s) ds = \int_{-\infty}^\infty f(u)L_t(u) du, \quad \forall t \in [0, T].
\]

(2.3)

**Empirical measure**

The empirical measure generalises the concept of local time to dimensions greater than 1 and unlike the local time, it is no longer absolutely continuous with respect to the Lebesgue measure. In Section 3.6.2 we will be dealing with $d$-dimensional diffusion processes, and require the empirical measure instead of the local time.

**Definition 12.** [73 p. 69] The empirical measure of a diffusion process $X$ at $x$ is defined as

\[
\Gamma_t(x) = \frac{1}{t} \int_0^t \mathbb{1}\{X_s < x\} ds, \quad \forall t \in [0, T],
\]

(2.4)

where $<$ is the vector inequality, i.e. $x < y$, $x, y \in \mathbb{R}^d$, if $x_1 < y_1, \ldots, x_d < y_d$.

Additionally, for any suitable continuous and bounded function $f$, the following can be derived

\[
\int_0^t f(X_s) dt = \int_{\mathbb{R}^d} f(x)d\Gamma_t(x), \quad \forall t \in [0, T].
\]
2.2 Gaussian measures

In Sections 3.5 and 3.6 for the purposes of drift inference, the drift term of the diffusion process of interest is specified as a priori coming from a Gaussian distribution. A Gaussian distribution on a function space is specified by a Gaussian measure, with characterising mean and covariance operator (or precision operator).

Gaussian measures

Definition 13. [38, p. 53] A probability measure $\mu$ on $(H, \mathcal{B}(H))$ is called Gaussian if for arbitrary $h \in H$ there exist $r \in \mathbb{R}$, $q \geq 0$, such that,

$$\mu \{ x \in H; \langle h, x \rangle \in A \} = N(r, q)(A), \quad A \subset \mathcal{B}(\mathbb{R}),$$

where $H$ is a Hilbert space, $\mathcal{B}(H)$ the Borel $\sigma$-algebra of $H$, $\langle \cdot, \cdot \rangle$ the inner product of $H$ and $N(r, q)(A)$ the Gaussian distribution on $A$ with mean $r$ and variance $q$.

The covariance operator and mean of a Gaussian measure are specified by:

Definition 14. [38, p. 54] If $\mu$ is Gaussian, then there exist an element $m \in H$ and a symmetric non-negative continuous operator $C$ such that:

$$\int_H \langle h, x \rangle \mu(dx) = \langle m, h \rangle, \quad \forall h \in H,$$

$$\int_H \langle h_1, x \rangle \langle h_2, x \rangle \mu(dx) - \langle m, h_1 \rangle \langle m, h_2 \rangle = \langle Ch_1, h_2 \rangle, \quad \forall h_1, h_2 \in H.$$

As will be seen in Section 3.5.2 however, it is more convenient for our purposes to specify a Gaussian measure by its precision operator instead of the covariance operator.

Definition 15. [109, p. 116] In the Hilbert space setting we refer to the inverse of the covariance operator $C$ as the precision operator.
2.3 Partial differential operators

In Section 3.6.4, we will show that the posterior mean of the drift can be obtained as the solution to a partial differential equation. Thus in this section we highlight some basic definitions for partial differential equations as well as numerical methods for solving them.

Sobolev space

Definition 16. [97, p. 116] The Sobolev space $H^k(\Omega, \mathbb{R}^d)$ is defined by

$$H^k(\Omega, \mathbb{R}^d) = \{ u : D^\alpha u \in L^2(\Omega, \mathbb{R}^d), \quad \forall 0 \leq |\alpha| \leq k \},$$

where $\Omega$ is an open subset of $\mathbb{R}^d$, $\alpha = (\alpha_1, \ldots, \alpha_d)$ is a multi-index with $|\alpha| = \alpha_1 + \ldots + \alpha_d$, and $D^\alpha$ is the $\alpha$-th partial derivative of $u$

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}.$$

Definition 17. [49, p. 245] We interpret $H^k_0(\Omega, \mathbb{R}^d)$ as comprising those functions $D^\alpha u \in H^k(\Omega, \mathbb{R}^d)$ such that

"$u = 0$ on $\partial \Omega$" for all $|\alpha| \leq k - 1$.

Partial differential operator

The precision operators of the Gaussian measures described in Sections 3.5.2 and 3.6.3 will in fact be elliptic partial differential operators.

Definition 18. [49, p. 293] Consider the boundary-value problem

$$Lu = f \quad \text{in } \Omega,$$

$$u = 0 \quad \text{on } \partial \Omega,$$

where $\bar{\Omega}$ denotes the closure of $\Omega$ and $u : \bar{\Omega} \to \mathbb{R}$ is the unknown function. Here $f : \Omega \to \mathbb{R}$ is known, and $L$ denotes a second-order partial differential operator
having the form

\[ Lu = - \sum_{j,k=1}^{d} a^{jk}(x) \frac{\partial^2}{\partial x_j \partial x_k} u + c(x)u, \]  

(2.6)

for given coefficient functions \( a^{jk} \) and \( c : \Omega \to \mathbb{R} \).

**Definition 19.** \([49, p. 294]\) We say the partial differential operator \( L \) is (uniformly) elliptic if there exists a constant \( \theta > 0 \) such that

\[ \sum_{j,k=1}^{d} a^{jk}(x) \xi_j \xi_k \geq \theta |\xi|^2, \]

for a.e. \( x \in \Omega \) and all \( \xi = (\xi_1, \ldots, \xi_d) \in \mathbb{R}^d \).

When we say that we want to find the solution to a partial differential equation, we mean the following:

**Definition 20.** \([49, p. 296]\) The bilinear form \( B[\cdot, \cdot] \) associated with the elliptic operator \( L \) defined by (2.6) is

\[ B[u, v] := \int_{\Omega} \left( - \sum_{j,k=1}^{d} a^{jk}(x) \frac{\partial}{\partial x_j} u \frac{\partial}{\partial x_k} v + cuv \right) \, dx, \]  

(2.7)

for \( u, v \in H^1_0(\Omega) \).

We say that \( u \in H^1_0(\Omega, \mathbb{R}^d) \) is a weak solution of the boundary-value problem (2.5) if

\[ B[u, v] = (f, v), \]  

(2.8)

for all \( v \in H^1_0(\Omega, \mathbb{R}^d) \), where \((,\) denotes the inner product in \( L^2(\Omega, \mathbb{R}^d) \).

We define here the spectral space representations for the function \( u \), which will be needed to solve the boundary-value problem (2.8) numerically using the spectral method. The spectral space representation of a function \( u \) uses basis functions that are non-zero over its domain \( \Omega \), for example the Fourier basis functions, which we will use here.

**Definition 21.** \([106, p. 5]\) We approximate the function of interest, \( u(x) \), with its
2.3. Partial differential operators

spectral space representation

$$u^{(N)}(x) = \sum_{n \in I} \hat{u}_n \phi_n(x), \quad x \in \Omega,$$  \hspace{1cm} (2.9)

where $$I := \{-\frac{N}{2}, \ldots, \frac{N}{2}\}^d \subset \mathbb{Z}^d$$, and the expansion coefficients are given by

$$\hat{u}_n = \int_{\Omega} \phi_n(x) u(x) dx.$$

The Fourier basis functions $$\phi_n(x)$$ given by

$$\phi_n(x) = \frac{1}{\sqrt{|\Omega|}} \exp \left\{ -i 2\pi x \cdot \frac{n}{N} \right\},$$

are orthonormal such that

$$\int_{\Omega} \phi_n(x) \phi_m(x) dx = \delta_{nm}, \quad \forall n, m \in I,$$

where $$\delta_{nm}$$ is the Kronecker delta function.

Spectral method

The spectral method (see Shizgal [106], Trefethen [112]) is used to numerically solve differential equations, by representing the solution of a differential equation as a sum of orthonormal basis functions which are non-zero over the whole domain.

As an example application of the spectral method, let us consider the operator (2.6) for constant functions $$a^{ij}$$ and $$c$$, which gives

$$f = -\sum_{j,k = 1}^{d} a^{jk} \frac{\partial^2}{\partial x_j \partial x_k} u + cu,$$  \hspace{1cm} (2.10)

Now to solve $$u$$ for the boundary-value problem associated with (2.10), we first approximate (2.10) by replacing $$u$$ with its respective spectral space representations
of \( u^{(N)} \) to give

\[
f(x) = - \sum_{j,k=1}^{d} a^{jk} \frac{\partial^2}{\partial x_j \partial x_k} u^{(N)}(x) + cu^{(N)}(x),
\]

\[
= - \sum_{j,k=1}^{d} a^{jk} \frac{\partial^2}{\partial x_j \partial x_k} \left( \sum_{n \in I} \hat{u}_n \phi_n(x) \right) + c \sum_{n \in I} \hat{u}_n \phi_n(x),
\]

where the equality for \( f \) in (2.11) holds weakly in the sense that, when integrated against any \( \phi_n \), the equality holds. We note that \( a^{jk} \) and \( c \) have been preserved since \( a^{jk} = a^{(N)jk}, c^{(N)} = c \), and that the second order partial differential operator in (2.11) can be reduced to

\[
\frac{\partial^2}{\partial x_j \partial x_k} \phi_n(x) = -(2\pi)^2 \frac{n_j n_k}{N^2} \phi_n(x).
\]

By inserting (2.12) into the inner product of (2.11) and \( \phi_m, m \in I \), over \( \Omega \), the Fourier coefficients of \( f \) can then be written as

\[
\hat{f}_m := \int_{\Omega} \phi_m(x)f(x)dx = \sum_{j,k=1}^{d} a^{jk} \hat{u}_m (2\pi)^2 \frac{n_j n_k}{N^2} + c \hat{u}_m,
\]

The strength of the spectral method here lies in the simplification of the second order partial differential operator \(-\frac{\partial^2}{\partial x_j \partial x_k}\) in the real basis into a vector multiplication involving \((2\pi)^2 \frac{n_j n_k}{N^2}\) in the Fourier basis, with an end result being a linear system of equations described by (2.13), which is in turn easily solvable for \( \hat{u} \) because the matrix involved is diagonal.
Pseudo-spectral method

Let us now consider a similar set up as in (2.10), but with a function $c(x)$ that is no longer constant, such that

$$f(x) = - \sum_{j,k=1}^{d} a^{jk} \frac{\partial^2}{\partial x_j \partial x_k} u(x) + c(x)u(x).$$

(2.14)

If we proceed along the lines of (2.11) to (2.13), we find that the spectral method is no longer effective, since the term involving $c(x)$ remains as

$$\int_{\Omega} \phi_{m}(x) c^{(N)}(x) \phi_{n}(x) dx = \int_{\Omega} \phi_{m}(x) \sum_{k \in I} \hat{c}_{k} \phi_{k}(x) \phi_{n}(x) dx.$$  

(2.15)

To numerically evaluate this integral involving triple-products of Fourier basis functions, a non-sparse matrix would have to be pre-computed and stored, especially if all (or most) of the $\hat{c}_{k}$ Fourier coefficients are non-zero, which will be true in our application in this thesis. This results in significant storage and computational costs.

Instead, we would like to evaluate $c(x)u(x)$ by remaining in the physical space. Let us make the dependence on the Fourier index explicit, such that $\hat{f}_{m} = \hat{f}(m)$. We denote by $\mathcal{F}$ and $\mathcal{F}^{-1}$ the usual Fourier transform operator and its inverse, such that

$$\mathcal{F}(f)(m) := \hat{f}(m),$$

and represent (2.13) as

$$\hat{f}(m) = \sum_{j,k=1}^{d} a^{jk} (2\pi)^{2} \frac{m_{j} m_{k}}{N^2} \hat{u}(m) + \mathcal{F} \left( c(x) \mathcal{F}^{-1}(\hat{u}) \right)(m).$$

(2.16)

Treating (2.16) as a system of equations allows us to think of $\sum_{j,k=1}^{d} a^{jk} (2\pi)^{2} \frac{m_{j} m_{k}}{N^2}$ and $c(x)$ as diagonal in the Fourier and the real basis representations respectively.
The resultant system of equations can be visualised by

\[
\hat{f} = \begin{pmatrix}
\ddots & & \\
& \ddots & \\
& & \ddots
\end{pmatrix} \hat{u} + \mathcal{F} \begin{pmatrix}
\ddots & & \\
& \ddots & \\
& & \ddots
\end{pmatrix} \mathcal{F}^{-1}(\hat{u}).
\]

This is an example of the pseudo-spectral method, see for example Fornberg [52] and Shizgal [106], which takes advantage of the simplifications resulting from the spectral method while retaining benefits from staying with the physical space representation. The system of equations (2.16) can then be solved using numerical iterative methods, which we will briefly cover in the next section.
2.4 Krylov subspace methods

Conjugate gradient algorithm

To solve partial differential equations which have been reduced to a system of equations of form (2.16), we refer to the family of Krylov subspace methods (see Trefethen [113]). Consider a non-singular system of equations $Ax = b$, with exact solution $x_\ast = A^{-1}b$. Denote by $\mathcal{K}_n$ the $n$-th Krylov subspace generated by $b$,

$$
\mathcal{K}_n(A, b) = \text{span}\{b, Ab, \ldots, A^{n-1}b\}
$$

The conjugate gradient algorithm is a Krylov subspace method for finding the solution $x_\ast$, such that at each iterative step $n$, $x_\ast$ is approximated by a vector $x_n \in \mathcal{K}_n$ that minimises the norm $\|b - Ax_n\|_A = \sqrt{(b - Ax_n)^\top A(b - Ax_n)}$. This algorithm is given by:

Algorithm 1. [113] p. 293 Conjugate gradient iteration

1. Initate: $x_0 = 0$, $r_0 = b$, $p_0 = r_0$

2. Iterate for $n = 1, 2, 3, \ldots$

   (a) $\alpha_n = (r_n^\top r_{n-1})/(p_{n-1}^\top Ap_{n-1})$

   (b) $x_n = x_{n-1} + \alpha_n p_{n-1}$

   (c) $r_n = r_{n-1} - \alpha_n A p_{n-1}$

   (d) $\beta_n = (r_n^\top r_n)/(r_{n-1}^\top r_{n-1})$

   (e) $p_n = r_n + \beta_n p_{n-1}$

The conjugate gradient algorithm allows us to treat $A$ as a black-box function rather than as an explicit matrix form since it does not enter the algorithm above unaccompanied. This is useful in particular for the pseudo-spectral method given in Section 2.3 since $A$ will have to be written as a combination of matrices and Fourier transforms.
2.4. Krylov subspace methods

Lanczos algorithm

In Section 3.6.5, we will need to calculate the product \( A^{-1/2}z \) for some vector \( z \) and a matrix \( A \) that is not easily invertible. Instead, we rely on the approximation of \( A^{-1/2}z \) which can be calculated using a matrix \( Q_n \) such that \( Q_n^T A Q_n \) is a symmetric tridiagonal matrix. This leads us to the Lanczos algorithm, which is an iterative method to reduce a matrix \( A \) to a symmetric tridiagonal matrix form \( T_n = Q_n^T A Q_n \), and is given by:

**Algorithm 2. [113] p. 277] Lanczos iteration**

1. Initialize Lanczos: \( q_0 = 0, q_1 = \frac{b}{\|b\|_2}, \beta_0 = 0 \)

2. Iterate for \( n = 1, 2, 3, \ldots \)
   
   (a) \( v = Aq_n \)
   
   (b) \( \alpha_n = q_n^T v \)
   
   (c) \( v = v - \beta_{n-1}q_{n-1} - \alpha_n q_n \)
   
   (d) \( \beta_n = \|v\|_2 \)
   
   (e) \( q_{n+1} = \frac{v}{\beta_n} \)

3. Return the tridiagonal matrix \( T = \begin{pmatrix} \alpha_1 & \beta_2 \\ \beta_2 & \alpha_2 & \beta_3 \\ & \beta_3 & \alpha_3 & \beta_4 \\ & & & \ddots & \ddots \end{pmatrix} \).
2.5 Markov chain Monte Carlo methods

Suppose there is a distribution \( \pi(\theta) \) from which it is impossible to sample the random variable \( \theta \) directly. Markov chain Monte Carlo (MCMC) methods rely on the construction of Markov chains whose stationary distribution is the target distribution \( \pi(\theta) \). By iteratively sampling from these Markov chains over a long enough period of time, the draws \( \theta \) can be considered as coming from the target distribution \( \pi(\theta) \).

Throughout this thesis we will rely on three commonly used MCMC methods, namely the Metropolis-Hastings algorithm, the Gibbs sampler and the Metropolis-within-Gibbs sampler to sample from posterior distributions with no closed form. For further descriptions of MCMC methods, see for example Liu [74], Gelman et al. [57] and Robert and Casella [94].

**Metropolis-Hastings algorithm**

The goal of the Metropolis-Hastings algorithm is to construct a Markov chain in the state space of \( \theta \), so that the stationary distribution of this chain is the target distribution \( \pi(\theta) \). At every step of the Markov chain, a proposal is drawn from a proposal distribution which depends on the previous state of the chain. This proposal is then accepted or rejected using a rule constructed by comparing the densities of the proposal and target distributions. The collection of states of the Markov chain can then be considered to come from the target distribution \( \pi(\theta) \).

**Algorithm 3.** [74] p. 111] Metropolis-Hastings algorithm

1. Target: \( \pi(\theta) \)

2. Initialisation: Set \( \theta^{(0)} \) at some initial value

3. Iteration \( j+1 \): Given current state \( \theta^{(j)} \)
   
   (a) Sample \( \theta' \) from a proposal distribution \( T(\theta'|\theta^{(j)}) \)
(b) Sample $u \sim \text{Uniform}[0,1]$ and update

$$
\theta^{(j+1)} = \begin{cases} 
\theta' & \text{if } u \leq \alpha(\theta^{(j)}, \theta') \\
\theta^{(j)} & \text{otherwise}
\end{cases}
$$

with acceptance ratio

$$
\alpha(\theta^{(j)}, \theta') = \min\left\{1, \frac{\pi(\theta')T(\theta^{(j)}|\theta')}{\pi(\theta^{(j)})T(\theta'|\theta^{(j)})}\right\}.
$$

**Gibbs sampler**

Suppose now that $\theta = (\theta_1, \ldots, \theta_d)$ is a $d$-dimensional random variable. To sample from the joint distribution $\pi(\theta_1, \ldots, \theta_d)$, the Gibbs sampler cycles between sampling the individual components, say $\theta_i$, from the distribution $\pi(\theta_i|\theta_{-i})$ conditional on all of the other components $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$.

**Algorithm 4. [74, p. 130] Systematic-scan Gibbs sampler**

1. Target: $\pi(\theta_1, \ldots, \theta_d)$
2. Initialisation: Set $(\theta_1^{(0)}, \ldots, \theta_d^{(0)})$ at some initial value
3. Iteration $j + 1$: Given current state $(\theta_1^{(j)}, \ldots, \theta_d^{(j)})$

   (a) Sample $\theta_i^{(j+1)}$ from $\pi(\theta_i|\theta_1^{(j+1)}, \ldots, \theta_i^{(j+1)}, \theta_{i-1}^{(j)}, \theta_{i+1}, \ldots, \theta_d^{(j)})$ sequentially for $i = 1, \ldots, d$.

**Metropolis-within-Gibbs sampler**

The Metropolis-within-Gibbs sampler (see Robert and Casella [94, p. 389]) combines both the Gibbs sampler and the Metropolis-Hastings algorithm, whereby step 3.(a) in Algorithm 4 is replaced by Algorithm 3 as necessary.
Chapter 3

Inference for diffusion processes

3.1 Introduction

As mentioned in the thesis introduction, diffusion processes have been used throughout science as statistical models to describe a vast variety of time series-based phenomena. Throughout this thesis, we consider diffusion process models of a specific form as described by Definition 3. In particular, suppose that we are provided with a low frequency, discretely observed \(d\)-dimensional trajectory \(\{X_t\}_{i=1}^n \subset \mathbb{R}^d\). We can fit to this data a \(d\)-dimensional diffusion process model specified as the solution \(X: [0, T] \rightarrow \mathbb{R}^d\) of the stochastic differential equation

\[
dX_t = b(X_t)dt + \sigma dW_t, \quad X_0 = x_0,
\]

with unknown drift \(b: \mathbb{R}^d \rightarrow \mathbb{R}^d\) and unknown constant diffusivity \(\sigma \in \mathbb{R}^{d \times d}\), where \(W\) is the \(d\)-dimensional Wiener process and \(t_1 = 0 < \ldots < t_n = T\). Given the provided trajectory, we are then interested in inferring the drift \(b\) and diffusivity \(\sigma\) of the diffusion process model via Bayesian methodology in this chapter.

The study of diffusion processes has a long and established history, but coefficient estimation of diffusion process models as described above remains relatively uncharted territory. We will provide a very brief summary of the available literature as categorised by frequentist and Bayesian approaches, of which the Bayesian works are of more relevance by the very nature of our work in this chapter. We further highlight the difficulties involved with the availability of continuous-time
versus discrete-time observations, parametric versus non-parametric inference of coefficients and 1-dimensional versus multi-dimensional diffusion process models.

**The frequentist approach**

The mathematics underlying the inference problem for the coefficients of 1-dimensional diffusion processes are well known asymptotically. Summaries of the progress made via the frequentist approach can be found in Kutoyants’ [73] overview of asymptotically efficient parameter estimation for continuously observed ergodic diffusion processes, and in Bishwal’s [19] textbook on the properties of parametric estimators for diffusion processes under the asymptotic frameworks involving either continuously observed trajectories over long time periods or discretely observed trajectories with decreasing inter-observation times.

While the assumption of continuously observed trajectories leads to pleasing theoretical results, it is simply not practical when actual data is concerned. Ait-Sahalia [4] was able to construct closed-form approximations to the transition density and likelihood of partially observed diffusion processes despite large time steps between observations, and this was followed by works such as those by Beskos et al. [17] who considered likelihood-based parametric inference for the drift and diffusivity of diffusion processes. Fuchs [54] provided an overview of practical methods to deal with discrete-time observations, including methods based on maximum likelihoods, indirect inference, estimating functions and method of moments.

It should be noted that for continuous-time observations, estimation of the diffusivity coefficient is trivial due to the availability of the quadratic variation (see for example Kutoyants [73]). Even for the discrete-time case in the literature stated above, the Lamperti transformation under certain conditions transforms a diffusion process with a non-unit diffusivity coefficient into one with a unit diffusivity coefficient, and the problem becomes that of drift inference once more, provided that the diffusivity coefficient that determines the Lamperti transformation has been estimated.

Further advances have then been made, with emphasis on non-parametric es-
3.1. Introduction

Estimators, which allowed for consideration of diffusion process models with coefficients of more general form. For example, Comte et al. [33] considered penalised non-parametric mean square estimators for the coefficients of diffusion processes. Cattiaux et al. [25] [26] [27] obtained central limit theorems for some non-parametric estimators of the drift and diffusivity of partially observed stochastic damping Hamiltonian systems (a specific use case for hypoelliptic diffusion processes). Comte and Genon-Catalot [32] investigated non-parametric drift estimation on non-compact support for univariate diffusion processes.

The multi-dimensional case on the other hand, has proven to be difficult, with relatively fewer theoretical contributions. For example, Dalalyan and Reiβ [40] investigated non-parametric drift estimation for continuously observed ergodic diffusion processes. Meanwhile, Schmisser [102] considered penalised non-parametric estimation of the drift for discretely observed ergodic diffusion processes.

The Bayesian approach

As Bayesian inference for the coefficients of diffusion process models relies on the likelihood being analytically available or on having continuously observed trajectories (in which case the likelihood can be numerically approximated), investigations into Bayesian inference for 1-dimensional diffusion processes did not begin in earnest until the problem of having partially (or discretely) observed trajectories was solved. Roberts and Stramer [95] overcame this missing data problem by introducing Markov chain Monte Carlo approaches to data augmentation using diffusion bridges. This alongside Aït-Sahalia’s [4] likelihood approximations allowed for the advancement of parametric Bayesian inference, a summary of which we once again defer to Fuchs [54].

The study of non-parametric Bayesian inference for 1-dimensional diffusion processes is relatively new, with pioneering work by Papaspiliopoulos et al. [81] who introduced a conjugate Gaussian prior for the drift. The asymptotic behaviour of the Gaussian posterior is then investigated in Pokern et al. [88]. van der Meulen et al. [116] extended Papaspiliopoulos et al.’s [81] framework to involve alternative
basis functions for the drift when constructing the prior, while van Waaij and van Zanten [119] further improve convergence rates of the Gaussian posterior by introducing hyper-priors on either the multiplicative scaling parameter or the regularity parameter of the Gaussian prior.

As with the frequentist approach however, diffusivity inference is not prevalent. Gugushvili et al. [60] studied non-parametric diffusivity inference using a histogram-type prior with piecewise constant partitions over the time interval of the trajectory.

Investigations into the asymptotic behaviour of the posterior distributions of the coefficients of diffusion processes remain ongoing. For example, van der Meulen and van Zanten [117] studied the posterior consistency for the drift with concrete priors, while Nickl and Söhl [78] proved that optimal contraction rates of posterior distributions of the drift and diffusivity can be achieved under certain conditions for the prior distributions.

Bayesian inference for multi-dimensional diffusion processes is also a recent development. Schauer et al. [101] introduced guided proposals for multi-dimensional diffusion bridges, leading to the works by van der Meulen and Schauer [114] [115] where these guided proposals were used to perform parametric drift inference for discretely observed diffusion processes. Meanwhile, Nickl and Ray [78] investigated the posterior contraction rates of a Bayesian maximum a posteriori (MAP) estimate of the drift stemming from a Gaussian prior.

**Thesis contribution and problem statement**

It has become clear that the missing next step is to be able to perform non-parametric Bayesian inference for the drift of partially observed multi-dimensional diffusion processes. Within this chapter we aim to provide a practical method to do so by extending Papaspiliopoulos et al.’s [81] framework to the multi-dimensional case.

Therefore the goal of this chapter is to construct an encompassing methodology to perform Bayesian inference for both the unknown drift $b$ and the unknown constant diffusivity $\sigma$. We have arranged this chapter so as to address each step of
the process, which can be summarised by the following: data augmentation (Section 3.2), reparameterisation (Section 3.3), diffusivity inference (Section 3.4) and drift inference for the 1-dimensional case (Section 3.5) and the multi-dimensional case (Section 3.6). The composite methodology (Section 3.7) and an evaluation of its performance (Section 3.8) are then presented preceding an applied example (Section 3.9).
3.2 Partially observed diffusion processes

First of all, suppose that we are provided with a low frequency, discretely observed $d$-dimensional trajectory $\{X_t\}_{t=0}^n = \{X_t\}_{t=0}^n \subset \mathbb{R}^d$. We fit to this data a $d$-dimensional diffusion process model specified as the solution $X : [0, T] \rightarrow \mathbb{R}^d$ of the stochastic differential equation

$$\text{d}X_t = b(X_t)\text{d}t + \sigma \text{d}W_t, \quad X_0 = x_0,$$

(3.1)

with unknown drift $b : \mathbb{R}^d \rightarrow \mathbb{R}^d$ and unknown constant diffusivity $\sigma \in \mathbb{R}^{d \times d}$, where $W$ is the $d$-dimensional Wiener process and $t_1 = 0 < \ldots < t_n = T$.

Let us assume for the moment that the diffusivity $\sigma$ is known. We would then want to perform inference for the drift given the low frequency trajectory and the known diffusivity $\sigma$. But as shall be seen in Sections 3.5 and 3.6, a continuous or relatively high frequency trajectory is required, firstly to ensure the availability of the likelihood of the drift, which is used in the specification of the conjugate prior for the drift and the posterior of the diffusivity, and secondly to perform the necessary calculations involved in the update equation for the resulting posterior mean of the drift. For practical reasons, in both cases a high frequency, approximately continuous trajectory is sufficient. For the first case, as the frequency of the trajectory increases, the Euler method (see Definition 10) allows the construction of the joint transition density of a high frequency trajectory $\{X_t\}_{t=1}^n$ to approximate the likelihood of the drift of the diffusion process model. For the second case, the update equation for the posterior mean consists of an integral over the time index, and simply requires a high enough frequency to allow for a good approximation through its corresponding Riemann sum.

Thus we need a method to supplement the low frequency trajectory with additional points to obtain an approximately continuous, high frequency trajectory. In other words, we need to be able to sample additional trajectories from the underlying diffusion process model to connect any two time-wise adjacent points from the original discretely observed trajectory $\{X_t\}_{t=1}^n$. This is known as a data aug-
3.2. Partially observed diffusion processes

3.2.1 Diffusion bridges

Once again, let \( X : [0, T] \to \mathbb{R}^d \) be a \( d \)-dimensional diffusion process specified as the solution to the stochastic differential equation

\[
dX_t = b(X_t)dt + \sigma dW_t, \quad X_0 = x_0,
\]

with drift \( b : \mathbb{R}^d \to \mathbb{R}^d \) and constant diffusivity \( \sigma \in \mathbb{R}^{d \times d} \), where \( W \) is the \( d \)-dimensional Wiener process. Now condition the process \( X \) to not only have a fixed value \( x_0 \) at time 0 but also \( v \) at time \( T \). The resultant conditioned process \( X^* := X\{X_0 = x_0, X_T = v\} \) is referred to as the diffusion bridge of (3.2) with starting point \( X_0 = x_0 \) and end-point \( X_T = v \), or the \((0, x_0, T, v)\)-diffusion bridge for short.

The \((0, x_0, T, v)\)-diffusion bridge process \( X^* \) can then be specified (see for example Rogers and Williams [98] and Lyon [75]) as the solution to the stochastic differential equation

\[
dX^*_t = (b(X^*_t) + \sigma \sigma^T \nabla_x \log(p(v, T | X^*_t, t))) dt + \sigma dW_t, \quad X^*_0 = x_0,
\]

where the coefficients \( b \) and \( \sigma \) in (3.3) remain identical to those in (3.2). The conditional density \( p(x_r, r | x_s, s) \) is defined as the transition density of the original diffusion process \( X \) from \( X_s = x_s \) at time \( s \) to \( X_r = x_r \) at time \( r \) with \( s \leq r \) and \( r, s \in [0, T] \). This transition density \( p(x_r, r | x_s, s) \) in the new drift term \((b(X^*_r) + \sigma \sigma^T \nabla_x \log(p(T, v | x^*_r)))\) is what guides the path of the diffusion bridge process towards the end-point \( X_T = v \).

When sampling from diffusion bridge processes, difficulties arise since the construction of a diffusion bridge process requires conditioning on the event that the path reaches a fixed end-point \( X_T = v \), which is an event of probability zero. This means that rudimentary sampling methods, for example a simple rejection algorithm based on the path reaching a fixed end-point, would be very computationally expensive. Additionally, the closed form of the transition density \( p(x_r, r | x_s, s) \)
of the original diffusion process $X$ and hence that of the bridge process $X^*$ are often unavailable in general. If they were, as in the case of the Ornstein-Uhlenbeck process (see for example Iacus [68]), then one would be able to simulate the diffusion bridge directly by sampling from its known transition density.

Therefore, we refer to existing Markov chain Monte Carlo (MCMC) accept-reject-type methods. The idea behind these methods is to instead sample from a proposal process $X^\circ$, whose law $P^\circ$ should be absolutely continuous with respect to the law $P^*$ of the target diffusion bridge process $X^*$. This proposal process $X^\circ$ should be chosen so that it is easy to sample from, and that an acceptance probability for the proposal samples can be constructed from the Radon-Nikodym derivative $\frac{dP^*}{dP^\circ}$. Samples from the target diffusion bridge process $X^*$ can thus be generated from the accepted proposals of $X^\circ$.

### 3.2.2 Random walk Metropolis-Hastings

One such MCMC method is the random walk Metropolis-Hastings sampler, which is a special case of the Metropolis-Hastings algorithm described in Section 2.5. In this sampler, successively new proposals are generated as a random perturbation of the previously accepted proposals.

**The target distribution**

We will firstly identify the law of the target $(0,x_0,T,v)$-diffusion bridge process along the lines of Roberts and Stramer [95] to allow its use in constructing an acceptance ratio for the random walk Metropolis-Hastings sampler.

Let us denote $G(X,b,\sigma)$ as the Radon-Nikodym derivative of the law $P$ of the original diffusion process $X$ of (3.2) with respect to the Wiener measure $\mathbb{W}_\sigma$ induced by the martingale $\sigma dW_t$, which is given as

$$G(X,b,\sigma) = \frac{dP}{d\mathbb{W}_\sigma} = \exp \left[ -\frac{1}{2} \int_0^T b(X_t) \cdot S^{-1} b(X_t) \, dt + \int_0^T S^{-1} b(X_t) \cdot dX_t \right],$$

where $S = \sigma \sigma^T$.

Consider now the $\sigma$-Brownian bridge $B$ as defined in Section 2.1 with start and
end-points $B_0 = x_0$ and $B_T = v$. Denote by $\mathbb{B}_\sigma(s, x_s; r, x_r)$ the law of the $\sigma$-Brownian bridge with start and end-points $X_s = x_s$ and $X_r = x_r$. The Wiener measure $\mathbb{W}_\sigma$ can then be factorised as

$$\mathbb{W}_\sigma = \mathbb{B}_\sigma(0, x_0; T, v) \otimes \text{Leb}_2(x) \times f(t, x, \sigma),$$

where $\text{Leb}_2$ is the 2-dimensional Lebesgue measure and $f$ is the density of the start and end-points with respect to the Lebesgue measure, under the Wiener measure $\mathbb{W}_\sigma$, given as

$$f(T, v, \sigma) = \frac{1}{(2\pi|S|)^{1/2}} \exp\left\{-\frac{|S^{-1}(v - x_0)|^2}{2T}\right\}.$$

Let us write $X_{\text{mis}} = \{X_t\}_{t \in [0, T]} \setminus \{X_0, X_T\}$ to emphasize that this is the continuous trajectory $\{X_t\}_{t \in [0, T]}$ with the exclusion of the start and end-points. The density of $X_{\text{mis}}$ conditioned on the start and end-points with respect to the Brownian bridge measure $\mathbb{B}_\sigma$ is given by

$$\frac{d\mathbb{P}_\sigma}{d\mathbb{B}_\sigma(0, x_0; T, v)}(X_{\text{mis}}|X_0, X_T) = G(X; b, \sigma) f(t, x, \sigma) g_b(t, x, \sigma) \propto G(X; b, \sigma), \quad (3.4)$$

where $g_b$ is the 2-dimensional Lebesgue density of $X$ under $\mathbb{P}_\sigma$.

Thus the Radon-Nikodym derivative $G(X; b, \sigma)$ can be used as a substitute for our target density of the $(0, x_0, T, v)$-diffusion bridge process.

**The proposals**

It can be shown (see Stuart et al. [110] and Hairer et al. [65]) that the law of the $(0, x_0, T, v)$-diffusion bridge is equivalent to that of the $\mathcal{L}^2([0, \infty) \times [0, T], \mathbb{R}^d)$-valued solution $Y$ to the stochastic partial differential equation

$$\frac{\partial}{\partial u} Y(u, t) = Y(u, t) + \sqrt{2C} \frac{\partial}{\partial u} W_u, \quad (3.5)$$

with $\mathcal{L}^2([0, T], \mathbb{R}^d)$-valued Wiener process $W$, boundary conditions $Y(u, 0) = x_0$ and $Y(u, T) = v$, and $C$ the covariance of the $(0, x_0, T, v)$-diffusion bridge.
3.2. Partially observed diffusion processes

We suppress the notation $t$ for convenience, and discretise (3.5) with respect to $u$ using the theta method (see [63, p. 204]) with $\theta = \frac{1}{2}$ and increment step $\Delta u$, to give

$$
\frac{y_u - y_{u+\Delta u}}{\Delta u} = -\frac{1}{2}(y_u + y_{u+\Delta u}) + \sqrt{\frac{2}{\Delta u}} \omega
$$

(3.6)

with Brownian bridge $\omega \sim \mathcal{B}_\sigma(0, x_0; T, v)$ which encapsulates the boundary conditions of $Y$ and the covariance $C$.

By rearranging (3.6), we can then denote our random walk Metropolis-Hastings proposal $y'$ given a previously accepted sample $y^{(j)}$ as

$$
y' = \sqrt{1 - \rho^2} y^{(j)} + \rho \omega,
$$

(3.7)

where $\rho$ controls the step size of the random walk.

Random walk Metropolis-Hastings

These proposals (3.7) can be considered as a special case arising from the Langevin MCMC method, see for example Stuart et al. [110] and Beskos et al. [18]. As demonstrated by Cotter [34], these proposals lead to a symmetric proposal distribution, so that the computation of the Metropolis-Hastings acceptance ratio involves the quotient of the Radon-Nikodym derivatives shown on the right hand side of (3.4).

The acceptance ratio can therefore be constructed as

$$
\alpha(y^{(j)}, y') = \min \left\{ 1, \frac{G(y', b, \sigma)}{G(y^{(j)}, b, \sigma)} \right\},
$$

and the prescription for the random walk Metropolis-Hastings sampler for $(0, x_0, T, v)$-diffusion bridge processes is then given as:

Algorithm 5. Random walk Metropolis-Hastings: diffusion bridge sampler

1. Target: $(0, x_0, T, v)$-diffusion bridge

2. Initialisation: Set $y^{(0)}$ at some initial value
3. Iteration $j + 1$: Given current state $y^{(j)}$

(a) Sample the $\omega$ Brownian bridge from $\mathbb{B}_\sigma(0, x_0; T, \nu)$

(b) Set $y' = \sqrt{1 - \rho^2} y^{(j)} + \rho \omega$

(c) Sample $\nu \sim \text{Uniform}[0,1]$ and update

$$y^{(j+1)} = \begin{cases} 
   y' & \text{if } \nu \leq \alpha(y^{(j)}, y') \\
   y^{(j)} & \text{otherwise}
\end{cases}$$

with acceptance ratio

$$\alpha(y^{(j)}, y') = \min \left\{ 1, \frac{G(y', b, \sigma)}{G(y^{(j)}, b, \sigma)} \right\}.$$ 

4. The accepted $\{y^{(j)}\}_{j=1}^n$ proposals then come from distribution of the $(0, x_0, T, \nu)$-diffusion bridge.

**Conditioning on a discrete path**

Referring back to the data augmentation problem described in the beginning of Section 3.4, we now wish to augment the low frequency, discretely observed path $\{X_{ti}\}_{i=0}^n$ with diffusion bridges to obtain a high frequency, approximately continuous trajectory which we denote as $X_{\text{aug}}$.

This can be achieved by first treating the individual points of $\{X_{ti}\}_{i=0}^n$ as pairs of start and end-points $(x_{ti-1}, x_{ti})$ so that the Wiener measure $\mathbb{W}_\sigma$ can now be factorised as

$$\mathbb{W}_\sigma = \prod_{i=1}^n \mathbb{B}_\sigma(t_{i-1}, x_{ti-1}; t_i, x_{ti}) \otimes \text{Leb}_n(x) \times f(t, x, \sigma),$$

where $\text{Leb}_n$ is the $n$-dimensional Lebesgue measure and $f$ is now the joint density of the start and end-points with respect to the Lebesgue measure, under the Wiener measure $\mathbb{W}_\sigma$, given as

$$f(t, x, \sigma) = \frac{1}{(2\pi|S|)^{n/2}} \exp \left\{ -\sum_{i=1}^n \frac{|S^{-1}(x_{ti} - x_{ti-1})|^2}{2(t_i - t_{i-1})} \right\}.$$
3.2. Partially observed diffusion processes

Let us denote $X_{\text{mis}} = X_{\text{aug}} \setminus \{X_{t_i}^n\}_{i=0}^n$ the augmented path with the collection of start and end-points $(x_{t_i-1}, x_{t_i})$ removed. The density of $X_{\text{mis}}$ conditioned on the low frequency trajectory $\{X_{t_i}^n\}_{i=0}^n$ with respect to the Brownian bridge measure $\mathbb{B}_\sigma$ is given by

$$
\frac{dP_\sigma}{\otimes_{i=1}^n \mathbb{B}_\sigma(t_{i-1}, x_{t_{i-1}}; t_i, x_{t_i})}(X_{\text{mis}} | \{X_{t_i}^n\}_{i=0}^n) = G(X_{\text{aug}}; b, \sigma) \frac{f(t, x, \sigma)}{g_b(t, x, \sigma)} \propto G(X_{\text{aug}}; b, \sigma),
$$

(3.8)

where $g_b$ is the $n$-dimensional Lebesgue density of $X$ under $P_\sigma$.

The proposals can be written as

$$
y' = \sqrt{1 - \rho^2} y(j) + \rho \omega,
$$

(3.9)

with $y(j) = \{y(j)_i\}_{i=1}^n$ and $y' = \{y'_i\}_{i=1}^n$ formed by concatenating independent diffusion bridges $y_i^{(j)}$, $y'_i$.

Algorithm 5 can then be easily modified to include the Radon-Nikodym derivative (3.8) and proposals (3.9) to allow sampling for the high frequency path $X_{\text{aug}}$ given a low frequency path $\{X_{t_i}^n\}_{i=0}^n$.

As an example, in Figure 3.1 (left), we simulate a high frequency path from the 2-dimensional toy diffusion process described in Section 2.1 and denote by $\{X_{t_i}^n\}_{i=0}^n$ a small subset of points (marked by red dots) which we wish to augment with diffusion bridges. In Figure 3.1 (right) we plot an accepted proposal from the random
walk Metropolis-Hastings sampler, which samples from the collection of diffusion bridges conditioned by the points \( \{X_t\}_{t=0}^n \).

### 3.2.3 Alternative diffusion bridge samplers

#### Independence sampler

The independence sampler is another special case of the Metropolis-Hastings algorithm whereby the new proposals are sampled independently of the currently accepted proposals. It differs from Algorithm 5 in that step 3.(b) is replaced by setting the proposal \( y' \) equal to the sample \( \omega \) from the Brownian bridge obtained in step 3.(a).

This independence sampler is easier to implement than the random walk Metropolis-Hastings sampler, but its acceptance rates suffer from the independence between accepted and newly proposed states especially if the target diffusion bridge process has a complicated drift or if the process is multi-dimensional. For example, consider the trajectory of the 2-dimensional toy diffusion process in Figure 3.1. A correctly realised diffusion bridge between any two points will look like an arc corresponding to the rotation enforced by the drift, especially over longer time periods, whereas a draw from the Brownian bridge will prioritise a more direct path between the two points. Without guidance from the previously accepted proposals, this lack of exploration will be compounded for higher dimensions.

![Figure 3.2: Computational times of our implemented Bayesian inference components for the capuchin monkeys data in Section 3.9.1. These components involve the diffusion bridge sampler, estimation of the constant diffusivity and non-parametric estimation of the drift.](image)

The random walk Metropolis-Hastings sampler construes an improvement over
the independence sampler and is fairly straightforward to implement, which is why we use this method as part of our implementation of the overall Bayesian inference methodology. However, let us consider Figure 3.2 which shows the computational costs of each of the components within our Bayesian inference methodology. The computational cost associated with the diffusion bridge sampler is slightly higher than that of the diffusivity inference component from Section 3.4 but lower than that of the drift inference component from Sections 3.5 and 3.6. For an algorithm as simple as the random walk Metropolis-Hastings to have such a high computational cost suggests that there is room for improvement.

**Guided proposals**

The guided proposals method is based once again on the Metropolis-Hastings algorithm, where the focus is on constructing proposals $X^\circ$ which sample from diffusion processes which are solutions to stochastic differential equations of a similar form to (3.3) of the $(0, x_0, T, v)$-diffusion bridge process. In particular, a guidance term is introduced in the drift to drive the trajectory towards the desired end-point.

Delyon and Hu [42], via generalising Clark’s [31] original method, considered the following proposals

\[
dX^\circ_t = \left( \frac{v - X^\circ_t}{T - t} \right) dt + \sigma dW_t, \tag{3.10}
\]

and

\[
dX^\circ_t = \left( b(X^\circ_t) + \frac{v - X^\circ_t}{T - t} \right) dt + \sigma dW_t, \tag{3.11}
\]

which gives us a sense of how the drift term of the proposal process can be manipulated to drive the trajectory towards an end-point $v$ at time $T$.

However, these two proposals have some limitations. If the difference between the drift of (3.10) and that of the target bridge is non-zero, it can lead to bad acceptance rates within the sampler. (3.10) and (3.11) can also fail to capture the true dynamics of the target bridge either by arriving in the proximity of the end-point $v$ too early, or by being unable to capture the multi-modality of the marginal
distributions of the target bridge.

Schauer et al. [101] instead proposed sampling from a process with a drift coefficient that guides the trajectory of its realisation to hit the designated end-point \( v \) at time \( T \) in a way that approximates the \( \nabla_x \log(p(v, T | X^*_t, t)) \) term in (3.3). Since the transition density \( p(v, T | X_t, t) \) is generally unknown, a known transition density \( \tilde{p}(v, T | X_t, t) \) is used instead. This leads the proposal diffusion process \( X^\circ \) given as the solution to

\[
\text{d}X^\circ_t = (b(X^\circ_t) + \sigma \sigma^T \nabla_x \log(\tilde{p}(v, T | X^\circ_t, t))) \text{d}t + \sigma \text{d}W_t, \tag{3.12}
\]

where the terms \( b \) and \( \sigma \) remain identical to that of (3.3). The new transition density \( \tilde{p}(v, T | X_t, t) \) comes from a chosen auxiliary diffusion process \( \tilde{X} \) specified as the solution to the stochastic differential equation

\[
\text{d}\tilde{X}_t = \tilde{b}(\tilde{X}_t) \text{d}t + \tilde{\sigma} \text{d}W_t.
\]

We thus have \( X \) the original diffusion process, \( X^* \) the diffusion bridge process, \( X^\circ \) the proposal process to approximate the bridge and \( \tilde{X} \) the auxiliary process whose transition density is used to construct the proposal process. Assuming that such a proposal process \( X^\circ \) can be constructed, the target bridge process \( X^* \) can then be sampled from since under certain assumptions (see Schauer et al. [101] Theorem 1, p. 11) the laws of the processes \( X^* \) and \( X^\circ \) are equivalent.

Additionally, a linear diffusion process \( \tilde{X} \) governed by the stochastic differential equation

\[
\text{d}\tilde{X}_t = \left( \tilde{B}(t)\tilde{X}_t + \tilde{\beta}(t) \right) \text{d}t + \tilde{\sigma} \text{d}W_t, \tag{3.13}
\]

with non-random matrix and vector coefficients \( \tilde{B}, \tilde{\beta} \) and \( \tilde{\sigma} \) has law that is equivalent to that of \( X^\circ \) and therefore that of the target bridge \( X^* \). This gives a large class of diffusion processes which can be used to construct the proposal process since the transition densities of linear diffusion processes are tractable.

As such, we believe Schauer et al’s guided proposals method to be a potential candidate to replace the current implementation in Section 3.2.2. This should lead to
3.2. Partially observed diffusion processes

A reduction of computational costs which would scale with dimensions when sampling for multi-dimensional diffusion bridge processes. For additional information on the sampling of diffusion bridge processes, we refer readers to the elucidation of problems and techniques by Papaspiliopoulos and Roberts [82].
3.3 Reparameterisation of diffusion processes

Let us now suppose that we have access to a high frequency, approximately continuous trajectory, as a result of the diffusion bridge augmentation technique described in Section 3.2 applied to a low frequency trajectory \( \{X_{t_i}\}_{i=0}^n \).

If we were to perform inference for the diffusivity on this augmented high frequency trajectory however, we would arrive at a degenerate posterior distribution for the diffusivity. This occurs because if we were to set a prior on the diffusivity, information about the diffusivity is provided to the posterior through the quadratic variation of the diffusion process \( X \) which satisfies

\[
[X,X]_T := \mathbb{P} - \lim_{\|\Pi_n\| \to 0} \sum_{i=1}^n (X_{t_i} - X_{t_{i-1}})(X_{t_i} - X_{t_{i-1}})^T = T \sigma \sigma^T, \tag{3.14}
\]

with partition \( \Pi_n := \{t_0 = 0 < \ldots < t_n = T\} \).

As the frequency of the trajectory increases, the variance of the posterior distribution of the diffusivity conditioned on the trajectory tends towards zero and this results in a distribution that is essentially a point mass on the value of \( \sigma \sigma^T \) initially obtained from the quadratic variation. Another way to view this problem is to note that the probability measures of any two diffusion processes with differing diffusivity coefficients are mutually singular and thus the Radon-Nikodym derivative used to transform between the two measures does not exist. Markov chain Monte Carlo algorithms with acceptance ratios constructed using said Radon-Nikodym derivatives will therefore exhibit extremely poor mixing.

Thus, to construct a non-degenerate posterior distribution for the diffusivity \( \sigma \), the dependence between the high frequency trajectory and the diffusivity has to be transformed in a way such that a valid Radon-Nikodym derivative can be found.

3.3.1 Reparameterisation

Roberts and Stramer’s [95] innovation was to transform the high frequency trajectory using the currently accepted estimate of the diffusivity into a new trajectory that has unit diffusivity, while retaining the quadratic variation of the underlying diffusion process.
3.3. Reparameterisation of diffusion processes

Let us introduce here $V : [0, T] \rightarrow \mathbb{R}^d$ the $d$-dimensional diffusion process specified as the solution to the stochastic differential equation

$$dV_t = \alpha(V_t)dt + dW_t, \quad V_0 = v_0,$$

with unit diffusivity, drift $\alpha : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $d$-dimensional Wiener process $W$ and transition density $p_V(x, t|y, s)$. The Lamperti transformation described in Definition 6 for constant $\sigma$ as applied to the diffusion process $X$ of (3.1) is simply the straightforward transformation

$$\eta(x; \sigma) = \sigma^{-1}x,$$

and by Itô’s formula the process $V = \eta(X; \sigma)$ solves

$$dV_t = \sigma^{-1}b(\sigma V_t)dt + dW_t$$

for drift $\alpha(v) = \sigma^{-1}b(\sigma v)$. We also have the inversion $b(x) = \sigma \alpha \{ \sigma^{-1}x \}$ to regain the original drift $b$ of $X$.

As before, we denote the low frequency trajectory as $\{X_t\}_{t=0}^n$. Let us now assume that we have access to the high frequency trajectory $V$, as the transformation $V = \eta(X_{\text{aug}}; \sigma)$, where $X_{\text{aug}}$ is the high frequency trajectory obtained by augmenting $\{X_t\}_{t=0}^n$ with diffusion bridges as in Section 3.2. Next, the centering transformation is given as

$$Z_t = V_t - \frac{t_t - t}{t_t - t_{t-1}} V_{t_{t-1}} - \frac{t - t_{t-1}}{t_t - t_{t-1}} V_{t_{t-1}}, \quad t_{t-1} \leq t \leq t_t,$$

with points $V_t = \eta(X_t)$. We can denote this centering transformation (3.16) by $Z = g(V)$ and its inversion by $V = h(Z; \sigma, \{X_t\})$. Each $i$-th segment $Z_s$, where $t_{i-1} \leq s \leq t_i$, is a diffusion bridge of unit diffusivity. The overall trajectory $Z$ will then be used in the posterior distribution of the diffusivity alongside the low frequency $\{X_t\}_{t=0}^n$ instead of the augmented trajectory $V$. This transformation of the trajectory to $Z$ results in a dominating measure that is independent of $\sigma$, which then allows the construction of the posterior of the diffusivity in Section 3.4.
Joint law and posterior densities

We will now write down the relevant laws and posterior densities required throughout this thesis (see Roberts and Stramer [95] and Papaspiliopoulos et al. [81] for more details).

Let us denote by $\mathbb{B}_{\sigma=1}(s,y;t,s)$ the standard Brownian bridge measure, $\text{Leb}_n$ the n-dimensional Lebesgue measure, $\{X_i\}_{i=0}^n = \{X_i\}_{i=0}^n$ the low frequency path, $V$ the augmented path of $\{X_i\}_{i=0}^n$ transformed to unit diffusivity, $V^{(i)} := \{V_s\}_{s \in [t_{i-1}, t_i]}$ the bridge segments of $V$, $Z$ the reparameterisation of $V$, $P_V(x,t|y,s)$ the transition density of $V$, $q(x,t|y,s)$ the transition density of Brownian motion, $P_V$ the probability measure of $V$, $W$ the Wiener measure, and $I(\alpha)$ the negative log-density between $P_V$ and $W$, so that

$$I(\alpha) = \frac{1}{2} \int_0^T |\alpha(V_t)|^2 \, dt - \int_0^T \alpha(V_t) \, dV_t. \quad (3.17)$$

We have the joint law

$$\mathbb{P}(Z, \{X_i\}, \alpha, \sigma) = \mathbb{P}(Z|\{X_i\}, \alpha, \sigma)\mathbb{P}(\{X_i\}|\alpha, \sigma)\mathbb{P}(\alpha)\mathbb{P}(\sigma), \quad (3.18)$$

where the independence of diffusion bridges gives

$$\mathbb{P}(Z|\{X_i\}, \alpha, \sigma) = \bigotimes_{i=1}^n \mathbb{P}(Z^{(i)}|X_{i-1}, X_i, \alpha, \sigma),$$

and the Markov property gives

$$\frac{d\mathbb{P}(\{X_i\}|\alpha, \sigma)}{d\text{Leb}_n} = \prod_{i=1}^n \sigma^{-1} p_{V}(V_{i}, t_{i}|V_{i-1}, t_{i-1}).$$

It can be derived from Papaspiliopoulos and Roberts [82] that

$$\frac{d\mathbb{P}(Z^{(i)}|X_{i-1}, X_i, \alpha, \sigma)}{d\mathbb{B}_{\sigma=1}(t_{i-1}, 0; t, 0)} = \frac{q(V_{i}, t_{i}|V_{i-1}, t_{i-1})}{p_{V}(V_{i}, t_{i}|V_{i-1}, t_{i-1})} \exp\{-I(\alpha; V^{(i)}|t_{i-1}, t_i)\}.$$
The posterior of the drift $\alpha$ can thus be written directly as

$$P(\alpha|Z, \{X_i\}, \sigma) = P(\alpha|V),$$

since $V$ encapsulates information about $Z, \{X_i\}$ and $\sigma$. For the augmented path $V$ we have

$$P(V|\{X_i\}, \alpha, \sigma) = \prod_{i=1}^{n} P(V^{(i)}|\alpha, V_{i-1}, V_i),$$

by the independence of diffusion bridges.

**Composite Bayesian inference methodology**

The composite Bayesian inference methodology for estimating the drift and diffusivity of a diffusion process in this thesis can thus be boiled down to sampling from the joint law $P(Z, \alpha, \sigma|\{X_i\})$, which can in turn be achieved by the use of a Metropolis-within-Gibbs algorithm (see Section 2.4):

**Algorithm 6. Composite Bayesian inference methodology**

1. Target: The $P(Z, \alpha, \sigma|\{X_i\})$ distribution.

2. Initialisation: Set $Z^{(0)}, \alpha^{(0)}, \sigma^{(0)}$ at some initial value.

3. Iteration $j+1$: Current state $Z^{(j)}, \alpha^{(j)}, \sigma^{(j)}$.

   (a) Sample $\sigma^{(j+1)}$ from $P(\sigma|Z^{(j)}, \{X_i\}, \alpha^{(j)})$ and set $V = h(Z^{(j)}; \alpha^{(j)}, \{X_i\}, \sigma^{(j+1)})$.

   (b) Sample $\alpha^{(j+1)}$ from $P(\alpha|Z^{(j)}, \{X_i\}, \sigma^{(j+1)}) = P(\alpha|V)$.

   (c) Sample $V$ from $P(V|\{X_i\}, \alpha^{(j+1)}, \sigma^{(j+1)})$ and set $Z^{(j+1)} = g(V)$.

4. The $Z, \alpha, \sigma$ samples then come from the $P(Z, \alpha, \sigma|\{X_i\})$ distribution.

To perform Step 3.(a) of Algorithm 5, we will show how to sample from $P(\sigma|Z, \{X_i\}, \alpha)$ in Section 3.4.
3.3. Reparameterisation of diffusion processes

For Step 3.(b), we note that since we have a direct relation between $\alpha$ and $b$, we can simulate $b$ instead from $P(b|\eta^{-1}(V; \sigma))$ and set $\alpha(v) = \sigma^{-1}b(\sigma v)$. By doing so we can proceed with Sections 3.5 and 3.6 where the high frequency $\eta^{-1}(V)$ is treated as an observation from $dX_t = b(X_t)dt + \sigma dW_t$, and sample from the posterior $p(b|X, \sigma)$ instead.

As for Step 3.(c), we have already shown in Section 3.2 how to sample diffusion bridges conditioned on a low frequency path.
3.4 Inference for constant diffusivity

3.4.1 Diffusivity inference

Let us suppose that we are again in the setup as described by Section 3.3.1, such that the augmented path $V$ and the reparameterised path $Z$ have been made available for the low frequency path $\{X_i\}_{i=0}^n$. Now assuming that the drift $b$, and therefore $\alpha$, is known, we can proceed to performing Bayesian inference for the unknown constant diffusivity $\sigma \in \mathbb{R}^{d \times d}$.

From (3.18), we can derive posterior of the diffusivity $\sigma$ as

$$
\mathbb{P}(\sigma|Z, \{X_i\}, \alpha) = \mathbb{P}(Z|\{X_i\}, \alpha, \sigma)\mathbb{P}(\{X_i\}|\alpha, \sigma)\mathbb{P}(\alpha)\mathbb{P}(\sigma),
$$

$$
\propto \left( \bigotimes_{i=1}^n \mathbb{P}(Z^{(i)}|X_{i-1}, X_i, \alpha, \sigma) \right) \mathbb{P}(\{X_i\}|\alpha, \sigma)\mathbb{P}(\alpha)\mathbb{P}(\sigma),
$$

$$
\propto \left( \bigotimes_{i=1}^n \mathbb{P}(Z^{(i)}|X_{i-1}, X_i, \alpha, \sigma) \right) \left( \prod_{i=1}^n \sigma^{-1} p_V(V_{i, t_i}|V_{i-1, t_i-1}) \right) \mathbb{P}(\sigma),
$$

$$
\propto \left( \prod_{i=1}^n \frac{q(V_{i, t_i}|V_{i-1, t_i-1})}{p_V(V_{i, t_i}|V_{i-1, t_i-1})} \exp\{-I(\alpha, V, t_i-1, t_i)\} \right) \left( \prod_{i=1}^n \sigma^{-1} p_V(V_{i, t_i}|V_{i-1, t_i-1}) \right) \mathbb{P}(\sigma),
$$

where $p_V(x, t|y, s)$ is the transition density of $V$, $I(\alpha)$ the negative log-density between $\mathbb{P}_\alpha$ and $\mathcal{W}$, given by (3.17), and $q(x, t|y, s)$ the transition density of Brownian motion such that

$$
\log \left( \prod_{i=1}^n \sigma^{-1} q(V_{i, t_i}|V_{i-1, t_i-1}) \right) = -\frac{1}{2} \sum_{i=1}^{n-1} \frac{|S^{-1}(X_{i+1} - X_i)|^2}{t_{i+1} - t_i} - \frac{n}{2} \log \left\{ 2\pi |\Sigma| (t_{i+1} - t_i) \right\}.
$$

We specify an Inverse Wishart prior for the diffusivity with scale parameter $\Psi$ and degrees of freedom $\nu$. Its density can be explicitly written as

$$
\mathbb{P}(\sigma) = \frac{|\Psi|^\frac{\nu}{2}}{2^{\nu d/2} I_d(\nu/2)} |\sigma|^{-\frac{\nu + d + 1}{2}} \exp \left\{ -\frac{1}{2} \text{trace}(\Psi \sigma^{-1}) \right\},
$$
3.4. Inference for constant diffusivity

with the multi-variate gamma function given as

$$\Gamma_d(v/2) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma \left( \frac{v}{2} + \frac{1-j}{2} \right).$$

The Inverse Wishart prior has support on the space of positive definite matrices, and we additionally choose a symmetric scale parameter $\Psi$ and degrees of freedom $\nu \geq 2$ to ensure the existence of a symmetric mean.

We can then implement a random walk Metropolis-Hastings algorithm to draw from the posterior distribution of the diffusivity $\sigma$, with proposals constructed as

$$\sigma' = \sigma^{(j)} + \rho R,$$

where $R \in \mathbb{R}^{d \times d}$ is a matrix given by $R = \frac{1}{2}(\bar{R} + \bar{R}^\top)$ to enforce symmetry, with random elements $\bar{R}_{i,j} \sim N(0,1)$, $i, j \in \{1, \ldots, d\}$, and $\rho$ controls the step size of the random walk. The acceptance ratio is then given by

$$\alpha(\sigma^{(j)}, \sigma') = \min \left\{ 1, \frac{p(\sigma'|Z, \{X_i\}, \alpha)}{p(\sigma^{(j)}|Z, \{X_i\}, \alpha)} \right\},$$

since the proposal density arising from the proposals (3.19) is symmetric.

The random walk Metropolis-Hastings algorithm to sample from the posterior distribution of $\sigma$ is therefore given by:

Algorithm 7. Random walk Metropolis-Hastings: diffusivity inference

1. Target: $p(\sigma|Z, \{X_i\}, \alpha)$

2. Initialisation: Set $\sigma^{(0)}$ at some initial value

3. Iteration $j + 1$: Given current state $\sigma^{(j)}$
   
   (a) Sample $\bar{R}_{i,j}$ from $\bar{R}_{i,j} \sim N(0,1)$
   
   (b) Set $R = \frac{1}{2}(\bar{R} + \bar{R}^\top)$
   
   (c) Set $\sigma' = \sigma^{(j)} + \rho R$
(d) Sample $\nu \sim \text{Uniform}[0,1]$ and update

$$
\sigma^{(j+1)} = \begin{cases} 
\sigma' & \text{if } \nu \leq \alpha(\sigma^{(j)}, \sigma') \\
\sigma^{(j)} & \text{otherwise}
\end{cases}
$$

with acceptance ratio

$$
\alpha(\sigma^{(j)}, \sigma') = \min \left\{ 1, \frac{p(\sigma' | Z, \{X_i\}, \alpha)}{p(\sigma^{(j)} | Z, \{X_i\}, \alpha)} \right\}
$$

4. The accepted $\{\sigma^{(j)}\}_{j=1}^n \}$ proposals then come from the distribution of $\mathbb{P}(\sigma | Z, \{X_i\}, \alpha)$.

We note that this algorithm is implemented in such a way that it will automatically reject non-positive definite proposals for $\sigma$, where the rejection is carried out before attempting to compute the acceptance ratio by testing for positive-definiteness using the trace and determinant.
3.5 Non-parametric drift inference: dimension one

Let us assume now that we have access to either a 1-dimensional continuous trajectory or a high frequency, approximately continuous trajectory. We fit to this trajectory a 1-dimensional diffusion process model $X: [0, T] \rightarrow \tau$ specified as the solution to the stochastic differential equation

$$dX_t = b(X_t)dt + \sigma dW_t, \quad X_0 = x_0,$$

(3.20)

with unknown drift $b: \tau \rightarrow \mathbb{R}$ and known, constant diffusivity $\sigma \in \mathbb{R}$, where $W$ is the 1-dimensional Wiener process. In this section, we are interested in performing Bayesian inference for the unknown drift $b$.

We note that we can consider two parameterisations of the diffusion process model. The first is simply the diffusion process as specified by (3.20) where we perform inference for the drift $b$. The second involves the transformation of this process into one with unit diffusivity

$$dV_t = \alpha(V_t)dt + dW_t,$$

as described by the reparameterisation technique in Section 3.3. The relationship between the two is given as $V_t = \eta(X_t; \sigma)$ where $\eta(x; \sigma) = \sigma^{-1}x$ and $\alpha(v) = \sigma^{-1}b(\sigma v)$. Inference is performed instead on the drift $\alpha$, which can then be easily transformed back to $b$. This parameterisation of $V$ is convenient since the drift inference step in the Metropolis-within-Gibbs sampler-based composite Bayesian inference methodology described in Section 3.3 uses the posterior $P(\alpha|V)$ without requiring additional work. However, the dependence of both the transformation $\eta$ and $\alpha$ on the diffusivity $\sigma$ may lead to a stronger correlation between $\alpha$ and $\sigma$, which in turn may result in poor mixing of the relevant Markov chains. Instead, we can include an additional transformation step involving $\eta$ in the Gibbs sampler to allow us to focus wholly on the parameterisation $X$ in (3.20) and perform inference on the drift $b$.

We additionally note here that for the purposes of drift inference the constant
diffusivity $\sigma$ is considered known since it can be provided via inference for constant diffusivity coefficients in Section 3.4. A high frequency trajectory can easily be constructed from a low frequency one using the diffusion bridge-based data augmentation technique described in Section 3.2.

In following subsections we will outline Papaspiliopoulos et al.’s [81] non-parametric inference methodology for the drift of the 1-dimensional diffusion process model as described by (3.20). In Section 3.5.1 the local time process is used to transform the time and stochastic integrals in the likelihood of the drift into a single encompassing spatial integral, which has a quadratic form with respect to the drift $b$. Using this likelihood a conjugate prior for the drift can then be specified in Section 3.5.2 as a Gaussian measure on the $L^2$ square-integrable function space. The precision operator of this Gaussian measure is specified in Section 3.5.3. The resultant posterior distribution described in Section 3.5.4 is likewise a Gaussian measure on the $L^2$ square-integrable function space.

### 3.5.1 The likelihood and local time

Papaspiliopoulos et al. [81] considered firstly the likelihood for the drift $b$ of (3.20), and assumed $b$ to be non-explosive and sufficiently regular such that at least the derivative $b'$ exists. Let $P_\sigma$ be the law of the diffusion process $X$ and $W_\sigma$ be the corresponding $\sigma$-Wiener measure. The Radon-Nikodym derivative of $P_\sigma$ with respect to $W_\sigma$ is given by

$$
\frac{dP_b}{dW_\sigma} = \exp \left\{ -\frac{1}{2} \int_0^T \sigma^{-2} |b(X_t)|^2 dt + \int_0^T \sigma^{-2} b(X_t) dX_t \right\}.
$$

We can denote $I(b)$ as the negative log-likelihood for the drift $b$ such that

$$
I(b) := \sigma^{-2} \left( \frac{1}{2} \int_0^T |b(X_t)|^2 dt - \int_0^T b(X_t) dX_t \right).
$$

(3.21)

Now denote by $B(x) = \int^x b(u)du$ the anti-derivative of $b$ and apply Itô’s formula to obtain

$$
dB(X_t) = b(X_t) dX_t + \frac{1}{2} b'(X_t) dt,
$$
which gives

\[ I(b) = \sigma^{-2} \left( \frac{1}{2} \int_0^T \{ |b(X_t)|^2 + b'(X_t) \} \, dt - B(X_T) + B(X_0) \right). \quad (3.22) \]

Recall the local time process \( L_t \) described by Definition [1], which we re-iterate here as

\[ L_t(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{2\varepsilon} \int_0^t \mathbb{1}_{\{ X_s \in (x - \varepsilon, x + \varepsilon) \}} \, ds, \quad \forall t \in [0, T], \]

where for any Borel measurable and bounded function \( f \) the following equality holds

\[ \int_0^t f(X_s) \, ds = \int_{-\infty}^{\infty} f(u) L_t(u) \, du. \quad (3.23) \]

A change of variables from the time index to the state space of the diffusion process can then be performed for (3.22) so that

\[ I(b) = \sigma^{-2} \left( \frac{1}{2} \int_{-\infty}^{\infty} \{ |b(u)|^2 L_T(u) - 2\chi(u)b(u) + b'(u)L_T(u) \} \, du \right), \quad (3.24) \]

where

\[ \chi(u) = \begin{cases} 
1, & X_0 < u < X_T, \\
-1, & X_T < u < X_0, \\
0, & \text{otherwise.}
\end{cases} \]

Ideally, we would want to perform an integration by parts on (3.24) to arrive at a quadratic form in \( b \). However, this is complicated by the regularity of the local time process, wherein it is continuous but not differentiable. However, to motivate the choice of prior for the drift, it is assumed here that the derivative \( L' \) does indeed exist. A formal calculation then yields the quadratic form

\[ I(b) = \sigma^{-2} \left( \frac{1}{2} \int_{-\infty}^{\infty} \{ |b(u)|^2 L_T(u) - 2\chi(u)b(u) + \frac{1}{2} L'_T(u) \} \, du \right), \quad (3.25) \]

where a simple rearrangement of (3.25) with a further restriction on the integral
3.5. Non-parametric drift inference: dimension one

gives

\[ I(b) \propto \sigma^{-2} \left( \frac{1}{2} \int_{\min(X_t)}^{\max(X_t)} L_T(u) \left( b(u) - \frac{\chi(u) + \frac{1}{2}L_T'(u)}{L_T(u)} \right)^2 du \right), \]

since the local time process has values zero beyond the range of the trajectory of \( X \).

3.5.2 The prior distribution

The representation of the log-likelihood \( I(b) \) in (3.25) thus suggests the use of a conjugate Gaussian prior distribution. This is given as a Gaussian measure on function spaces specified by its mean \( m_0 \) and precision operator \( A_0 \). We define Gaussian measures in Section 2.2 and refer to da Prato and Zabczyk [38] for additional details. Since the properties of the drift \( b \) are determined by its prior, and are carried forward into the posterior, we will focus here on choosing an appropriate mean \( m_0 \) and precision operator \( A_0 \) so that functions sampled from the associated Gaussian measure will have desirable properties.

Recall the beginning of Section 3.5.1 where the drift \( b \) was assumed to be non-explosive and sufficiently regular such that \( b' \) exists. Suppose now that the drift \( b \) is fitted with a prior distribution specified as the Gaussian measure on \( L^2(\tau, \mathbb{R}) \) with precision operator \( A_0 \) with domain \( D(A_0) \subset L^2 \). The choice of precision operator here is what guarantees the non-explosiveness and regularity of \( b \).

To see why, we turn to Hairer et al. [65] who demonstrated how certain conditioned diffusion processes have unique invariant measures which coincide with Gaussian measures. Consider the diffusion process \( Y : [0, 1] \to \mathbb{R} \) specified by the stochastic differential equation conditioned with a fixed starting point

\[ dY_u = AY_u du + BdW_u, \quad Y_0 = y_0, \tag{3.26} \]

where \( A, B \in \mathbb{R} \) and \( W \) is the 1-dimensional Wiener process. This diffusion process \( Y \) is in fact a Gaussian process with mean

\[ \mathbb{E}[Y_u] = \exp\{uA\}y_0, \]
3.5. Non-parametric drift inference: dimension one

and covariance function

\[ C(u, v) = \exp\{(u + v)A\} \left( \int_0^{u\wedge v} \exp\{-2rA\}B^2 dr \right), \quad \forall u, v \in [0, 1]. \]

The precision operator is given as the second order differential operator \( A \), where

\[ A = B^{-2} \left( \frac{d}{du} + A \right)^2, \]

with domain \( \mathcal{D}(A) = \{ f \in L^2([0, 1], \mathbb{R}) | f(0) = 0, \frac{d}{du}f(1) = Af(1) \} \). It follows from [65, Lemma 3.2, p. 6] that the function \( C \) is the Green’s function (see for example Duffy [44]) for \(-A\). This result is generalised for the multi-dimensional case as [65, Theorem 3.3, p. 6], which is also available for diffusion processes conditioned with a Gaussian starting point as [65, Theorem 3.4, p. 7] and for diffusion bridges as [65, Theorem 3.6, p. 8].

For example, consider the stationary 1-dimensional Ornstein-Uhlenbeck process \( \gamma : [0, 1] \to \mathbb{R} \) specified as the solution to the stochastic differential equation

\[ d\gamma_u = -\left( \frac{\varepsilon}{\eta} \right)^{1/2} \gamma_u du + \eta^{-1/2}dW_u, \quad \gamma_0 \sim N(0, 2^{-1}(\eta \varepsilon)^{-1/2}), \quad (3.27) \]

where \( \varepsilon \geq 0 \) and \( \eta > 0 \). The invariant measure of this Ornstein-Uhlenbeck process is a Gaussian measure (see Karatzas and Shreve [71]) with zero mean and covariance function given by

\[ C(u, v) = (2\eta(\varepsilon/\eta)^{1/2})^{-1} \exp\{-\left( \frac{\varepsilon}{\eta} \right)^{1/2}|u - v|\}, \quad \forall u, v \in [0, 1], \]

which is the Green’s function of the precision operator

\[ A = -\eta \frac{d^2}{du^2} + \varepsilon = -\eta \Delta + \varepsilon, \]

with domain \( \mathcal{D}(A) = \{ f \in H^2([0, 1], \mathbb{R}) | \frac{d}{du}f(0) = ( \varepsilon/\eta )^{1/2}f(0), \frac{d}{du}f(1) = -(\varepsilon/\eta)^{1/2}f(1) \} \), where \( \Delta \) is the Laplacian operator and \( H^2([0, 1], \mathbb{R}) \) is the Sobolev space of functions with weak derivatives of orders up to 2 (see Definition
Hairer et al. [64] also showed that higher order differential operators can arise as precision operators for diffusion processes specified as solutions to conditioned, linear stochastic differential equations with degenerate diffusion matrices, for example the integrated Brownian motion. Following this reasoning, to encompass processes with stronger regularity, a precision operator can be chosen as

$$A = \eta (-1)^k \frac{d^{2k}}{du^{2k}} + \epsilon = \eta (-\Delta)^k + \epsilon, \quad \eta > 0, \epsilon \geq 0, k \in \mathbb{N},$$

(3.28)

with domain $D(A) \subset H^{2k}_{\text{per}}(\tau, \mathbb{R})$ where $H^{2k}_{\text{per}}(\tau, \mathbb{R})$ is the Sobolev space of functions with weak derivatives of orders up to $2k$ and periodic boundary conditions, which is enforced by the torus $\tau$. This choice of Sobolev space will simplify matters later in Section 3.5.4, where the periodic boundary conditions allow the posterior mean of the drift to be found as a solution of the boundary-value problem (3.37).

By setting a Gaussian measure with precision operator (3.28) as the prior for the drift $b$, its regularity can be guaranteed. Since the mean $m_0$ of the Gaussian measure encodes information about the shape of the drift $b$, a simple and safe choice for a prior would be to set $m_0 = 0$. The prior for the drift $b$ is thus specified as the Gaussian measure on $L^2(\tau, \mathbb{R})$ with mean $m_0$ and precision operator $A_0 = \eta (-\Delta)^k + \epsilon$, $\eta > 0, \epsilon \geq 0, k \in \mathbb{N}$ with domain $D(A_0) = H^{2k}_{\text{per}}(\tau, \mathbb{R})$. We can think of the index $u$ as used in (3.26), (3.27) and (3.28) as a spatial index instead of the usual time index, and it can be seen that the drift $b$ will behave as a conditioned diffusion process with a degenerate diffusivity coefficient indexed by the range of the diffusion process $X$ specified by (3.20).

3.5.3 The precision operator

To further illustrate the properties of the precision operator $A_0$ and the properties of the samples $b$ drawn from the corresponding Gaussian measure $N(m_0, A_0)$, Pokern et al. [88] illustrated the following train of thought where the Gaussian measure can be represented as the law of a centred Gaussian process.
The domain of the precision operator is stated as

\[ \mathcal{D}(A_0) \subset H^{2k}(\tau, \mathbb{R}) = \{ y : D^\alpha y \in L^2(\tau, \mathbb{R}), \quad \forall 0 \leq |\alpha| \leq 2k \}, \]

where \( D^\alpha \) is the \( \alpha \)-th order differential operator (see Definition 16), and \( \tau \in \mathbb{R} \).

Then denote by \( \phi_j \) an orthonormal basis of \( L^2(\tau, \mathbb{R}) \) given as

\[ \begin{align*}
\phi_{2j}(x) &= \sqrt{2} \cos(2\pi jx), \\
\phi_{2j-1}(x) &= \sqrt{2} \sin(2\pi jx), \\
j &\in \mathbb{N}.
\end{align*} \tag{3.29} \]

Now consider the covariance operator \( C_0 \) with its inverse given by

\[ A_0 = \eta(-\Delta)^k + \epsilon, \quad \eta > 0, \epsilon \geq 0, k \in \mathbb{N}, \]

such that

\[ A_0 \phi_{2j} = (\eta(4\pi^2 j^2)^k + \epsilon) \phi_{2j}, \]

\[ A_0 \phi_{2j-1} = (\eta(4\pi^2 j^2)^k + \epsilon) \phi_{2j-1}. \]

This shows that the covariance operator \( C_0 \) is the operator on \( L^2(\tau, \mathbb{R}) \) diagonalised by the basis functions \( \phi_j \), with eigenvalues given by

\[ \lambda_j = \left( \eta \left( 2\pi \left\lfloor \frac{j}{2} \right\rfloor \right)^{2k} + \epsilon \right)^{-1}. \tag{3.30} \]

The Gaussian measure with precision operator \( A_0 \) is therefore the law of the centred Gaussian process \( b \) defined by

\[ b(x) = \sum_{j \in \mathbb{N}} \sqrt{\lambda_j} \phi_j(x) Z_j, \]

via the Karhunen-Loève expansion where \( Z_j, j \in \mathbb{N} \) are i.i.d. standard Gaussian random variables.

It can be seen from (3.30) that the \( k, \eta \) and \( \epsilon \) terms, and therefore the selfsame terms in (3.28), can be adjusted to further characterise the Gaussian measure. The \( k \) term dictates the regularity of samples from the Gaussian measure, while \( \eta \) and \( \epsilon \)
3.5. Non-parametric drift inference: dimension one

together control the speed of mean reversion to the mean.

**Hyper-parameters: the prior**

First of all, we focus on the regularity term $k$. For the case when $k = 1$, the precision operator $\mathcal{A} = -\eta \Delta + \epsilon$ is exactly the precision operator obtained from the stationary Ornstein-Uhlenbeck process defined by (3.27). We can see that given a fixed regularity term $k = 1$ and fixed $\eta$, the $\epsilon$ term controls the speed of mean reversion to the prior mean. Whereas for fixed $k = 1$ and fixed $\epsilon$, the $\eta$ term determines the overall scale of functions drawn from the prior in terms of how close to the prior mean it will remain. The regularity term $k$ manifests by controlling how smooth the functions drawn from the prior will be. These almost sure effects of $\eta$ and $k$ carry on to the posterior since the prior and posterior measures are absolutely continuous.

Let us pretend now that we can sample from Gaussian measures (see Section 3.6.5) on $L^2(\tau, \mathbb{R}^d)$. In particular, we set $d = 2$ so that the samples correspond to drifts from a 2-dimensional diffusion process. To further visualise the effects of how the hyper-parameters $k$, $\eta$ and $\epsilon$ affect the samples, we will draw samples from the Gaussian measure characterised by mean zero and precision operator given by

$$
\mathcal{A}_0 = \eta (-\Delta)^k + \epsilon, \quad \eta > 0, \epsilon \geq 0, k \in \mathbb{N}.
$$

for different hyper-parameter values.

In Figure 3.5 the hyper-parameters $\eta$ and $\epsilon$ are set to the fixed values of $\eta = 5 \times 10^{-5}$ and $\epsilon = 5 \times 10^{-3}$, while the regularity term $k$ is varied over the values 1, 2, 3 and 4. As expected, as $k$ increases the draws become smoother, and in the case of $k = 4$ the draw becomes very smooth, with only very slight perturbations away from the zero mean.

In Figure 3.4 the hyper-parameters $k$ and $\epsilon$ are set to the fixed values of $k = 3$ and $\epsilon = 5 \times 10^{-3}$, while the $\eta$ is varied over the values $5 \times 10^{-l}$, for $l \in 3, 5, 7, 9$. As $\eta$ decreases in value, the speed of mean reversion increases, but we see an increase in scale of perturbations as well.

In Figure 3.3 the hyper-parameters $k$ and $\eta$ are set to the fixed values of $k = 3$,
3.5. Non-parametric drift inference: dimension one

Figure 3.3: The $x$-component of samples from the mean zero Gaussian measure with precision operator $A_0 = \eta (-\Delta)^k + \epsilon$, for fixed $\eta = 5 \times 10^{-5}$, $\epsilon = 5 \times 10^{-3}$, and for varying $k = 1, 2, 3$ and 4 (a,b,c,d).

$\eta = 5 \times 10^{-7}$, while $\epsilon$ is varied over the values $5 \times 10^{-l}$, for $l \in 1, 3, 5, 7$. When $\epsilon$ becomes larger, the speed of mean reversion increases while the scale of the samples decreases.
Figure 3.4: The x-component of samples from the mean zero Gaussian measure with precision operator $A_0 = \eta (-\Delta)^k + \varepsilon$, for fixed $k = 3$, $\varepsilon = 5 \times 10^{-3}$, and for varying $\eta = 5 \times 10^{-l}$, where $l = 3, 5, 7$ and 9 (a,b,c,d).
Figure 3.5: The $x$-component of samples from the mean zero Gaussian measure with precision operator $A_0 = \eta(-\Delta)^k + \epsilon$, for fixed $k = 3$, $\eta = 5 \times 10^{-7}$ and for varying $\epsilon = 5 \times 10^{-l}$, where $l = 1, 3, 5$ and 7 (a,b,c,d).
3.5. Non-parametric drift inference: dimension one

3.5.4 The posterior distribution

As per the usual method of deriving the parameters of a conjugate Gaussian distribution, the mean $m_1$ and precision operator $A_1$ of the posterior Gaussian measure can be found by completing the square on the log-likelihood (3.25) and the prior Gaussian measure to give the prior-posterior update equations

$$(A_0 + L_T)m_1 = A_0m_0 + \chi + \frac{1}{2}L_T', \quad (3.32)$$

and

$$A_1 = A_0 + L_T. \quad (3.33)$$

With our choice of prior in Section 3.5.2 for the drift as the Gaussian measure on $L^2(\tau, \mathbb{R})$ with mean $m_0 = 0$ and precision operator $A_0 = \eta(-\Delta)^k + \epsilon = \eta(-1)^k \frac{d^{2k}}{du^{2k}} + \epsilon$ with domain $\mathcal{D}(A) = H^k_{\text{per}}(\tau, \mathbb{R})$, the prior-posterior update equations (3.32) and (3.33) can then be rewritten as

$$A_1m_1 = \left\{ \eta(-1)^k \frac{d^{2k}}{du^{2k}} + \epsilon \right\} m_0 + \chi + \frac{1}{2}L_T', \quad (3.34)$$

and

$$A_1 = \eta(-1)^k \frac{d^{2k}}{du^{2k}} + \epsilon + L_T. \quad (3.35)$$

and the precision operator $A_1$ of the posterior given as

$$A_1 = \eta(-1)^k \frac{d^{2k}}{du^{2k}} + \epsilon + L_T. \quad (3.36)$$

The equations (3.34) and (3.35) can be reformulated by taking their inner product with some test function $y \in H^k_{\text{per}}$, which gives (3.34) as

$$r(y) = \int_{\tau} \left\{ \eta \frac{d^k m_0}{du^k} (u) \frac{d^k y}{du^k} (u) + \epsilon m_0(u) y(u) + \chi(u) y(u) - \frac{1}{2} L_T(u) y'(u) \right\} du,$$

where the problematic $L_T'$ term has been removed via an integration by parts. Simi-
3.5. Non-parametric drift inference: dimension one

larly, (3.55) is given as

$$a(x, y) := \int_{\tau} \left\{ \eta \frac{d^{k}x}{du^{k}}(u) \frac{d^{k}y}{du^{k}}(u) + \varepsilon x(u)y(u) + L_T(u)x(u)y(u) + \varepsilon x(u)y(u) + L_T(u)x(u)y(u) \right\} du.$$  

with \( x \in H^k \). Together with the periodic boundary conditions of \( H^k_{\text{per}} \), \( a(x, y) = r(y) \) is then considered a boundary-value problem, as described by Definition 18.

We reproduce here a theorem from Papaspiliopoulos et al. [81], the proof of which can be found in the paper by Pokern et al. [88], which not only guarantees the existence and conjugacy of the Gaussian measure posterior for the drift, but also allows one to obtain the posterior mean \( m_1 \) as a weak solution to the boundary-value problem

$$a(x, y) = r(y). \quad (3.37)$$

**Theorem 2.** [81, Theorem 1, p. 519] Consider a prior Gaussian measure \( \mu_0 = N(m_0, C_0) \) with mean \( m_0 \in H^l_\text{per} \) with \( l \geq k \) and precision operator \( A_0 \) specified as in (3.28) with domain \( H^k_{\text{per}} \). Then the posterior measure for the drift \( b \in L^2 \), given a sample path \( X \), is a Gaussian measure \( \mu_1 = N(m_1, C_1) \). The mean \( m_1 \) is the unique weak solution of (3.37) where the precision operator given by (3.36) has domain \( H^{2k}_{\text{per}} \), the mean itself is an element of \( H^s_{\text{per}} \) for \( s = \min\{l, 2k - 1/2 - \varepsilon\} \) for any \( \varepsilon > 0 \). Furthermore, \( \mu_1 \) and \( \mu_0 \) are equivalent.

Solving for the posterior mean \( m_1 \) then becomes a numerical problem, which Papaspiliopoulos et al. [81] tackle by using a finite element method, where the boundary-value problem (3.37) is represented via a Galerkin approximation using a basis with piecewise third-order polynomials.

**Hyper-parameters: the posterior**

As an extension to the visualisations in Section 3.5.3 to examine the effects of the hyper-parameters \( \eta \) and \( \varepsilon \) of the precision operator \( A_0 = \eta (-\Delta)^k + \varepsilon I \) on samples from the Gaussian measure prior, we will now examine the effects of said hyper-parameters on samples from the corresponding Gaussian measure posterior.
3.5. Non-parametric drift inference: dimension one

We draw a sample path from the 2-dimensional toy diffusion process described in Section 2.1. Let us assume for the moment that we are able to perform drift inference for 2-dimensional diffusion processes (to be seen later in Section 3.6), such that a conjugate Gaussian measure prior for the drift is specified with a precision operator $A_0 = \eta (-\Delta)^k + \varepsilon$. The posterior precision operator will be of a form analogous to the 1-dimensional posterior precision operator

$$A_1 = \eta (-\Delta)^k + \varepsilon + L_T.$$ 

We then fix the regularity term $k = 2$ and perform drift inference on the sample path, with results displayed in Figures 3.6 and 3.7.

In Figure 3.6, the hyper-parameter $\varepsilon$ is set to a fixed value of $\varepsilon = 5 \times 10^{-3}$, while $\eta$ is varied over the values $5 \times 10^{-l}$, for $l \in 3,5,7,8$. We can see that $\eta$ alongside the $k$ term controls the regularity of the samples, such that when $\eta$ decreases in value, the samples become rougher. As with the samples from the prior in Section 3.5.3, the scale of the perturbations increases as well when $\eta$ decreases.

In Figure 3.7, the hyper-parameter $\eta$ is set to a fixed value of $\eta = 5 \times 10^{-7}$, while $\varepsilon$ is varied over the values $5 \times 10^{-l}$, for $l \in 2,3,4$. We see that as $\varepsilon$ decreases in value the scale of the samples increases while preserving its regularity, as in the case with the prior samples.

It can be seen from the posterior precision operator that decreasing either $\eta$ or $\varepsilon$ will lead to the information from the data provided through $L_T$ being weighted less compared to the prior precision operator. This is evidenced in both Figures 3.6 and 3.7, where the scales of the samples exceed the scale of the actual drift, which can be found in Figure 2.4.
Figure 3.6: The full (left column) and x-component (right column) of the posterior mean of the drift, with hyper-parameter $\eta$ set to $\eta = 5 \times 10^{-l}$, for $l = 3, 5, 7$, and 9 (top to bottom), for fixed $\varepsilon = 5 \times 10^{-3}$. The colour bar denotes the posterior variance.
Figure 3.7: The full (left column) and x-component (right column) of the posterior mean of the drift, with hyper-parameter $\varepsilon$ set to $\varepsilon = 5 \times 10^{-l}$, for $l = 2, 3$ and 4 (top to bottom), for fixed $\varepsilon = 5 \times 10^{-7}$. The colour bar denotes the posterior variance.
3.6 Non-parametric drift inference: dimension greater than one

In this section we will be generalising the non-parametric Bayesian drift inference methodology for 1-dimensional diffusion processes previously introduced in Section 3.5 for the multi-dimensional case.

Once again, let us assume that we have access to a continuously observed path. We fit to it a diffusion process model $X : [0, T] \to \tau$ characterised by the stochastic differential equation

$$dX_t = b(X_t)dt + \sigma dW_t, \quad X_0 = x_0,$$  

(3.38)

with $\tau$ a $d$-dimensional torus, unknown drift $b : \tau \to \mathbb{R}^d$, known positive-definite symmetric constant diffusivity $\sigma \in \mathbb{R}^{d \times d}$ and the $d$-dimensional Wiener process $W$.  

3.6.1 The prior distribution

The Gaussian conjugate prior

The drift $b$ is defined component-wise such that $b := (b_1, \ldots, b_d)^T$ where $b_l : \tau \to \mathbb{R}$, $l \in \{1, \ldots, d\}$. We can proceed as in the 1-dimensional case, and assume that the drift $b$ follows a Gaussian prior with mean function $b_0$ and precision operator $A_0 = \eta (-\Delta)^k + \varepsilon$ with domain $H^{2k}_{\text{per}}(\tau, \mathbb{R}^d)$, where the periodic boundary conditions are characterised by the torus $\tau$. The precision operator $A_0$ acts on $b$ as

$$A_0 b = (A_0 b_1, \ldots, A_0 b_d)^T,$$

so that $b$ is independent component-wise. Formally, this Gaussian prior can be represented by the density

$$\pi_0(b) \propto \exp \left[ -\frac{1}{2} \int_\tau (b - b_0)(u) \cdot A_0 (b - b_0)(u) du \right],$$  

(3.39)

where $\cdot$ denotes the standard inner product in $\mathbb{R}^d$. The prior mean $b_0$ is then set to zero barring additional prior knowledge, which helps ease further notation.
3.6. Non-parametric drift inference: dimension greater than one

Global Lipschitz condition

To show that the drift $b$ satisfies the global Lipschitz condition (see Definition 3), we can proceed similarly to Section 3.5.3 by considering the multivariate Karhunen-Loève expansion of $b$ (see Theorem 5.2.2 from [122, p. 204] and Remark 5.2.2 from [122, p. 207]).

Let us now consider the centered (zero-mean) Gaussian process $b : \tau \rightarrow \mathbb{R}^d$ which can be represented component-wise via the Karhunen-Loève expansion as

$$b_l(x) = \sum_{j \in \mathbb{N}^d} \sqrt{\lambda_j} \phi_j(x) Z_{j,l}, \quad \forall x \in \tau,$$

where $l \in \{1, \ldots, d\}$ and $Z_{j,l} \sim i.i.d. N(0,1)$. We will first show that $b \in H^s_{\text{per}}$, which requires $\|D^\alpha b\|_{L^2} < \infty$, $\forall 0 \leq |\alpha| \leq s$. It is sufficient to require $\|D^\alpha b_l\|_{L^2} < \infty$, $\forall 0 \leq |\alpha| \leq s$, for each $l \in \{1, \ldots, d\}$.

In the following notation we will denote $C > 0$ to be some arbitrary constant. We can choose $\phi_j(x)$ to be an orthonormal Fourier basis of $L^2(\tau, \mathbb{R})$ such that its partial derivatives are bounded above by

$$D^\alpha \phi_j(x) \leq C \left( \prod_{l=1}^d (2\pi j_l)^\alpha \right) \phi_j, \quad C > 0.$$

For example, the bases used for the 1- and 2-dimensional case can be found in Sections 3.5.3 and 3.6.4. For our choice of precision operator $A_0 = \eta(-\Delta)^k + \epsilon$, we have

$$A_0 \phi_j \leq C \left( \eta \sum_{l=1}^d (2\pi j_l)^{2k} + \epsilon \right) \phi_j, \quad C > 0,$$

so that the corresponding eigenvalues $\lambda_j$ are bounded above by

$$\lambda_j \leq C \left( \eta \sum_{l=1}^d (2\pi j_l)^{2k} + \epsilon \right)^{-1}, \quad C > 0.$$

Since by orthonormality we have $\phi_m \perp \phi_n$, $\forall m \neq n$, and $\|\phi_j\|_{L^2} = 1$, the eigen-
values are further bounded above by

\[ \lambda_j \leq C \left( \sum_{l=1}^{d} j_l^{2k} \right)^{-1} \leq C d^{2k-1} \left( \sum_{l=1}^{d} j_l \right)^{-2k} \propto |j|^{-2k}, \quad C > 0, \]

by Jensen’s inequality. We can then write

\[
\|D^\alpha b_l\|_{L^2}^2 = \left\| \sum_{j \in \mathbb{N}^d} Z_{j,l} \sqrt{\lambda_j} D^\alpha \phi_j \right\|_{L^2}^2,
\]

\[
= \left\| \sum_{j \in \mathbb{N}^d} Z_{j,l} \sqrt{\lambda_j} \prod_{l=1}^{d} (2\pi i j_l)^{\alpha_l} \phi_j \right\|_{L^2}^2,
\]

\[
= \sum_{j \in \mathbb{N}^d} Z_{j,l}^2 \lambda_j \prod_{l=1}^{d} (2\pi i j_l)^{2\alpha_l},
\]

\[
\leq C \sum_{j \in \mathbb{N}^d} Z_{j,l}^2 |j|^{-2k} |j|^{2|\alpha|}, \quad C > 0.
\]

Taking its expectation gives

\[
\mathbb{E} \left( \|D^\alpha b_l\|_{L^2}^2 \right) \leq C \sum_{j \in \mathbb{N}^d} |j|^{-2k+2|\alpha|}, \quad C > 0.
\]

We can then transform this sum into an integral over the polar coordinates in the \(d\)-dimensional Euclidean space of the form

\[
\int_{\Psi} \int_{0}^{\infty} \gamma^{d-1} V(\psi) d\gamma d\psi,
\]

where \(\gamma\) is the radial coordinate, \(\psi = (\psi_1, \ldots, \psi_{d-1})^T \in \Psi\) the angular coordinates and \(\gamma^{d-1} V(\psi)\) the Jacobian of the transformation. This gives us

\[
\sum_{j \in \mathbb{N}^d} |j|^{-2k+2|\alpha|} \propto \int_{0}^{\infty} \gamma^{-2k+2|\alpha|} \gamma^{d-1} d\gamma.
\]

This integral converges if \(-2k+2|\alpha|+d-1<-1\) and therefore \(\mathbb{E} \left( \|D^\alpha b_l\|_{L^2}^2 \right) < \infty\) if

\[ |\alpha| < \frac{2k-d}{2}, \]
3.6. Non-parametric drift inference: dimension greater than one

which in turn implies $P(\|D^\alpha b\|_{L^2} < \infty) = 1$.

By additionally setting $s > d/2 + 1$, we can apply the Sobolev embedding theorem for periodic domains (see Corollary A.2 from [97, p. 436]), so that $b \in C^1(\tau)$ and $\|b\|_{C^1} \leq C\|b\|_{H^s}$ for some constant $C > 0$. This gives

$$D^\alpha b \in C^0(\tau), \quad |\alpha| \leq 1,$$

$$\|D^\alpha b\|_{\infty} = \sup_{x \in \tau} |D^\alpha b(x)| < \infty, \quad |\alpha| \leq 1,$$

and thus

$$\frac{|b(y) - b(x)|}{|y - x|} \leq \max_{|\alpha| \leq 1} \left( \sup_{x \in \tau} |D^\alpha b(x)| \right), \quad x, y \in \tau,$$

by definition of $D^\alpha b$.

The global Lipschitz condition is therefore satisfied for the drift $b$ when $s, k$ and $d$ are chosen such that $d/2 + 1 < s < k - d/2$.

3.6.2 The likelihood and empirical measure

The likelihood

Following Definition[1] the likelihood of the drift can be written in the form

$$l(b) \propto \exp \left[ -\frac{1}{2} \int_0^T b(X_t) \cdot S^{-1} b(X_t) dt + \int_0^T S^{-1} b(X_t) \cdot dX_t \right], \quad (3.40)$$

where $S = \sigma \sigma^\top$.

However, unlike in Section 3.5.1, we cannot move forward with the change of variables involving the local time process $L_T$. Firstly, the transformation of the stochastic integral in (3.40) into a convenient form in the same vein as equation (3.25) cannot in general be achieved by applying the multi-dimensional Itô formula onto the anti-derivative of $b$.

To see why, let us for the moment assume that the diffusion process $X$ has unit diffusivity (as a result of the Lamperti transform from Definition 6, for example). Now suppose we have a function $B : \tau \rightarrow \mathbb{R}^d$. Applying the multi-dimensional Itô
3.6. Non-parametric drift inference: dimension greater than one

formula gives

\[ dB(X_t) = \sum_{i=1}^{d} \frac{\partial B}{\partial u_i}(X_t) dX_{i,t} + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2 B}{\partial u_i u_j}(X_t) dX_{i,t} dX_{j,t}, \]

where \( X_t = (X_{1,t}, \ldots, X_{d,t})^{\top} \). Taking the integral and re-arranging gives

\[ \sum_{i=1}^{d} \int_0^T \frac{\partial B}{\partial u_i}(X_t) dX_{i,t} = \int_0^T dB(X_t) - \frac{1}{2} \sum_{i,j=1}^{d} \int_0^T \frac{\partial^2 B}{\partial u_i u_j}(X_t) dX_{i,t} dX_{j,t}. \] (3.41)

However, the equality

\[ \int_0^T b(X_t) dX_t = \sum_{i=1}^{d} \int_0^T \frac{\partial B}{\partial u_i}(X_t) dX_{i,t} \] (3.42)

only holds if the drift \( b \) is the gradient of \( B \), that is \( b = -\nabla B \), so the step to transform the left hand side of (3.41) into the stochastic integral we seek to eliminate cannot be performed in general.

Secondly, the empirical measure \( \Gamma_T \) (see Definition 12) replaces the local time process \( L_T \) for dimensions greater than 1. This empirical measure allows for the change of variables given by

\[ \int_0^T \phi(X_t) dt = \int_T \phi(u) d\Gamma_T(u), \] (3.43)

for all bounded, measurable test functions \( \phi \). Thus a change of variables allows us to rewrite the first integral in the above likelihood (3.40) as a spatial integral over \( \tau \) with respect to the empirical measure \( \Gamma_T \) of the path \( X \). Unlike the local time process however, the empirical measure is not absolutely continuous with respect to the Lebesgue measure.

The likelihood (3.40) then becomes

\[ l(b) \propto \exp \left[ -\frac{1}{2} \int_{\tau} b(u) \cdot S^{-1} b(u) d\Gamma_T(u) + \int_0^T S^{-1} b(X_t) \cdot dX_t \right]. \] (3.44)

In comparison with the 1-dimensional case, in which the local time process \( L_T \)
combined with the start and endpoints $X_0$ and $X_T$ form a sufficient statistic for the drift $b$, the multi-dimensional case here requires the entire trajectory of $X$.

### 3.6.3 The posterior distribution

Combining (3.44) with the prior density (3.39) gives the posterior density of form

$$
\pi(b|X) \propto \exp \left[ -\frac{1}{2} \int_\tau b(u) \cdot A_0 b(u) \, du - \frac{1}{2} \int_\tau b(u) \cdot S^{-1} b(u) \, d\Gamma_T(u) \right] 
$$

$$
+ \int_0^T S^{-1} b(X_t) \cdot dX_t ,
$$

$$
= \exp \left[ -\frac{1}{2} (b, A_0 b)_{L^2} - \frac{1}{2} (b, S^{-1} b)_{\Gamma_T} + \int_0^T S^{-1} b(X_t) \cdot dX_t \right] ,
$$

where $(f, g)_{L^2}$ is the $L^2$-inner product and $(f, g)_{\Gamma_T}$ is the $\Gamma_T$-inner product characterised by $\int_\tau f(u) \cdot S^{-1} g(u) \, d\Gamma_T(u)$, for some $f$, $g$.

In lieu of a rigorous derivation, we proceed by discretising (3.46), which results in a matrix quadratic form of the log-posterior

$$
\log \pi(b|X)_{\text{disc}} \propto -\frac{1}{2} b_{\text{disc}}^T A_0 b_{\text{disc}} - \frac{1}{2} b_{\text{disc}}^T G_T b_{\text{disc}} + rb_{\text{disc}},
$$

where $b_{\text{disc}}$ is the discretisation of $b$, the matrices $A_0$, $G_T$ and $r$ denote the actions of the $L^2$-, $\Gamma_T$-inner products and the stochastic integral respectively on $b$ as present in (3.45) and (3.46). Note that we have considered $S$ to be absorbed into the action of the matrices $G_T$ and $r$. This quadratic form (3.47) is convenient as it not only makes it intuitively clear that the Gaussian prior is indeed conjugate, but also interfaces well with the development of the numerical methods in Section 3.6.4 where we explain the discretisation step of (3.47) in detail.

The relationship between the the continuous and discretised forms can thus be summarised as

$$
b_{\text{disc}}^T A_0 b_{\text{disc}} := (b, A_0 b)_{L^2},
$$

$$
b_{\text{disc}}^T G_T b_{\text{disc}} := (b, S^{-1} b)_{\Gamma_T},
$$
3.6. Non-parametric drift inference: dimension greater than one

and

\[ rb_{\text{disc}} := \int_0^T S^{-1} b(X_t) \cdot dX_t. \]

Henceforth we will drop the \( b_{\text{disc}} \) notation in favour of simply writing \( b \) when the context makes it clear that we are working in a discretised setting.

Completing the square for (3.47) gives us

\[
\log \pi(b|X)_{\text{disc}} = -\frac{1}{2} [b - (A_0 + G_T)^{-1} r]^T (A_0 + G_T) [b - (A_0 + G_T)^{-1} r],
\]

and matching terms leads to the prior-posterior update equations

\[
A_1 = A_0 + G_T \quad \text{and} \quad b_1 = (A_0 + G_T)^{-1} r,
\]

for the discretised precision operator \( A_1 \) and posterior mean \( b_1 \).

**Variational formulation of the posterior mean**

To corroborate the update equation \( b_1 = (A_0 + G_T)^{-1} r \), let us consider the function \( b \) that maximises the log-posterior (3.44). Denote \( F[b] := \log \pi(b|X) \) such that

\[
F[b] = -\frac{1}{2} (b, A_0 b)_{\mathcal{L}^2} - \frac{1}{2} (b, S^{-1} b)_{\Gamma_T} + \int_0^T S^{-1} b(X_t) \cdot dX_t.
\]

The functional derivative \( \frac{\delta F}{\delta b} \) of \( F[b] \) is defined as

\[
\int \frac{\delta F}{\delta b}(u) \psi(u) du = \lim_{\varepsilon \to 0} \frac{F[b + \varepsilon \psi] - F[b]}{\varepsilon},
\]

for some test function \( \psi \). To find the function \( b \) that maximises \( F[b] \), we evaluate (3.50) via the right hand side and set \( \frac{\delta F}{\delta b} \) equal to zero, resulting in

\[
\lim_{\varepsilon \to 0} \frac{F[b + \varepsilon \psi] - F[b]}{\varepsilon} = - (\psi, A_0 b)_{\mathcal{L}^2} - (\psi, S^{-1} b)_{\Gamma_T} + \int_0^T S^{-1} \psi dX_t,
\]
since $A_0$ and $S^{-1}$ are positive definite. Re-arranging gives

$$
(\psi, A_0 b)_{L^2} + (\psi, S^{-1} b)_{\Gamma_T} = \int_0^T S^{-1} \psi dX_t. \tag{3.51}
$$

The weak solution to (3.51) is thus the function $b$ that maximises $F[b]$.

In principle, a more thorough analysis is required to establish that the posterior implied by the likelihood (3.44) and the Gaussian prior density (3.39) is indeed Gaussian with mean given by the solution of (3.51). Similarly to the 1-dimensional treatment, an argument along the lines of Pokern et al. [88] leading up to Theorem 2 modified for the multi-dimensional case should be warranted, but we delegate this as future work and proceed according to the discretisation presented in (3.48) instead.

### 3.6.4 Discretisation of the drift and the pseudo-spectral method

Obtaining the posterior mean $b_1$ from the posterior update equation

$$
b_1 = (A_0 + G_T)^{-1} r, \tag{3.52}
$$

or equivalently finding the weak solution $b$ of

$$
(\psi, A_0 b)_{L^2} + (\psi, S^{-1} b)_{\Gamma_T} = \int_0^T S^{-1} \psi dX_t, \tag{3.53}
$$

is not a trivial matter, but can be approached using the pseudo-spectral method for solving partial differential equations, where a brief explanation and example can be found in Section 2.3. Our aim is to reduce (3.52) and (3.53) into a finite system of linear equations of a form not dissimilar to $(A_0 + G_T)b = r$, where the operators, integrals and functions involved are represented as matrices and vectors over a fixed grid, which can then be solved iteratively using a conjugate gradient algorithm.

In fact, the discretisation used for (3.52) allows us to write (3.53) as

$$
\psi^T A_0 b + \psi^T G_T b = r \psi. \tag{3.54}
$$
We then apply the pseudo-spectral method, as a special case of the example in Section 2.3 where the second-order partial differential term is replaced with a $k$-th order Laplacian operator, by representing the terms in (3.54) using the natural and Fourier basis functions, resulting in

$$
\hat{\psi}^\top \hat{A}_0 \hat{b} + \hat{\psi}^\top \hat{G}_T \hat{b} = \hat{r} \hat{\psi},
$$

(3.55)

or, equivalently

$$
\hat{\psi}^\top \left( \hat{A}_0 \hat{b} + \hat{G}_T \hat{b} \right) = \hat{r} \hat{\psi},
$$

(3.56)

where $\hat{\cdot}$ denotes its Fourier transform. Under the Fourier basis, the discretised differential operator $A_0$ becomes a diagonal matrix $\hat{A}_0$, which makes $\hat{A}_0 \hat{b}$ easy to handle. The discretised empirical measure operator $G_T$ however is a sparse matrix only in the natural basis, hence the choice of notation $\hat{G}_T \hat{b}$, which amounts to first applying the inverse Fourier transform to $\hat{b}$, multiplying by $G_T$, then applying the Fourier transform back to obtain $\hat{G}_T \hat{b}$. We can write this formally in terms of the Fourier transform operator $F$ as

$$
\hat{\psi}^\top \left( \hat{A}_0 \hat{b} + \mathcal{F}(G_T \mathcal{F}^{-1}(\hat{b})) \right) = \hat{r} \hat{\psi}.
$$

(3.57)

By Parseval’s theorem, solving for (3.54) is equivalent to solving for (3.57). With this, we can focus on solving the system of equations represented by

$$
\hat{\psi}^\top \left( \hat{A}_0 \hat{b} + \mathcal{F}(G_T \mathcal{F}^{-1}(\hat{b})) \right) = \hat{r} \hat{\psi},
$$

where $\hat{A}_0$ and $G_T$ are both diagonal in their respective Fourier and natural basis representations, which can be readily solved with Krylov subspace based iterative methods. Finally, applying the inverse Fourier transform to the solution $\hat{b}$ will then yield our desired posterior mean $b$. In the following sections we will describe our implementation of the pseudo-spectral method as restricted to the 2-dimensional case due to our practical application to be seen in Section 3.9, but note that this implementation can extended to the multi-dimensional case without too much difficulty.
The drift and its Fourier transform

First, the torus $\tau$ is discretised into a grid with $k \times k$ points. The discretisation of the drift $b$ can thus be written as

$$b = \begin{pmatrix} b_x \\ b_y \end{pmatrix},$$

where $b \in \mathbb{R}^{2k \times k}$ and $b_x, b_y \in \mathbb{R}^{k \times k}$.

Let us represent the discrete Fourier transform operator and its inverse as matrices $\mathcal{F}, \mathcal{F}^{-1} \in \mathbb{R}^{k \times k}$. Additionally, the Kronecker product of two matrices $S \in \mathbb{R}^{2 \times 2}$ and $B \in \mathbb{R}^{k \times k}$ is given as the $2k \times 2k$ block matrix

$$S \otimes B = \begin{pmatrix} S_{11}B & S_{12}B \\ S_{21}B & S_{22}B \end{pmatrix},$$

Thus the Fourier transform of the drift can be represented by

$$(I_2 \otimes \mathcal{F})b = \hat{b} = \begin{pmatrix} \hat{b}_x \\ \hat{b}_y \end{pmatrix},$$

where $\hat{b} \in \mathbb{R}^{2k \times k}, \hat{b}_x, \hat{b}_y \in \mathbb{R}^{k \times k}$ and $I_2 \in \mathbb{R}^{2 \times 2}$ is the 2-dimensional identity matrix.

The PDE

The PDE (3.56) for the posterior mean can be written more carefully as

$$\hat{\psi}^T \left[ \begin{pmatrix} \hat{A}_0 \circ \hat{b}_x \\ \hat{A}_0 \circ \hat{b}_y \end{pmatrix} + \begin{pmatrix} S_{11}^{-1} \mathcal{F}G_T \circ (\mathcal{F}^{-1} \hat{b}_x) + S_{12}^{-1} \mathcal{F}G_T \circ (\mathcal{F}^{-1} \hat{b}_y) \\ S_{21}^{-1} \mathcal{F}G_T \circ (\mathcal{F}^{-1} \hat{b}_x) + S_{22}^{-1} \mathcal{F}G_T \circ (\mathcal{F}^{-1} \hat{b}_y) \end{pmatrix} \right] = \sum_i \hat{\psi}(X_i)S^{-1}(X_{i+1} - X_i),$$

(3.58)

where $\circ$ denotes the Hadamard or element-wise product of two equally-sized matrices. Note that we have decoupled $S$ from $G_T$ to make it clear that $S$ is simply a constant matrix. We have also used the Riemann sum approximation of the stochastic integral $r\psi := \int_0^T S^{-1} \psi dX_t \approx \sum_i \psi(X_i)S^{-1}(X_{i+1} - X_i)$.

To ease notation, we can instead transform $\hat{b}$ into a $2k^2 \times 1$ vector formed by
3.6. Non-parametric drift inference: dimension greater than one

stacking the columns of \( \hat{b}_x \) and \( \hat{b}_y \), such that now

\[
\hat{b} = \begin{pmatrix}
\text{vec}(\hat{b}_x) \\
\text{vec}(\hat{b}_y)
\end{pmatrix},
\]

where \( \hat{b} \in \mathbb{R}^{2k^2} \), and \( \text{vec}(\hat{b}_x), \text{vec}(\hat{b}_y) \in \mathbb{R}^{k^2} \).

The PDE (3.58) can then be written more succinctly as

\[
\psi^\top (I \otimes \hat{A}_0 + S^{-1} \otimes (\mathcal{F}G_T\mathcal{F}^{-1}))\hat{b} = \sum_i \psi(X_{t_i})S^{-1}(X_{t_{i+1}}-X_{t_i}),
\]

where \( \hat{A}_0, G_T \) are now \( 2k^2 \times 2k^2 \) diagonal matrices and \( \mathcal{F}, \mathcal{F}^{-1} \) the Fourier and inverse Fourier transformation matrices adjusted for the above vectorisation of \( \hat{b} \).

**Choice of Fourier basis**

We additionally attach to \( \tau \) the parameterisation by \( [0,x_s] \times [0,y_s] \), where \( x_s, y_s \in \mathbb{R}^+ \), and discretise \( \tau \) through the use of a regular grid of \( k \times k \) points, where \( k = 2^l \), for some \( l \in \mathbb{N} \), to simplify the use of a fast Fourier transform algorithm. We then expand \( b \) as

\[
b(x,y) = \sum_{j,k=0}^{2^l-1} c_{j,k} \cos \left( \frac{2\pi jx}{x_s} + \frac{2\pi ky}{y_s} \right) + d_{j,k} \sin \left( \frac{2\pi jx}{x_s} + \frac{2\pi ky}{y_s} \right),
\]

\[
+ e_{j,k} \cos \left( \frac{2\pi jx}{x_s} - \frac{2\pi ky}{y_s} \right) + f_{j,k} \sin \left( \frac{2\pi jx}{x_s} - \frac{2\pi ky}{y_s} \right),
\]

using real Fourier basis functions with corresponding coefficients \( c_{j,k}, d_{j,k}, e_{j,k} \) and \( f_{j,k} \). The approximations of the \( \Gamma_T \)-inner product and the stochastic integral for \( G_T \) and \( r\phi \) are described as

\[
(u,u)_{\Gamma_T} \approx \Delta t \sum_i u^2 \left( \frac{x_s}{m} \left\lfloor \frac{mX_{1,t_i}}{x_s} \right\rfloor, \frac{y_s}{m} \left\lfloor \frac{mX_{2,t_i}}{y_s} \right\rfloor \right),
\]

and

\[
\int_0^T \psi(X_t) \cdot dX_t \approx \sum_i \psi \left( \frac{x_s}{m} \left\lfloor \frac{mX_{1,t_i}}{x_s} \right\rfloor, \frac{y_s}{m} \left\lfloor \frac{mX_{2,t_i}}{y_s} \right\rfloor \right) (X_{t_{i+1}}-X_{t_i}),
\]
3.6. Non-parametric drift inference: dimension greater than one

where $X_{1,t}$, $X_{2,t}$ refer to the first and second dimension components of the observed point $X_t = (X_{1,t}, X_{2,t})^\top$, and $\lfloor \cdot \rfloor$ is the floor function.

**System of linear equations**

Finally, solving for the vector $\hat{b}$ in (3.59) amounts to solving a system of equations, which can be achieved by using an iterative solver. We note that $[I \otimes A_0 + S^{-1} \otimes (F GF^{-1})]$ is in fact positive definite symmetric, which allows the use of the conjugate gradient method, which we describe in Section 2.4. In brief, it is an iterative method for solving a linear system of equations of the form $Ab = r$.

The conjugate gradient method is convenient, as it allows us to treat $A$ as a black-box function rather than as an explicit matrix form. This suits us well since the discretised matrix $[I_2 \otimes A_0 + S^{-1} \otimes (F GF^{-1})]$ is not explicitly calculated. Rather, it is implemented as the diagonal matrix $A_0$ and a function $FGF^{-1}$. We also note that the conjugate gradient method implemented is in fact the preconditioned conjugate gradient method to ensure fast convergence to the solution.

**Preconditioning for conjugate gradient**

When solving for $b$ in the system $Ab = r$ using iterative methods, in this case via the conjugate gradient algorithm, the properties of $A$ are important to the convergence of the solution. Iterative solvers suffer from lack of robustness, especially for large systems of equations, which is the case for the system of equations (3.59) that we consider here. Therefore, effective preconditioning where $A$ is modified to have suitable properties is important for the successful application of iterative methods.

According to Trefethen and Bau [113], the idea behind preconditioning is to move from solving $Ab = r$ to solving a modified system of the form

$$M^{-1}Ab = M^{-1}r,$$

(3.60)

which has the same solution as $Ab = r$. The convergence of the solution however depends on the properties of $M^{-1}A$ instead of $A$. Ideally the matrix $M$ should be
chosen to be close enough to $A$ such that the product $M^{-1}A$ is close to $I$, non-
singular and is inexpensive to compute.

Since the matrix $A$ in our case is positive definite symmetric, the preconditioner

 chosen to be close enough to $A$ such that the product $M^{-1}A$ is close to $I$, non-
singular and is inexpensive to compute.

Since the matrix $A$ in our case is positive definite symmetric, the preconditioner

can be chosen as $M = CC^T$ for some $C$, so that solving [3.60] is equivalent to solving

$$[C^TAC]C^{-1}b = Cr,$$

where $C^TAC$ is in turn positive definite symmetric as well.

While there are a variety of possibilities for the choice of preconditioner $C$, we
will consider just two choices here.

**The current preconditioner**

Recall the posterior precision matrix in [3.59] given by

$$I_2 \otimes A_0 + S^{-1} \otimes (FG_T F^{-1}).$$

Now denote by $U \in \mathbb{R}^{j \times k}$ a projector matrix with $j < k$ columns composed of
the first $j$ standard basis vectors $e_j \in \mathbb{R}^j$. $U \hat{b}$ is therefore the projection of $\hat{b}$ onto the
low ($j$-) dimensional Fourier subspace corresponding to the first $j$ Fourier modes.
A low rank approximation of the posterior precision matrix $A_1$ can be represented
by

$$A_{1,j} = I_2 \otimes A_{0,j} + S^{-1} \otimes (U^T F_j G_{T,j} F_j^{-1} U),$$

where $A_{0,j}, G_{T,j}, F_j, F_j^{-1} \in \mathbb{R}^{j \times j}$ are the submatrices of $A_0, G_T, F, F^{-1}$ formed
from their respective first $j \times j$ terms.

We can then construct a preconditioner from this low rank approximation as

$$C = I_2 \otimes A_{0,j}^{-1/2} + (I_2 \otimes U) \left[ \text{chol} (A_{1,j}) \right]^{-1} (I_2 \otimes U^T),$$

where $A_{0,j}^{-1/2}$ is the matrix $A_0^{-1/2}$ with the first $j$ terms of its diagonal entries set to
zero and $\text{chol} (\cdot)$ is the incomplete Cholesky factorisation.

This (3.62) is the preconditioner that we use for our implementation. However,
3.6. Non-parametric drift inference: dimension greater than one

Let us once again examine Figure 3.2. The green line tracks the computation time of our drift inference implementation, including the conjugate gradient algorithm for solving the posterior mean. The periodic spikes represent the moment where the preconditioner $C$ is recalculated to account for new estimates of the diffusivity $\sigma$, and hence for $S$. These computational costs are relatively high, and so in an attempt to improve the performance of the conjugate gradient, we also consider here an alternate preconditioner. However, we emphasize that it does not perform well in our current implementation, but should serve as a foundation to discover potentially better preconditioners.

An experimental preconditioner

Let us consider once more the projector matrix $U \in \mathbb{R}^{j \times k}$, and write an alternative low rank approximation of the posterior precision matrix as

$$A_1 \approx I_2 \otimes A_0 + (I_2 \otimes U)(S^{-1} \otimes D)(I_2 \otimes U^T),$$  \hspace{1cm} (3.63)

where

$$D = U^T \mathcal{F}_T \mathcal{F}^{-1} U.$$  \hspace{1cm}

The Woodbury matrix identity, see for example Golub and Van Loan [59], gives the following inverse expansion

$$(A + UDU^T)^{-1} = A^{-1} - A^{-1}U(D^{-1} + U^TA^{-1}U)^{-1}U^TA^{-1}$$  \hspace{1cm} (3.64)

for matrices $A$, $U$ and $D$ of correct sizes. The inverse of (3.63) can then be expanded as

$$\left( I_2 \otimes A_0 + (I_2 \otimes U)(S^{-1} \otimes D)(I_2 \otimes U^T) \right)^{-1}
= I_2 \otimes A_0^{-1/2} \left[ I - (I_2 \otimes (A_0^{-1/2}U)) (S \otimes D^{-1} + I_2 \otimes (U^TA_0^{-1/2})) \right] I_2 \otimes A_0^{-1/2}. \hspace{1cm} (3.65)$$

Ambikasaran [7] provided a factorisation of matrices of form $I + UKU^T$ in the
3.6. Non-parametric drift inference: dimension greater than one

following theorem.

**Theorem 3.** [7, p. 3] For the $m \times m$ identity matrix $I$, matrices $U \in \mathbb{R}^{m \times n}$ and $K \in \mathbb{R}^{n \times n}$, the symmetric positive definite matrix $I + UKU^T$ can be symmetrically factored as

$$I + UKU^T = (I + UXU^T)(I + UXU^T)^T,$$

(3.66)

where $X = L^{-T}(M - I)L^{-1}$, $LL^T = U^TU$ and $MM^T = I + L^TKL$.

Using Theorem 3, the right hand side of (3.65) can be written as

$$(I_2 \otimes A_0 + (I_2 \otimes U)(S^{-1} \otimes D)(I_2 \otimes U^T))^{-1}$$

$= I_2 \otimes A_0^{-1/2} (I + (I_2 \otimes U)X(I_2 \otimes U^T)) \left[ I_2 \otimes A_0^{-1/2} (I + (I_2 \otimes U)X(I_2 \otimes U^T)) \right]^T,$

(3.67)

where $X = \sqrt{K + I} - I$ since $U^TU = I$, and

$$K = - \left( S \otimes (A_{0,j}^{1/2} D^{-1} A_{0,j}^{1/2} + I) \right)^{-1},$$

with $A_{0,j}$ the restriction of $A_0$ to act only on the first $j$ Fourier modes, such that $A_0U = UA_{0,j}$.

We can then set the pre-conditioner as

$$C_{\text{new}} = I_2 \otimes A_{0,j}^{-1/2} (I + (I_2 \otimes U)X(I_2 \otimes U^T)).$$

However, this preconditioner does not perform as well as the preconditioner given by (3.62). We suspect that the matrix $D = U^T F M_{T,j} F^{-1} U$ remains relatively ill-conditioned, and that improvements can be made by modifying this matrix $D$.

### 3.6.5 Sampling from the prior and posterior distributions

**Sampling from Gaussian measures**

Sampling $b$ from $N(0,A^{-1})$ amounts to evaluating $b = A^{-\frac{1}{2}}z$ where $z$ is a sample from $N(0,I)$. To do so, we will utilise a combination of the Lanczos algorithm to
reduce $A$ into an easily invertible form, the Lanczos approximation of $A^{-\frac{1}{2}}z$ and the conjugate gradient algorithm to determine convergence and accuracy of the Lanczos algorithm and approximation respectively. A discussion of this combined method can be found in Aune et al. [11].

In brief, the Lanczos algorithm first takes two inputs, a vector $z$ and a Hermitian matrix $A$, then after $n$ iterations returns the Lanczos vectors $\{q_1, \ldots, q_n\}$, which form an orthonormal basis of the Krylov subspace $\mathcal{K}_n(A, z) = \text{span}\{z, Az, \ldots, A^{n-1}z\}$, and a symmetric tridiagonal matrix $T_n = Q_n^\top AQ_n$, with $Q_n$ the column matrix formed by the aforementioned Lanczos vectors. The Lanczos algorithm is outlined as Algorithm 2 in Section 2.4 and we refer to Trefethen and Bau [113] for an easy introduction.

The product $A^{-\frac{1}{2}}z$ can then be approximated using the Lanczos approximation, see for example van der Vorst [118] or Frommer and Simoncini [53], which is given by

$$f(A)z \approx \|z\|Q_n f(T_n)e_1,$$

(3.68)

for any matrix function $f$ and $\| \cdot \|$ the Euclidean norm. In our case, we simply have $f(A) = A^{-\frac{1}{2}}$.

Additionally, Simpson et al. [107] presented the following theorem which addresses not only the accuracy of the Lanczos approximation (3.68) but also provides a stopping criteria for the Lanczos algorithm.

**Theorem 4.** Let $A$ be a symmetric positive definite matrix with its smallest eigenvalue denoted as $\lambda_{\text{min}}$. Then

$$\left\| A^{-\frac{1}{2}}z - \|z\|Q_n T_n^{-\frac{1}{2}}e_1 \right\| \leq \lambda_{\text{min}}^{-\frac{1}{2}}\|r_n\|,$$

(3.69)

where $r_n$ is the residual obtained after $n$ iterations of conjugate gradients to solve $Ax = z$.

According to the treatment by Arbenz [9], the Lanczos algorithm has the curious property of converging if and only if the Lanczos vectors lose orthogonality. In
3.6. Non-parametric drift inference: dimension greater than one

addition to this, we can utilise the conjugate gradient residual \( r_n \) to indicate convergence of the Lanczos approximation (3.68).

Since the conjugate gradient algorithm can be derived from the Lanczos algorithm, Saad [100] provided explicit formulae relating the Lanczos and conjugate gradient terms. The conjugate gradient coefficients and therefore the residual \( r_n \) can be obtained with minimal computational cost when computed in parallel with the Lanczos algorithm.

We express here the conjugate gradient terms with respect to the Lanczos vectors \( \{q_1, \ldots, q_n\} \) as

\[
\begin{align*}
    r_n &= \sum_{k=1}^{n+1} \rho^n_k q_k, \\
    x_n &= \sum_{k=1}^{n+1} \chi^n_k q_k, \\
    p_n &= \sum_{k=1}^{n+1} \pi^n_k q_k,
\end{align*}
\]

where \( r_n, p_n \) and \( x_n \) (correspondingly \( \rho^n, \chi^n \) and \( \pi^n \)) are the conjugate gradient residual, search direction and solution terms, see for example Algorithm [1]. We will additionally use the notation \( \rho^n_0 \) to refer to the entire vector of coefficients \( \rho^n_0 = \{\rho^n_0, \ldots, \rho^n_{n+1}\} \).

Finally, as with the conjugate gradient algorithm, we pre-apply the left- and right-preconditioner \( C \) given by (3.62) to both \( z \) and \( A \), which requires a modification of the Lanczos approximation (3.68) as follows

\[
A^{-\frac{1}{2}}z = C(C^TAC)^{-\frac{1}{2}}(C^Tz) \approx \beta_1 CQ_n^Tn^{-\frac{1}{2}}e_1,
\]

where \( \beta_1 = \|Cz\| \). The amalgamate pre-conditioned Lanczos algorithm for sampling from \( b \sim N(0,A^{-1}) \) with conjugate gradient residual stopping criteria is thus given as Algorithm [2]

**Algorithm 8. Lanczos algorithm for sampling** \( b \sim N(0,A^{-1}) \)

1. Target: \( b \sim N(0,A^{-1}) \)

2. Initialize Lanczos: \( z \sim N(0,I), q_0 = 0, q_1 = \frac{Cz}{\|Cz\|}, \beta_0 = 0, \beta_1 = \|Cz\| \)

3. Initialize CG: \( \chi_0^0 = 0, \rho_1^0 = \|Cb\|, \pi_1^0 = \|Cb\|, \) and set all other entries of \( \pi_0^0 \) and \( \rho_0^0 \) to zero.
3.6. Non-parametric drift inference: dimension greater than one

4. Iterate for $n = 1, 2, 3, \ldots$ until $(\rho_n^T \rho_n)_{\rho_0^T \rho_0} < \text{designated tolerance}$:

(a) $v = CAC^T q_n$

(b) $\alpha_n = q_n^T v$

(c) $v = v - \beta_{n-1} q_{n-1} - \alpha_n q_n$

(d) $\beta_n = \|v\|$

(e) $q_{n+1} = \frac{v}{\beta_n}$

(f) $c_o = \pi_{o}^{n-1} \alpha_o + \pi_{o+1}^{n-1} \beta_o + \pi_{o-1}^{n-1} \beta_{o-1}$

(g) $\gamma_n = \frac{(\rho_{n-1}^{o})^T \rho_{n-1}^{o}}{(\pi_{o}^{n-1})^T c_o}$

(h) $\chi_o^n = \chi_o^{n-1} + \gamma_n \pi_{o}^{n-1}$

(i) $\rho_o^n = \rho_o^{n-1} - \gamma_n c_o$

(j) $\delta_n = \frac{(\chi_o^n)^T \chi_o^{n-1}}{(\chi_o^{n-1})^T \chi_o^{n-1}}$

(k) $\pi_o^n = \rho_o^n + \delta_n \pi_{o}^{n-1}$

5. Return $b = \beta_1 CQ_n T_n^{-\frac{1}{2}} e_1$, with column matrix $Q_n = \begin{bmatrix} q_1 & \ldots & q_n \end{bmatrix}$, tridiagonal matrix $T = \begin{pmatrix} \alpha_1 & \beta_2 \\ \beta_2 & \alpha_2 & \beta_3 \\ & \beta_3 & \alpha_3 & \beta_4 \\ & & \ddots & \ddots & \ddots \end{pmatrix}$, and $e_1 = (1, 0, \ldots, 0)^T \in \mathbb{R}^n$.

3.6.6 Conditioning on conservative drift

The Helmholtz decomposition

Thus far we have only considered a drift of general form, which may not be necessary for some datasets as it may be more desirable to fit a model with certain restrictive properties instead. As a pointed example, in a pioneering paper on applying non-trivial diffusion process models to animal movement ecology research, Brillinger [22] fitted diffusion process models with conservative drift $b = -\nabla \phi$ to
the movement data of elk, *Cervus elaphus*, and a point of attraction can be clearly identified as a result. In fact, referring back to the factorisation of the likelihood of the drift in Section 3.6.2, a conservative drift \( b = -\nabla \phi \) would allow the equality of (3.42), and therefore allow a transformation of the stochastic integral (3.40) into a temporal integral as in (3.25) to allow for an easier factorisation of the likelihood. It is therefore of interest to be able to perform inference for a drift that is constrained to be conservative, that is \( b = -\nabla \phi \) for some scalar field \( \phi \), for a diffusion process characterised by

\[
dX_t = -\nabla \phi(X_t) dt + \sigma dW_t.
\]

According to the Helmholtz decomposition, see for example Arfken et al. [10], any vector field can be expressed as the sum of a conservative (or curl-free) vector field and a rotational (or divergence-free) vector field. For a vector field \( b \), these two components can be written as

\[
b = -\nabla \phi + \nabla \times H,
\]

where we denote by \( b_\parallel = -\nabla \phi \) the conservative component and \( b_\perp = \nabla \times H \) the rotational component. \( b_\parallel \) and \( b_\perp \) are orthogonal in \( L^2 \), where the direct sum of their respective subspaces form the \( L^2 \)-space, that is \( \{b_\parallel\} \oplus \{b_\perp\} = L^2 \). The Helmholtz decomposition allows for a convenient projection of \( b \) onto the \( \{b_\perp\} \) space. The solutions \( \tilde{\phi} \) and \( \tilde{H} \) of (3.70) however are non-unique since \( \tilde{\phi} = \phi + \text{const.} \) and \( \tilde{H} = H + \nabla \psi \) for some scalar field \( \psi \).

Two seemingly reasonable choices for performing inference for the case when the drift \( b = b_\parallel \) is conservative would be to:

1. perform the Helmholtz decomposition on the posterior mean as described by (3.48) and discard the rotational component, or to

2. condition the prior to the case where \( b = b_\parallel \), which as the Helmholtz decomposition allows, is equivalent to setting \( b_\perp = 0 \).

The first choice seems most straightforward, since the differences between the
two resultant posterior means are not immediately apparent. In a special case which we will describe later with Figure 3.8, these two means will in fact coincide. However, as shown later in Figure 3.9, the truncated mean obtained via the first choice can in general be shown to differ from the posterior mean as proposed by the second choice.

Therefore we will proceed with the second choice, and do so by extending the inference methodology prescribed in Sections 3.6.2-3.6.5 while considering instead the conditioned prior $\pi_0(b|b_\perp = 0)$.

**Helmholtz projectors and the Fourier basis**

We additionally introduce the projectors $P_\parallel$ and $P_\perp$ which project a vector field onto the subspaces spanned by conservative and rotational vector fields respectively, such that $P_\parallel b = b_\parallel$ and $P_\perp b = b_\perp$. Since $\nabla \cdot (\nabla \times H) = 0$, the projection $P_\parallel b$ is equivalent to solving for $\phi$ in the Poisson equation $\nabla \cdot b = -\Delta \phi$, when taking the periodic boundary conditions established in 3.6.2 into consideration, with the solution being unique up to an arbitrary additive constant, and then taking the gradient of the solution (see for example Evans [49] or Robinson [97] for additional details). Formally, we write this as

$$P_\parallel b = \nabla (\Delta^{-1}(\nabla \cdot b)) = \nabla (\Delta^{-1}(-\Delta \phi)). \quad (3.71)$$

In particular, the Fourier basis representation for the 2-dimensional case $\hat{b} = (\hat{b}_x, \hat{b}_y)^T$ of $b = (b_x, b_y)^T$ allows for the discretisation of (3.71), such that the projector $P_\parallel$ and correspondingly $P_\perp$ have the following forms

$$P_\parallel \hat{b}(k_x, k_y) = \begin{pmatrix} \frac{k_x^2}{k_x^2 + k_y^2} & \frac{k_x k_y}{k_x^2 + k_y^2} \\ \frac{k_y k_x}{k_x^2 + k_y^2} & \frac{k_y^2}{k_x^2 + k_y^2} \end{pmatrix} \hat{b}(k_x, k_y), \quad (3.72)$$

and

$$P_\perp \hat{b}(k_x, k_y) = \begin{pmatrix} \frac{k_x^2}{k_x^2 + k_y^2} & -\frac{k_x k_y}{k_x^2 + k_y^2} \\ \frac{k_y k_x}{k_x^2 + k_y^2} & -\frac{k_y^2}{k_x^2 + k_y^2} \end{pmatrix} \hat{b}(k_x, k_y), \quad (3.73)$$

since we have discretisation of the gradient as $\nabla v(k_x, k_y) = (k_x v_x, k_y v_y)^T$, the solution
3.6. Non-parametric drift inference: dimension greater than one

to the Laplacian as $\Delta^{-1}c = (\frac{1}{k_x^2+k_y^2}c, \frac{1}{k_x^2+k_y^2}c)^\intercal$ and the divergence as $\nabla \cdot v(k_x,k_y) = k_xv_x + k_yv_y$ for some vector $v$ and scalar $c$. Naturally we have $P_\parallel = P_\parallel^\intercal$, $P_\perp = P_\perp^\intercal$, $P_\parallel + P_\perp = I$, $P_\parallel^n = P_\parallel$ and $P_\perp^n = P_\perp$ for any $n \in \mathbb{N}$.

The prior conditioned on $b_\perp = 0$

Let us now consider the prior conditioned on $b = b_\parallel$, denoted by $\pi_0(b|b_\perp = 0)$, with conditional density given by

$$
\pi_0(b|b_\perp = 0) \propto \exp \left[ -\frac{1}{2} \int b_\parallel(u) \cdot (A_0b_\parallel)(u) du \right].
$$

The new posterior can then be represented by

$$
\pi(b|X,b_\perp = 0) = \frac{P(X|b,b_\perp = 0)\pi_0(b|b_\perp = 0)}{P(X|b_\perp = 0)}.
$$

Since the conditioning on $b_\perp = 0$ is equivalent to having $b = b_\parallel = P_\parallel b$, we insert this into the discretised log-posterior (3.47) from Section 3.6.3, which leads to

$$
\log \pi(b|X,b_\perp = 0)_{\text{disc}} \approx -\frac{1}{2} b^\intercal P_\parallel A_0 P_\parallel b - \frac{1}{2} b^\intercal P_\parallel G_T P_\parallel b + r^\intercal P_\parallel b,
$$

(3.74)

$$
= -\frac{1}{2} \left[ b - (A_0 + G_T)^\dagger P_\parallel r \right]^\intercal P_\parallel (A_0 + G_T) P_\parallel \left[ b - (A_0 + G_T)^\dagger P_\parallel r \right] + \text{const.,}
$$

(3.75)

where $(A_0 + G_T)^\dagger$ denotes the inverse of $A_0 + G_T$ constrained to the subspace $b_\perp = 0$, since $(P_\parallel(A_0 + G_T)P_\parallel)^{-1}$ is ill-defined, and $b$ is such that $b_\perp = 0$, otherwise $\log \pi(b|X)_{\text{disc}}$ is simply zero. More precisely, we define $(A_0 + G_T)^\dagger$ as the linear operator which maps any $r$ to the unique $b$ with $b_\perp = 0$ such that

$$
P_\parallel(A_0 + G_T)P_\parallel b = P_\parallel r,
$$

(3.76)

where $(A_0 + G_T)^\dagger$ can be represented as a Moore-Penrose inverse (a generalisation
3.6. Non-parametric drift inference: dimension greater than one

of the inverse, see for example Ben-Israel [15, p. 40]). The conditioned posterior mean can thus be given as the solution (3.76), or more succinctly as

\[ b_c = (A_0 + G_T)\dagger P_\parallel r. \]  

(3.77)

We have established in Section 3.6.4 through the use of the pseudo-spectral method that \((A_0 + G_T)\) can be considered to be positive definite symmetric. The Fourier representation of \(P_\parallel\) in (3.72) is positive definite symmetric, and so \((A_0 + G_T)\dagger P_\parallel r\) in (3.77) can be viewed as positive definite symmetric as well from a numerical solver point of view.

Solving for \(b_c\) can thus proceed according to the conjugate gradient scheme prescribed in Section 2.4 for the unconstrained drift case, but with an adjustment to the pre-conditioning terms, which we will describe later as (3.89).

**Orthogonality of \(b_\parallel\) and \(b_\perp\)**

Let us consider now the orthogonality in \(L^2\) of the two terms obtained from the Helmholtz decomposition of the drift \(b\)

\[ b = b_\parallel + b_\perp = -\nabla \phi + \nabla \times H. \]  

(3.78)

A corollary from the divergence theorem states that

\[ \int_\tau [G \cdot (\nabla \times F) - F \cdot (\nabla \times G)] du = \int_\zeta (F \times G) \cdot d\zeta, \]

for vector fields \(F, G\) and the “boundary” of the torus \(\zeta = \partial \tau\). Taking the inner product of \(-\nabla \phi\) and \(\nabla \times H\) and the above corollary gives

\[ \int_\tau (-\nabla \phi) \cdot (\nabla \times H) du = \int_\tau H \cdot (\nabla \times \nabla \phi) du + \int_\zeta H \times (\nabla \phi) \cdot d\zeta = 0, \]  

(3.79)

where the identity \(\nabla \times (\nabla \phi) = 0\) applies to the first integral and the surface integral goes to zero due to periodic boundary conditions imposed by our parameterisation of the torus \(\tau\).
3.6. Non-parametric drift inference: dimension greater than one

The un-conditioned prior log-density \((3.39)\) can then be expanded as

\[
\tau (\|b\| + b_\perp) (u) \cdot (A_0 (\|b\| + b_\perp)) (u) du = \int \tau (\|b\|) (u) \cdot (A_0 b_\parallel) (u) + b_\perp (u) \cdot (A_0 b_\perp) (u) du,
\]

where the mixed terms \(b_\parallel \cdot A_0 b_\perp\) and \(b_\perp \cdot A_0 b_\parallel\) vanish in the same vein as \((3.79)\) to give

\[
\int \tau (\|b\|) (u) \cdot (A_0 b_\perp) (u) du = \int (-\nabla \phi) (u) \cdot (\Delta (\nabla \times H)) (u) du,
\]

\[
= \int (-\nabla \phi) (u) \cdot (\nabla (\Delta \times H)) (u) du,
\]

using the additional identities for the Laplacian operator present in the prior precision operator \(A_0\) given by

\[
\Delta (\nabla \times H) = \nabla \times (\Delta H), \quad \Delta (\nabla \phi) = \nabla (\Delta \phi).
\]

We have shown that \(b_\perp\) and \(b_\parallel\) are independent a priori, but will show now that this is not the case a posteriori. To attempt a similar derivation of \((3.80)\)–\((3.82)\) for the un-conditioned Gaussian posterior \((3.47)\), let us assume just for this particular case that the Radon-Nikodym derivative of the empirical measure \(\Gamma_T\) with respect to the Lebesgue measure \(d\Gamma_T\) exists. We know this to be not true, since \(\Gamma_T\) is not absolutely continuous with respect to the Lebesgue measure, but we tolerate this transgression here to illustrate that the orthogonality of the Helmholtz decomposition terms do not hold under \(\Gamma_T\).

We treat \(\Gamma_T\) as piecewise constant and \(\nabla \Gamma_T\) in the distributional sense. Let us
now focus on the term in (3.45) involving the empirical measure

\[
\int b_\parallel(u) \cdot b_\perp(u) \Gamma_T(u) du = \int (-\nabla \phi) \cdot (\nabla \times H) \Gamma_T(u) du,
\]

\[
= \int \phi \cdot [\nabla \cdot (\nabla \times H) \Gamma_T(u)] du,
\]

\[
= \int \phi \cdot [(\nabla \cdot (\nabla \times H)) \Gamma_T(u) + (\nabla \times A) \cdot (\nabla \Gamma_T(u))] du,
\] (3.83)

where the first term in the sum vanishes due to the identity \( \nabla \cdot (\nabla \times A) = 0 \) but the second term may persist since the derivative of \( \hat{\Gamma}_T \) is non-zero somewhere. For example, let us suppose that we have \( \nabla \Gamma_T \neq 0 \) at some subspace \( P \subset \tau \), and some \( (\phi, H) \) such that \( \phi \neq 0 \) only in the vicinity of \( P \). Hence (3.83) will be non-zero for this choice of \( (\phi, H) \) whereas (3.82) will be zero regardless of the choice of \( (\phi, H) \). This result can also be reached by making a similar integration-by-parts argument starting from the discretisation (3.48), but we digress.

Since (3.83) may be non-zero, we have shown that the empirical measure \( \Gamma_T \), and therefore the posterior precision operator characterised by \( ((\cdot, A_0)_{L^2} + (\cdot, S^{-1})_{\Gamma_T}) \), does not factorise. A posteriori, \( b_\perp \) and \( b_\parallel \) will not be independent, as long as \( \nabla \Gamma_T \neq 0 \), which is generally the case.

**Conditioning versus projecting**

Let us denote the alternative posterior mean presented at the start of this section as the projected posterior mean \( b_p \), such that

\[
b_p = P_\parallel (A_0 + G_T)^{-1} r.
\] (3.84)

This projected posterior mean \( b_p \) differs from the conditioned posterior mean

\[
b_c = (A_0 + G_T)^{\dagger} P_\parallel r,
\]

due to the non-commutativity of the operators \( (A_0 + G_T)^{-1} \) and \( P_\parallel \).

We introduce here a projection operator \( R \) which maps any \( b \in V \) into a sub-
space $V_{\frac{1}{2}}$ of half the dimension of $V$, such that the projection $Rb \in V_{\frac{1}{2}}$ corresponds to the Helmholtz project projection $P\|b \in V$ where the only difference is in the dimension of the subspaces in which the projections reside. We can then write $P\| = R^\top R$ and therefore (3.76) as

$$R^\top R(A_0 + G_T)R^\top Rb = R^\top Rr.$$  (3.85)

Let us now write $A := (A_0 + G_T)$ for ease of notation, and write the conditioned posterior mean $b_c$ in terms of the Moore-Penrose inverse representation of $A^\dagger := (A_0 + G_T)^\dagger$.

**Proposition 1.** Suppose $A$ is positive definite symmetric. Then, the unique drift $b$ maximising the posterior under $P\|b = b$ is given by the Moore-Penrose inverse

$$b_c = R^\top (RAR^\top)^{-1} Rr.$$  (3.86)

**Proof.** Since $P\|b = b$, we can replace $b = R^\top \xi$, with $\xi \in V_{\frac{1}{2}}$, in the log-posterior of the drift (3.75) and expand it as

$$\log \pi(b|X, b_\perp = 0) = -\frac{1}{2} (R^\top \xi - A^{-1}r)^\top A(R^\top \xi - A^{-1}r),$$

$$= -\frac{1}{2} \xi R^\top A R^\top \xi + r R^\top \xi + \text{const.}$$

Maximising for $\xi \in V_{\frac{1}{2}}$ then leads to

$$R^\top \xi = R^\top (RAR^\top)^{-1} Rr.$$  

Note that the matrix $RAR^\top$ is positive definite symmetric on $V_{\frac{1}{2}}$, which follows from assuming $A$ being positive definite symmetric on $V$.  

Using (3.86) we can now establish the conditions under which the equality $b_c = b_p$ holds.

**Proposition 2.** The equality of the conditioned posterior mean and projected posterior mean $b_c = b_p$ holds if $AP\| = P\|A$ or $P\|A^{-1} = A^{-1}P\|$. More precisely, $b_c = b_p$
if and only if \( r \in \ker(R(P \parallel A - AP \parallel)A^{-1}) \).

**Proof.** Note that \( P \parallel = R^\top R \) is not invertible, whereas \( RR^\top : V_{\frac{1}{2}} \rightarrow V_{\frac{1}{2}} \) is invertible and is the identity on \( V_{\frac{1}{2}} \). Since both \( b_c \) and \( b_p \) are conservative, it suffices to consider \( R(b_c - b_p) \) since it is non-zero if and only if \( b_c - b_p \) is zero.

\[
R(b_c - b_p) = R(R^\top(RA)^{-1}R - R^\top RA^{-1})r,
\]

\[
= RR^\top(RA)^{-1}(R - RA^\top RA^{-1})r,
\]

\[
= (RA^\top)^{-1}(RR^\top RA - RA^\top R)A^{-1}r,
\]

\[
= (RA^\top)^{-1}R(P \parallel A - AP \parallel)A^{-1}r,
\]

where \( RR^\top = I \).

The requirement of \( AP \parallel = P \parallel A \) in Proposition \textbf{2} is equivalent to requiring \((A_0 + G_T)P \parallel = P \parallel (A_0 + G_T)\), and we have shown with \textbf{(3.83)} that \((A_0 + G_T)P \parallel \neq P \parallel (A_0 + G_T)\) in general. However, we can show how the equality might be achieved as special case.

Recall the log posterior of the drift after completing the square \textbf{(3.48)} which we give here as

\[
\log \pi(b|X) \propto -\frac{1}{2} [b - (A_0 + G_T)^{-1}r]^\top (A_0 + G_T) [b - (A_0 + G_T)^{-1}r],
\]

which leads to the posterior mean

\[
\hat{b} = \arg \max_b (\log \pi(b|X))_\text{disc},
\]

such that

\[
\log \pi(b|X) \propto -\frac{1}{2} [b - \hat{b}]^\top (A_0 + G_T) [b - \hat{b}].
\]

Let us now formally write the log posterior in its functional form

\[
\log \pi(b|X) = -\frac{1}{2} \int_{\tau} [b - \hat{b}]^\top (A_0 + S^{-1} \gamma_T) [b - \hat{b}] du,
\]
where the action of the empirical measure $\Gamma_T$-inner product and the squared diffusivity term $S^{-1}$ has been decoupled from $GT$ and $GT$ is resolved instead as a density $\gamma_T$ on the torus $\tau$. Of course, the empirical measure is not absolutely continuous with respect to the Lebesgue measure, and we ignore this impropriety for illustrational purposes here.

The drift conditioned to be conservative that maximises the log posterior can then be considered the solution, unique up to additive constant under periodic boundary conditions, to the following optimisation problem

$$\hat{\phi}_c = \arg\max_\phi -\frac{1}{2} \int_\tau (\nabla \phi - \hat{b})^\top [A_0 + S^{-1} \gamma_T] (\nabla \phi - \hat{b}) du,$$

(3.87)

whereas the Helmholtz projection of the posterior mean $\hat{b}$ solves

$$\hat{\phi}_p = \arg\max_\phi -\frac{1}{2} \int_\tau (\nabla \phi - \hat{b})^\top (\nabla \phi - \hat{b}) du.$$

(3.88)

The solution $\hat{\phi}_c$ in (3.87) is found by minimising the distance between $\nabla \phi$ and $\hat{b}$ in the $L^2(\tau, [A_0 + S^{-1} \gamma_T] du)$ sense, as opposed to the $L^2(\tau, du)$ sense for solution $\hat{\phi}_p$ in (3.88). In statistical terms, the deviation of $\nabla \phi$ from $\hat{b}$ is penalised uniformly for the Helmholtz projection (3.88), whereas for (3.87) deviations of $\nabla \phi(u)$ away from $\hat{b}(u)$ at every point $u$ is penalised differently depending on the direction of the deviation. Directions corresponding to the larger eigenvalues of $S^{-1} \gamma_T$ are penalised strongly whereas directions corresponding to the smaller eigenvalues are least penalised.

This phenomenon can be observed in Figures 3.8 and 3.9, where we have sampled paths from the toy diffusion process from Section 2.1 with two different diffusivity coefficients and estimated their drifts for both the unconditioned drift, which we denote here by $b_{\text{full}} = b_{\parallel} + b_{\perp}$, and conservative drift cases.

For the results in Figure 3.8, the diffusivity $\sigma$ of the toy diffusion process is set to be a multiple of the identity, which in this case is $\sigma = (1.4, 0; 0, 1.4)$. Figure 3.8(a) displays the true drift of the toy diffusion process, and Figure 3.8(b) is an estimate of the unconditioned drift $\hat{b}_{\text{full}}$. Unsurprisingly they are both quite similar.
3.6. Non-parametric drift inference: dimension greater than one

Figure 3.8: The toy diffusion process with diffusivity as a multiple of the identity $\sigma = (1.4, 0; 0, 1.4)$ was used to generate the input trajectory used for drift estimation. We have here (a) the true drift, (b) an estimate of the unconditioned drift $\hat{b}_{\text{full}}$, (c) the projection of $\hat{b}_{\text{full}}$, and (d) an estimate of the drift conditioned to be conservative $\hat{b}_c$.

Figure 3.8(c) is the projection of the unconditioned drift estimate $P_{\|} \hat{b}_{\text{full}}$, and Figure 3.8(d) is an estimate of the drift conditioned to be conservative $\hat{b}_C$. The similarities of $P_{\|} \hat{b}_{\text{full}}$ and $\hat{b}_c$ is due to the choice of $\sigma$ here, such that perturbations attributed to the Wiener process in the toy diffusion process will be dimensionally independent. This results in a sampled path, and in turn the empirical measure, which exhibits a rotational symmetry.

Let us now consider the toy diffusion process with diffusivity set to $\sigma = (1.4, -0.25; -0.25, 1.4)$, with the estimates and projection of the drift displayed in Figure 3.9. While the estimate of the unconditioned drift $\hat{b}_{\text{full}}$ and its projection $P_{\|} \hat{b}_{\text{full}}$ in Figures 3.9(b) and (c) remain similar to the above case where the diffusivity is set to a multiple of the identity, the estimate of the conditioned drift $\hat{b}_C$ in
3.6. Non-parametric drift inference: dimension greater than one

![True drift](a)

![Estimated drift](b)

![Projected estimated drift](c)

![Conditioned estimated drift](d)

Figure 3.9: The toy diffusion process with non-symmetric diffusivity $\sigma = (1.4, -0.25; -0.25, 1.4)$ was used to generate the input trajectory used for drift estimation. We have here (a) the true drift, (b) an estimate of the unconditioned drift $\hat{b}_{\text{full}}$, (c) the projection of $\hat{b}_{\text{full}}$, and (d) an estimate of the drift conditioned to be conservative $\hat{b}_c$.

Figures 3.9(d) is no longer rotationally symmetric. It is readily clear that the perturbations of the sampled path are no longer dimensionally independent, leading to an empirical measure that is not rotationally symmetric as well.

Preconditioning for conservative drift

To numerical solve the system of equations associated with (3.77), a conjugate gradient algorithm is applied, which in turn requires the use of a preconditioner, as established in Section 3.6.4. We note that we are in the same setting as Section 3.6.4, but will circumvent the heavy notation in favour of ease of exposition here.
3.6. Non-parametric drift inference: dimension greater than one

For a system of equations given by

\[ P^T A P \parallel b = P \parallel r, \]

where \( A = (A_0 + \Gamma_T) \), we can apply the preconditioner \( M^{-1} = CC^\top \), so that

\[ C^\top P^T A P \parallel CC^{-1}b = C^\top P \parallel r. \]

A naive choice of preconditioner \( C \) would be to take the incomplete Cholesky factorisation of \((P^T A P \parallel)^\dagger\), but since this is not directly available, we use instead the Moore-Penrose inverse given by (3.86), such that

\[ (P^T A P \parallel)^\dagger \approx R^\top (RAR^\top)^{-1}R. \]

We can then set the preconditioner \( C \) to

\[ C = A_{0,-j}^{-1/2} + L^\top \text{chol}(LAL^\top)^{-1}L, \quad (3.89) \]

where \( L \) is a low rank version of the projector matrix \( R \), associated with the first \( j \) Fourier modes, and \( A_{0,-j}^{-1/2} \) is the matrix \( A_0^{-1/2} \) with the first \( j \) terms of its diagonal entries set to zero.
3.7 Composite methodology

In this section we consolidate the inference components described by Sections 3.2, 3.3, 3.4 and 3.6 into a composite Bayesian inference methodology, whereby non-parametric inference for the drift and inference for the constant diffusivity is performed for a multi-dimensional diffusion process model, given a partially observed trajectory.

We reproduce here Algorithm 6 from Section 3.3.1 with some modifications, to give the Metropolis-within-Gibbs sampler loop which forms the core of our Bayesian inference methodology.

Algorithm 9. Composite Bayesian inference methodology

1. Target: The $P(Z, b, \sigma | \{X_i\})$ distribution.

2. Initialisation: Set $Z^{(0)}, b^{(0)}, \sigma^{(0)}$ at some initial value.

3. Iteration $j + 1$: Current state $Z^{(j)}, b^{(j)}, \sigma^{(j)}$.
   
   (a) Sample $\sigma^{(j+1)}$ from $P(\sigma | Z^{(j)}, \{X_i\}, \alpha)$ and set
       $V = h(Z^{(j)}; \alpha, \{X_i\}, \sigma^{(j+1)})$.

   (b) Sample $b^{(j+1)}$ from $P(b|\eta^{-1}(V))$ and set
       $\alpha(v) = \sigma^{-1}b(\sigma v)$

   (c) Sample $X_{\text{aug}}$ from $P(X_{\text{aug}}|\{X_i\}, b^{(j+1)}, \sigma^{(j+1)})$ and set
       $Z^{(j+1)} = g(\eta(X_{\text{aug}}))$.

4. The $(Z, b, \sigma)$ samples then come from the $P(Z, b, \sigma | \{X_i\})$ distribution.

To illustrate the use of Algorithm 9, we once again refer to the toy diffusion process from Section 2.1, where the diffusivity has been set to $\sigma = (1.5, 0.4; 0.4, 1.2)$ . A single path is simulated from the toy diffusion process with size $n = 801$ and total time $T = 20$, and is shown in Figure 3.10.

We will run Algorithm 9 twice given this trajectory, first for the diffusion process model with unconditioned drift, and then for the model with drift conditioned to be conservative, with results for both displayed in Figure 3.11.
According to Figures 3.11(a)-(b), we can see that the computational times for the diffusion bridge and diffusivity samplers are indistinguishable for both models, but the computational time for the unconditioned drift sampler is nearly twice that of the conservative drift sampler.

For the unconditioned drift model, the posterior mean of the drift and the posterior density of the diffusivity provide a good estimate of the true drift and diffusivity, as seen in Figures 3.11(c) and (d). For the conservative drift model however, the posterior mean of the drift is unsurprisingly a poor estimate of the true drift, since the true drift does in fact have a strong rotational component. As a consequence, the posterior density of the diffusivity is overestimated. We can see these results in Figures 3.11(d) and (f).
Figure 3.11: These are the results of the composite Bayesian inference methodology for the models where the drift is unconditioned (left column) and where the drift is conditioned to be conservative (right column). (a)-(b) The computational times of the individual Metropolis-within-Gibbs sampler components, which are respectively the diffusion bridge, diffusivity and drift samplers. (c)-(d) The posterior mean of the drift. (e)-(f) The posterior densities of the diffusivity.
3.8 MCMC diagnostics

In this section we will evaluate the performance of our composite Bayesian inference methodology, which was established in Section 3.7 as Algorithm 9, by investigating the convergence diagnostics of its individual components as well as its overall performance.

3.8.1 Diffusion bridge

Given known low frequency observed points, drift and diffusivity terms, step 3.(c) of Algorithm 9 consists of sampling diffusion bridges from \( P(X_{\text{aug}}|\{X_i\}, b, \sigma) \), which in turn relies on the use of the random walk Metropolis-Hastings diffusion bridge sampler described by Algorithm 5.

Trace and autocorrelation plots

To evaluate Algorithm 5, a single path consisting of only the start- and end-points over a time interval \([0, T]\) is sampled from the toy diffusion process from Section 2.1. Diffusion bridges consisting of 10 observed points (including the given start- and end-points) are then sampled over 1000 iterations according to Algorithm 5. As shown in Figure 3.12, the trace plots for the middle-points of the sampled diffusion bridges show good mixing, with autocorrelation decaying almost immediately.

This procedure is then repeated for diffusion bridges consisting of 100 and 1000 observed points and provides similar results as shown in Figure 3.12, which suggests that a relatively low frequency augmentation using diffusion bridges can be adequate if a lower computational cost is desired.

Acceptance rate and step size

We have chosen a step-size \( \rho = 0.3 \) for the random walk update step in Algorithm 5. This corresponds to average acceptance rates of 0.8317, 0.7997 and 0.6170 when sampling diffusion bridges where the underlying diffusion processes have:

1. a drift that is a sample from a Gaussian process with mean zero and precision operator \( \mathcal{A} = \eta(-\Delta)^k + \epsilon \), where \( k = 4, \eta = 0.005 \) and \( \epsilon = 0.05 \),
3.8. MCMC diagnostics

![Image of trace plots and autocorrelation plots for X and Y mid-points of diffusion bridges with 10, 100, and 1000 observed points.](image)

**Figure 3.12**: The trace plots (left) of the x- (top) and y-components (bottom) of the mid-point of diffusion bridges with 10, 100 and 1000 observed points. The autocorrelation plots (right) correspond to the diffusion bridge with the worst autocorrelation, which in this case is the one with 10 observed points.

2. an Ornstein-Uhlenbeck drift of form \( \alpha (\beta - X_t) \), where \( \beta = (4.5, 4.5)' \), \( \alpha = (3, 0.5; 0.5, 3) \), and

3. the drift of the toy diffusion process from Section 2.1

and diffusivity \( \sigma = (1.5, 0.4; 0.4, 1.2) \).

### 3.8.2 Drift

To evaluate the performance of the drift inference procedure, we turn to steps 3.(b) and (c) of Algorithm 9 which alternate between sampling diffusion bridges from \( \mathbb{P}(X_{\text{aug}} | \{X_i\}, b, \sigma) \) and sampling the drift from \( \mathbb{P}(b | \eta^{-1}(V)) \) given an observed path and known diffusivity, where the latter additionally relies on the numerical solver in Section 3.6.4 and the Gaussian sampler given by Algorithm 8.
3.8. MCMC diagnostics

Trace and autocorrelation plots

A single path is sampled from the toy diffusion process, with diffusivity $\sigma = (1.5, 0.4; 0.4, 1.2)$, $n = 4001$ observed points and total time $T = 200$. Given that the diffusivity is known, we alternate between sampling diffusion bridges and sampling the drift from its Gaussian posterior, via steps 3.(b) and (c) of Algorithm 9, over 200 iterations. This was then repeated for different initial states of the drift.

Figure 3.13: The trace plots (left) of the x- (top) and y-components (bottom) of the drift evaluated at $(5.7656, 5.7656)$, with traces for the different initial states. The autocorrelation plots (right) correspond to the trace with the worst autocorrelation.

Figure 3.13 (left) displays the traces of the drift evaluated at a single point $(5.7656, 5.7656)$. The traces shows good mixing of the drift samples stabilising around the true value of the drift with an almost immediate decay of autocorrelation, despite the unfavourable initial states chosen for the sampler (up to 3 orders of magnitude larger than the true value). We note that this point $(5.7656, 5.7656)$ was
selected in such a way as to be within the neighbourhood of an observed data point. While not shown here, we obtain comparable results for the trace and autocorrelation plots of other similarly selected points.

Prior reproduction

To further evaluate the performance of the drift inference procedure, we perform a prior reproduction test as follows:

Algorithm 10. Prior reproduction test

1. Fix $\sigma$. For $j = 1, \ldots, N$:

(a) Sample $\tilde{b}^{(j)}$ from $\sim \pi(b)$.

(b) Sample $\{X_i\}$ from $P(X|\tilde{b}^{(j)}, \sigma)$.

(c) Sample $b^{(j)}$ from $P(b|\{X_i\}, \sigma)$.

If the drift inference procedure is implemented correctly, the samples $\{b^{(j)}\}$ should be distributed according to the prior $\pi(b)$. Here we set $\pi(b)$ to be the Gaussian measure with mean zero and precision operator $\mathcal{A} = \eta (-\Delta)^k + \epsilon$, where $k = 4$, $\eta = 0.005$ and $\epsilon = 0.05$. The diffusivity is set to $\sigma = (1.5, 0.4; 0.4, 1.2)$. The paths $\{X_i\}$ are sampled with $n = 4001$ and $T = 200$. Algorithm 10 is then run for $N = 500$ iterations.

Figure 3.14 shows a sample of the drift from the prior (top) with a corresponding sample from the posterior (bottom) for a randomly selected path. At a glance the prior and posterior samples both display similar sizes and shapes.

We perform a more thorough comparison, where the prior and posterior samples of the drift are compared by evaluating the samples at 200 randomly selected points on the grid and performing the two-sample Kolmogorov-Smirnov test (KS2-test) for each of the 200 points. To achieve this, 10000 samples were additionally drawn from the prior to reduce the dependence of $b^{(j)}$ on $\tilde{b}^{(j)}$. Figure 3.15 (top) provides a visual comparison of the marginal prior and posterior densities via a scatter plot of the samples evaluated at $(4.7813, 2.6719)$, where 500 samples were
chosen randomly out of the 10000 drawn from the prior and plotted against the $N = 500$ samples from the posterior. Figure 3.15 (middle) details the QQ-plots of the prior and posterior samples evaluated at $(4.7813, 2.6719)$, with the KS2-tests giving $p$-values of 0.45243 and 0.13095 for the $x$- and $y$-components respectively, which suggests that the marginal prior and posterior distributions do indeed match at that particular point.

Figure 3.15 (bottom) graphs the histograms of $p$-values obtained for the KS2-tests for prior and posterior samples evaluated at all 200 points, which further suggests that samples from the posterior distribution do indeed match those from the prior distribution.
3.8. MCMC diagnostics

![Drift b_x: prior vs. posterior samples at (4.7813, 2.6719)](image)

(a)

![Drift b_y: prior vs. posterior samples at (4.7813, 2.6719)](image)

(b)

![Drift b_x: prior vs. posterior samples at (4.7813, 2.6719)](image)

(c)

![Drift b_y: prior vs. posterior samples at (4.7813, 2.6719)](image)

(d)

![Drift b_x: p-values of KS2-tests](image)

(e)

![Drift b_y: p-values of KS2-tests](image)

(f)

**Figure 3.15:** Comparison of the prior and posterior distributions for the x- (left) and y- components (right) of the drift using scatter plots (top) and QQ-plots (bottom) for one random point on the grid, and histograms (bottom) of p-values of KS2-tests for 200 randomly sampled points on the grid. 10000 samples were drawn from the prior for the KS2-tests, 500 of which were coupled with the 500 posterior samples to generate the scatter plots.

The conservative drift case

We repeat the investigation of the trace and autocorrelation plots and prior reproduction test for the case when the drift inference procedure is conditioned to be
3.8. MCMC diagnostics

3.8.3 Diffusivity

As for the performance of the diffusivity inference procedure, we look at steps 3.(a) and (c) of Algorithm 9 which alternates between sampling diffusion bridges from $P(X_{\text{aug}}|\{X_i\}, b, \sigma)$ and sampling the diffusivity from $P(\sigma|Z, \{X_i\}, \alpha)$, where the latter utilises the diffusivity sampler given by Algorithm 7.

Trace and autocorrelation plots

A single path is sampled from the toy diffusion process, with diffusivity $\sigma = (0.5631, 0.0128; 0.0128, 0.6102)$, $n = 4001$ observed points and total time $T = 200$. Given that the drift is known, we alternate between sampling diffusion bridges and sampling the diffusivity, via steps 3.(a) and (c) of Algorithm 9 over 200 iterations. This was then repeated for different initial states of the diffusivity.

Figure 3.16 (left) displays the trace plots of the $\sigma_{1,1}$, $\sigma_{1,2}$ and $\sigma_{2,2}$ components of the diffusivity $\sigma = (\sigma_{1,1}; \sigma_{1,2}; \sigma_{2,1}, \sigma_{2,2})$. The trace stabilises at the true values of the diffusivity components, even for very different initial states, with the largest deviation requiring only 11 iterations before reaching the correct values. We have chosen to display the autocorrelation plot corresponding to the trace started at the largest value in Figure 3.16 (right). For other initial states with smaller values, the autocorrelation decays almost immediately.

Acceptance rate and step size

We have chosen a step-size $\rho = 0.03$ for the random walk update step, which corresponds to average acceptance rates of 0.5664, 0.5872 and 0.5574 when sampling diffusion bridges where the underlying diffusion processes have

1. a drift that is a sample from a Gaussian process with mean zero and precision operator $A = -\eta(-\Delta)^k + \epsilon$, where $k = 4$, $\eta = 0.005$ and $\epsilon = 0.05$,

2. an Ornstein-Uhlenbeck drift of form $\alpha(\beta - X_t)$, where $\beta = (4.5, 4.5)'$, $\alpha = (3, 0.5; 0.5, 3)$, and
Figure 3.16: The trace plots (left) of the $\sigma_{11}$ (top), $\sigma_{12}$ (middle) and $\sigma_{22}$ components of the diffusivity, with traces for the different initial states. The autocorrelation plots (right) correspond to the trace with the worst autocorrelation.

3. the drift of the toy diffusion process from Section 2.1

and diffusivity $\sigma = (0.5631, 0.0128; 0.0128, 0.6102)$. 
3.8. MCMC diagnostics

Prior reproduction

As with the drift inference procedure, we perform a prior reproduction test as described by Algorithm 10 for the diffusivity inference procedure, where the prior $\pi(\sigma)$ is set to the Inverse-Wishart distribution with $\nu = 12$ and $\psi = 9I$, and the drift is set to be a sample from the Gaussian measure with mean zero and precision operator $A = \eta(-\Delta)^k + \varepsilon$, where $k = 4$, $\eta = 0.005$ and $\varepsilon = 0.05$. The paths $\{X_t\}$ are sampled with $n = 4001$ and $T = 200$. Algorithm 10 is then run for $N = 500$ iterations.

![Prior densities of $\sigma$](a) ![Posterior densities of $\sigma$](b)

Figure 3.17: Prior (left) and posterior (right) densities of the diffusivity components $\sigma_{11}$, $\sigma_{12}$ and $\sigma_{22}$.

To reduce dependencies of the posterior samples on the prior, we additionally draw 10000 samples from the prior. Figure 3.17 then compares the prior and posterior densities of the diffusivity components, where they both present strong similarities with the prior densities displaying higher peaks.

For a visual comparison, Figure 3.18 (left) compares 500 out of the 10000 samples from the prior against the $N = 500$ posterior samples via a scatter plot. Figure 3.18 (right) displays the QQ-plots of the prior and posterior samples of the diffusivity, with KS2-tests giving $p$-values of 0.10398, 0.13632 and 0.19755 respectively for $\sigma_{1,1}$, $\sigma_{1,2}$ and $\sigma_{2,2}$. This suggests that the posterior distribution might indeed match the prior distribution.
3.8. **MCMC diagnostics**

![Diagram](image)

**Figure 3.18:** Comparison of the prior and posterior distributions for the $\sigma_{11}$ (top), $\sigma_{12}$ (middle) and $\sigma_{22}$ components of the diffusivity using scatter plots (left) and QQ-plots (right). 10000 samples were drawn from the prior for the KS2-tests, 500 of which were coupled with the 500 posterior samples to generate the scatter plots.

### 3.8.4 Metropolis-within-Gibbs

Finally, we evaluate Algorithm 9 as a whole by looking at the trace and autocorrelation plots of the diffusion bridge, drift and diffusivity samplers.
3.8. MCMC diagnostics

Trace and autocorrelation plots

A single path with \( n = 4001 \) observed points and total time \( T = 200 \) is sampled from a diffusion process where the drift is set to be a sample from the Gaussian measure with mean zero and precision operator \( \mathcal{A} = \eta (-\Delta)^k + \varepsilon \), where \( k = 4 \), \( \eta = 0.005 \) and \( \varepsilon = 0.05 \), and the diffusivity is set to a sample from the Inverse-Wishart distribution with \( \nu = 12 \) and \( \psi = 9I \). Algorithm 9 is then run for 200 iterations.

Figures 3.19, 3.20 and 3.21 showcase the trace and autocorrelation plots for a randomly selected diffusion bridge, the drift evaluated at a randomly selected point and the diffusivity. As with the results in Sections 3.8.1, 3.8.2 and 3.8.3, we see that the trace plots display good mixing, where the drift and diffusivity traces stabilise quickly towards the true values. The autocorrelation plots show almost immediate decay as well for all samplers.

Similar results are obtained for other randomly selected diffusion bridges and the drift evaluated at other random points. Therefore, we have reason to believe that our composite inference methodology that is Algorithm 9 performs reasonably well.

The conservative drift case

We repeat this investigation for the case when the inferred drift is conditioned to be conservative in Appendix A.2 with similar results.

![Figure 3.19: Trace (left) and autocorrelation (right) plots for the middle point of a randomly selected diffusion bridge.](image)
3.8. MCMC diagnostics

**Figure 3.20:** Trace (left) and autocorrelation (right) plots for the drift evaluated at a randomly selected point on the grid.

**Figure 3.21:** Trace (left) and autocorrelation (right) plots for the diffusivity.
3.9 Application

In this section, we will apply the composite Bayesian inference methodology described in Section 3.7 to a real data example coming from the area of animal movement ecology.

For motivation, let us consider the use of diffusion process models in the context of analysing animal movement data. Firstly, we refer to Shimatani et al.’s [105] statement about quantifying a conceptual framework for movement ecology, which states that a model of animal movement must have an ecological basis such that a broad range of movement patterns can be explained by the model, and that parameters used in the model must have ecological interpretations. Therefore there should be no surprise that within the context of ecological science, there is some interest in diffusion process models for describing the movements of animal species. The usage of diffusion processes to describe such systems were initially limited to simpler models such as the pure diffusion process \( dX_t = \sigma dW_t \) whereby the drift is set to zero, and the mean-reverting Ornstein-Uhlenbeck process \( dX_t = \alpha(\beta - X_t)dt + \sigma dW_t \) to study and estimate the home range of free-ranging animals. Otherwise considered are parametric models which give way to frequentist methods of estimating the underlying parameters, such as calculating the maximum likelihood estimates. Preisler et al. [91] summarise some of the earlier usage of diffusion process models by environmental scientists to describe movements of animal species, in particular those of female bark beetles, elk and deer.

However, the constraint to a parametric inference framework, as well as the difficulty of implementing diffusion process models, has hampered their wide-spread use by non-specialists. Thus the application of diffusion process models, in particular models that are not conventionally specified by parametric drift, as applied to animal movement, has not been extensively studied within the area of animal movement ecology.

As an alternative to the usual parametric diffusion process models, Brillinger [22] considered the gradient of a potential as the drift, that is the drift is a conservative vector field \( b = -\nabla \phi \). However this leads to restrictions on the animal
3.9. Application

movement to be modelled since conservative drift fields do not allow for itinerant, circular movement, for example as brought up by Wall et al.’s [121] analysis of elephant movement. This leads us once more to the idea of a non-parametric drift specification whereby its structure can be more appropriately influenced by the animal movement data via Bayesian inference, which is exactly the methodology outlined in this chapter.

3.9.1 Capuchin monkey movement data

A single female white-faced capuchin monkey, *Cebus capuchinus*, from Barro Colorado Island in Panama was fitted with a GPS tracker and tracked over a period of two months in the year 2009 as part of a pilot test of digital telemetry technology [36]. The location of the monkey was set to be recorded every 10 minutes, from 10:00 am to 00:00 am. An examination of the data reveals that due to recording errors or otherwise, the observations are available for 46 days, for a total of 3033 observations, with mean, minimum and maximum time differences of 11.98, 7 and 120 minutes between observations. A histogram of the time differences is shown in Figure 3.22(d).

The longitudinal and latitudinal location data of the capuchin monkey is scaled in such a way to fit within a torus parameterised by $[0,9]^2 \in \mathbb{R}^2$. The resultant location data is displayed below for the entirety of the 3033 observations in Figure 3.22(a), with a trajectory corresponding to observations recorded over a single day shown in Figure 3.22(b), and the same 3033 observations colour-coded to distinguish their time indexing in Figure 3.22(c) with the start-points of the trajectories in blue and the end-points in yellow.

Unconditioned drift

We can then fit to the data our diffusion process model with non-parametric drift $b$ and constant diffusivity $\sigma$ described by (3.38) for the 2-dimensional case. We set the hyper-parameters of the Gaussian measure prior precision operator as $\eta = 5 \times 10^{-7}$, $\varepsilon = 5 \times 10^{-3}$ and the regularity term at $k = 4$. The hyper-parameters of the Inverse Wishart prior are set to $\nu = 8$ and $\Psi = 6I$. We can now apply the composite
Bayesian inference methodology described by Algorithm 9 with the results given by Figure 3.23.

The posterior mean of the drift is presented in Figure 3.23(b). As suggested by the finite range of exploration by the data, there are strong boundary forces present all around, and the rotational aspect of the data present in Figure 3.22(c) manifests itself in the posterior mean of the drift as well. There also appears to be an attraction point in the top right of Figure 3.23(b), and we can see from the data that there is a corresponding concentration of observed points in the area, which may be the resting area of the capuchin monkey, as can be distinguished by the start and end-points of the trajectories.
3.9. Application

Figures 3.23(c) and (d) are the projections of the posterior mean estimate of the drift into the $x$- and $y$-dimensions, and the peaks correspond to the boundary, rotational and attracting components of the drift as identified from Figure 3.23(b).

Figure 3.23(e) displays the posterior densities of the diffusivity, and Figure 3.23(f) shows the computational costs of the components of the Metropolis-within-Gibbs sampler given by Algorithm 9. We note that compared to the toy diffusion process example in Section 3.7, the diffusion bridge sampler component here takes up a significantly longer amount of computational time, simply because the capuchin monkey data has nearly 4 times as many data points as in the toy diffusion process example, therefore there are 4 times as many diffusion bridges to be sampled.

Drift conditioned to be conservative

Now let us fit to the capuchin monkey movement data the diffusion process model where the drift is conditioned to be conservative, as in Section 3.6.6. Applying Algorithm 9 once more gives the results in Figure 3.24.

The posterior mean of the drift is shown in Figure 3.24(b). This time, the rotational aspect of the data is no longer present in the posterior mean when compared to the posterior mean of the unconditioned drift. Instead, the boundary forces become more prominent relative to the forces found in the interior. The attraction point in the top right is also more pronounced.

The peaks of the projections of the posterior mean drift into the $x$- and $y$-dimensions in Figures 3.24(c) and (d) now identify the boundary and attraction components from Figure 3.24(b).

Figure 3.24(e) displays the posterior densities of the diffusivity, and Figure 3.24(f) shows the computational costs of the components of the Metropolis-within-Gibbs sampler given by Algorithm 9. As with the toy diffusion process example in Section 3.7, we note that there is an improvement of the drift inference component when compared to that of the unconditioned drift in Figure 3.23(f).
3.9. Application

Figure 3.23: (a) Movement data of the capuchin monkey. (b) Posterior mean of the drift. (c), (d): $x$ and $y$-components of the posterior mean. (e): Posterior densities of the components of the diffusivity. (f) Computational times of the Gibbs sampler components.

A brief conclusion

When comparing both diffusion process models, it seems clear from the posterior means of the drift in Figures 3.23(b) and 3.24(b) that the rotational behaviour present in the capuchin monkey movement data should preclude the choice of drift
conditioned to be conservative. While the posterior mean of the unconditioned drift is unable to identify the attracting component as well as the posterior mean of the drift conditioned to be conservative, it seems reasonable to believe that the posterior mean of the unconditioned drift is better equipped to capture the characteristics of

Figure 3.24: (a) Movement data of the capuchin monkey. (b) Posterior mean of the drift conditioned to be conservative. (c), (d): x and y-components of the posterior mean. (e): Posterior densities of the components of the diffusivity. (f) Computational times of the Gibbs sampler components.

3.9. Application

141
the capuchin monkey movement data as a whole.

For a formal means of evaluating how well these two models fit the capuchin monkey movement data, we point forward to the next chapter which describes how model checking can be performed for diffusion processes.

### 3.9.2 MCMC diagnostics

As in Section 3.8.4, we can evaluate the performance of the composite inference methodology as applied to the capuchin monkey movement data by exploring the trace and autocorrelation plots of the diffusion bridge, drift and diffusivity samplers.

**Trace plots**

According to Figures 3.25 and 3.26, the randomly selected diffusion bridge and the drift evaluated at a random point display good mixing and an almost immediate decay of autocorrelation. However, Figures 3.27 show that the diffusivity displays slower autocorrelation decay than in the toy examples in Sections 3.8.1 and 3.8.3, although this can be mitigated by enacting a burn-in time of at least 5 MCMC iterations. For the results in Section 3.9.1, a burn-in time of 20 MCMC iterations was used.

![Trace plots](image)

**Figure 3.25:** Trace (left) and autocorrelation (right) plots for the middle point of a randomly selected diffusion bridge.

We obtain similar results for other randomly selected diffusion bridges and the drift evaluated at other randomly selected points. Therefore, as in the toy example in Section 3.8.4, we are reasonably certain that our composite inference methodology performs well for the capuchin monkey movement data.
3.9. Application

Figure 3.26: Trace (left) and autocorrelation (right) plots for the drift evaluated at a randomly selected point on the grid.

Figure 3.27: Trace (left) and autocorrelation (right) plots for the diffusivity.

The conservative case

We point forward to Appendix A.3 for further MCMC diagnostics for the case when the inferred drift is additionally conditioned to be conservative.
Chapter 4

Model checking

4.1 Introduction

4.1.1 Introduction to model checking

As with the previous chapter, suppose that we are provided with a low frequency, discretely observed \( d \)-dimensional trajectory. We fit to this data a \( d \)-dimensional diffusion process model specified as the solution \( X : [0, T] \rightarrow \tau \) to the stochastic differential equation

\[
dX_t = b(X_t)dt + \sigma dW_t, \quad X_0 = x_0,
\]

with unknown drift \( b : \tau \rightarrow \mathbb{R}^d \) and unknown constant diffusivity \( \sigma \in \mathbb{R}^{d \times d} \), where \( W \) is the \( d \)-dimensional Wiener process. In Sections 3.6 and 3.4, we provided methods to estimate both the drift \( b \) and the diffusivity \( \sigma \). In this chapter, we are now interested in determining how well our diffusion process model equipped with the estimated drift and diffusivity fits the provided data.

In this introductory section, we provide a brief overview of model checking or model validation in a general context, starting with the idea of informal model checking, followed by formal hypothesis testing from both the frequentist and Bayesian points of view. We also summarise some of the hypothesis tests available for evaluating diffusion process models, be it for parametric or non-parametric models. Linking back to our capuchin monkey movement data example in Section 3.9, we
also highlight some works from the animal movement ecology literature where diffusion process models have been increasingly utilised, with the caveat that model checking has been lacking in comparison.

**All models are wrong**

The concept that “all models are wrong” is common adage to in statistics, with a possible follow-up question being: just how far away from the truth can these models be before they cease to be illuminating and useful? Model checking is therefore an important component in statistical analysis in determining whether one or more models provide a good fit to the data at hand, and if not, in what way does it differ from the truth and can this difference be quantified?

Gelman et al. [57] provided an authoritative overview of the place of model checking in applied Bayesian statistics and argue the following:

“We do not like to ask, ‘Is our model true or false?’, since probability models in most data analyses will not be perfectly true. Even the coin tosses and die rolls ubiquitous in probability theory texts are not truly exchangeable. The more relevant question is, ‘Do the model’s deficiencies have a noticeable effect on the substantive inferences?’”

We bear this philosophy in mind for the latter parts of this chapter, and contend that model checking does not end at identifying discrepancies between model and data, but should be used to further improve the fit of the models considered.

**Informal model checking**

Prior to applying, or in the absence of, formal model testing methods, it is oftentimes simpler to check if the data satisfies certain properties of a model in consideration before actually implementing said model. For example, one could calculate the quadratic variation of a trajectory to determine if a diffusion process model would be suitable for it. Commonly used informal model checks also involve graphical checks, where for example one might plot the residuals from a fitted linear regression model to check for patterns or lack thereof. One can also consider posterior
predictive checks where observed data is plotted against simulated data to detect incongruencies. However, analysing raw data can be difficult, especially for large data sets, and this is where formal model testing becomes useful.

4.1.2 Hypothesis testing

Hypothesis testing: the frequentist approach

For the majority of this chapter we will be focussing on one of the most commonly used formal model testing methods: goodness-of-fit hypothesis testing.

The hypothesis testing problem under the Neyman-Pearson framework (see Young and Smith [124, p.65] or Cox and Hinkley [35, p.64]) can be formulated by considering a model \( f(X; \psi) \) with parameter space \( \Psi \) containing disjoint subsets \( \Psi_0 \) and \( \Psi_1 \). The parameter \( \psi \in \Psi \) can be partitioned into \( \psi = (\theta, \nu) \) where \( \theta \) is the primary parameter and \( \nu \) is the nuisance parameter. The null and alternative hypotheses are given as

\[
H_0 : \psi \in \Psi_0 \text{ vs. } H_1 : \psi \in \Psi_1.
\]

Next, a function of the data is chosen as a test statistic \( T(X) \), leading to the \( p \)-value

\[
p = \sup_{\psi \in \Psi_0} P\{T(X) > T(x) | H_0\},
\]

or equivalently

\[
p = \sup_{\psi \in \Psi_0} P\{T(X) > T(x) | f(X; \psi), \psi \in \Psi_0\},
\]

as a measure of compatibility between the data and the model under the null hypothesis. When the nuisance parameter \( \nu \) is equal to zero under the null hypothesis, this \( p \)-value will in principle follow the uniform distribution, which makes it straightforward to determine whether or not to reject the null hypothesis under the usual significance level regime.

However, the presence of nuisance parameter \( \nu \) will introduce systematic un-
certainties, which in turn changes the distribution of the \( p \)-value. The conditioning method, supremum method, confident set method and bootstrap method are some common methods (see for example Demortier \[43\]) used to reduce the influence of the nuisance parameter on the \( p \)-value.

**Plug-in (bootstrap) \( p \)-value**

In particular, Davison and Hinkley’s \[41\] bootstrap method combined with the plug-in principle described by Efron and Tibshirani \[47\], whereby the nuisance parameter is estimated point-wise (usually with the MLE), is often used due to its simplicity. When properly performed, this method should preserve the uniformity of the distribution of the \( p \)-value.

This plug-in \( p \)-value is given as

\[
p_{\text{plug}} = P\{T(X) > T(x)|\hat{H}_0\} = P\{T(X) > T(x)|f(X|\hat{\psi}), \hat{\psi} \in \Psi_0\},
\]

and is taken over \( f(X|\hat{\psi}) \), where \( \hat{\psi} \) is an estimate of \( \psi \). In our application with the null hypothesis and test statistic to be described in Sections \ref{sec:4.2.1} and \ref{sec:4.2.2} we set \( \hat{\psi} \) to be the penalised MLE which coincides with the posterior mean of \( \psi = (b, \sigma) \). In particular, the penalised likelihood for the drift \( b \) has form

\[
l_{\text{pen}}(b) := l(b) - \frac{1}{2} \int b A_0 b.
\]

We note that the choice of penalisation will affect the distribution of \( p_{\text{plug}} \).

Additionally, the plug-in \( p \)-value uses the data twice, once when estimating the nuisance parameter with \( \hat{\psi} \) and once again when evaluating \( p_{\text{plug}} \) itself, which would result in a \( p \)-value which tends towards not rejecting the null hypothesis. Bayarri and Berger \[13\] argue that the plug-in \( p \)-value fails to account for the uncertainty involved in estimating the nuisance parameter, and that this problem is “rather involved”. As demonstrated by Demortier \[43\] and Robins \[96\], the other methods mentioned above (conditioning, supremum and confident set methods) also lead to their own share of problems, which we do not elaborate here.
4.1. Introduction

Hypothesis testing: the Bayesian approach

The problem of having a nuisance parameter becomes more manageable under a Bayesian framework, since the \( p \)-value can suitably modified to include reference densities other than \( f(X; \hat{\psi}) \), \( \hat{\psi} \in \Psi_0 \), as with the plug-in bootstrap method. This allows the consideration of Bayesian prior-posterior distribution models or the removal of the nuisance parameter via integration. Robins et al. [96] provided an overview of such Bayesian \( p \)-values and investigate their asymptotic distributions as compared with the plug-in \( p \)-value, but we will later in Section 4.3 focus on the Bayesian discrepancy \( p \)-value.

In particular, Meng [76] introduced the discrepancy statistic \( D(X, \psi) \) as a replacement for the test statistic \( T(X) \). It is a function of both the data and parameters, and gives rise to the discrepancy \( p \)-value

\[
p_{\text{disc}} = P\{D(X, \psi) > D(x, \psi) \mid f(X \mid \psi) \pi(\psi \mid x)\}.
\]

This discrepancy \( p \)-value is taken over the joint posterior distribution of \((X, \psi)\) and allows one to measure the discrepancy between sample and population quantities when checking the discrepancy between the data and the assumptions made.

Unlike uniformly distributed frequentist \( p \)-values however, the concept of model rejection based on the comparison of the evaluated discrepancy \( p \)-value with a chosen significance level does not necessarily make much sense. Gelman [56] illustrated how the traditional choice of the significance level does not have much value in the Bayesian setting by showing that the discrepancy \( p \)-value will not necessarily follow a uniform distribution. However, since we do not wish to strictly adhere to the arbitrariness involved with significance level-based model rejection, and instead wish to include the information provided by the Bayesian inference paradigm in the form of the posterior distribution of parameters, the discrepancy \( p \)-values provides a suitable alternative. Of course, a fully Bayesian approach to hypothesis testing such as Bayes factors might be prefereable but that lies beyond the scope of this thesis.
4.1.3 Goodness-of-fit testing for diffusion process models

Informal model checking

Diffusion process models have long been used in mathematical finance to describe the dynamics of interest rate derivatives, price of European-style options and so on. The earliest such models include the Black-Scholes, Vasicek, and Cox-Ingersoll-Ross models. As a holdover from the initial papers describing these models, attempts at constructing goodness-of-fit testing methodology or fitting improved models have often included them as baseline graphical comparisons.

Tests for model fit beyond that proposed by Aït-Sahalia [3] and Hong and Li [66] are not prevalent, once the model becomes more sophisticated and too complex to be described by the usual parametric models. For example, in the absence of applicable formal tests in their setting, García-Portugués [55] evaluated the goodness-of-fit of their parametric model via graphical comparisons of the parametric fits of the drift and diffusion coefficients with their non-parametric estimates.

Hypothesis testing

Goodness-of-fit testing literature for diffusion process models tend to focus on testing parametric-versus-parametric or parametric-versus-non-parametric hypotheses. The tests themselves are oftentimes parametric, for example maximum likelihood ratio tests which relies on maximum likelihood estimators that generally do not exist for non-parametric models.

Severini and Wong [104] improved upon MLE-based tests by constructing semi-parametric estimators and Fan et al. [51] extended this further with generalised likelihood ratio estimators which allows testing for non-parametric models. However, the Wilks phenomenon that allows its use has to be verified for the specific problems at hand. Fan [50] also provided a further overview of model validation methods for non-parametric-versus-parametric hypothesis tests.

Aït-Sahalia [3] introduced a non-parametric test for diffusion process models which compares the marginal density estimator of a parametric model with a non-parametric counterpart. However, Pritsker [92] showed that this test has poor fi-
nite sample performance due in part to persistent dependence in the data and slow convergence of the non-parametric density estimator. The use of a Vasicek model showed that the test also tends to over-reject the null hypothesis.

Building on Aït-Sahalia’s marginal density test, Hong and Li [66] developed a non-parametric test which compares instead the transition density of a parametric model with its non-parametric counterpart. This test improves upon the size and power performance of the marginal density test and is robust against persistent dependence in the data. Hong and Li also expect the non-parametric transition density estimator to perform well for finite samples.

Animal movement models in movement ecology

In a recent overview for animal movement modelling, Patterson et al. [86] argued that diffusion process models should be among the core components for quantitative researchers working on stochastic modelling of animal movement, in particular for animal movement data collected at the individual level. Historically, this has been the case, as can be seen by the various models used including the random walk model, Brownian bridges (Horne et al. [67]), Ornstein-Uhlenbeck position model (Dunn and Gipson [45], Börger et al. [20]), and diffusion process models with drift arising as the gradient of a potential function (Brillinger and Stewart [24], Brillinger et al. [23], Preisler et al. [90] [70]). Patterson et al. acknowledged that while there exists more sophisticated diffusion process models, these are rarely used within the context of animal movement ecology.

However, even with the increased usage of diffusion process models, investigations regarding model fit are not prevalent, with comparisons often made superficially in terms of whether a model is overly simplistic or otherwise. To illustrate this lack of statistically inclined criticism towards goodness-of-fit of diffusion process models in the literature, we turn to an example of how model fit in animal movement ecology is approached for the case of the Lévy-walk hypothesis used to model random foraging behaviour in animals where Pyke (2015) [93] argued based on biological inference, by scrutinising the assumption that the animals are indeed
searching for food or water by comparing their body lengths versus the step lengths observed by GPS technology.

4.1.4 Thesis contribution and problem statement

It has become apparent to us that while there exist several goodness-of-fit hypothesis testing methods for diffusion process models, there is also a lack of application of such methods in existing literature beyond those from mathematical finance. Even with these testing methods however, they are overwhelmingly in the frequentist camp, with few attempts made to provide a Bayesian alternative. Once model fit has been determined, improving model fit does not appear to be a main concern as well.

In this chapter we will attempt to tackle this deficiency via a holistic approach, first by setting up a goodness-of-fit hypothesis testing framework to allow the implementation of Hong and Li’s [66] transition density-based test in Section 4.2. Then as a compromise between the frequentist and Bayesian points of views, we modify our hypothesis testing framework to treat the transition density-based test statistic as a discrepancy variable instead, and move from the plug-in \( p \)-value to Meng’s [76] Bayesian discrepancy \( p \)-value in Section 4.3. We will then evaluate the model fit of the capuchin monkey movement data using both the plug-in and discrepancy \( p \)-values. Finally, in Section 4.5 we outline how model fit can be improved via identification of discrepancies between model and data. We do so by outlining informal model checks that are applicable to diffusion process models, constructing an outlier detection method, and establishing that identification of multi-scale properties in the data may allow the fitting of a diffusion process model.
4.2  Hypothesis testing

4.2.1  Historical hypothesis testing for diffusion processes

It is usually the case that both the null and alternative hypotheses have a common set of restrictions or underlying assumptions on the distribution of the random variable \( X \) under scrutiny. The union of the null and alternative hypotheses, therefore does not represent all possible variations on the distribution of \( X \). An obvious example is given by the classic hypothesis of specifying the mean of a normal population, which evidently makes the underlying assumption of the distribution of the sample either following a normal distribution or being i.i.d. and large enough such that the Central Limit Theorem holds. While these sort of assumptions are often taken for granted in the construction of simpler hypotheses, care must be taken as modifications to these assumptions will lead to modifications to the hypotheses themselves.

Therefore let us begin by stating the following assumption, as a modified reproduction of Definition 3.

**Assumption 1.** Let the true process \( X : \tau \rightarrow \mathbb{R}^d \) be a \( d \)-dimensional diffusion process specified as the solution to the stochastic differential equation

\[
dX_t = b_0(X_t)dt + \sigma_0 dW_t, \quad X_0 \sim \mathcal{X},
\]

where \( b_0 : \tau \rightarrow \mathbb{R}^d \) and \( \sigma_0 \in \mathbb{R}^{d \times d} \) are the true drift and diffusivity coefficients, and \( \mathcal{X} \) some probability measure on the torus \( \tau \).

This should make it clear that the data is expected to come from a diffusion process regardless of the hypothesis under question. We also adopt the convention of \( X_0 \sim \mathcal{X} \) here, and as a matter of convenience, will not be further concerned with it. In fact, all of the hypotheses and consequently the test statistics considered below will ignore the marginal density of \( X_0 \) completely. This is partly due to the fact that constructing a test which incorporates \( \mathcal{X} \) is difficult, and also due to the realisation of \( X_0 \) being a single data point is not very informative relative to the rest of the trajectory under scrutiny.
In existing literature on hypothesis testing for diffusion processes, as shown in Section 4.1.3, the drift and diffusivity are treated as nuisance parameters, and are often assumed to belong to some parametric family $\Psi_0$, such that $(b, \sigma) \in \Psi_0$. For example, the 1-dimensional Ornstein-Uhlenbeck diffusion process gives rise to $\Psi_0 = \{(b(x), \sigma) = (\alpha(\beta - x), \sigma); (\alpha, \beta, \sigma) \in R \subset \mathbb{R}^3\}$, for some $R$.

Thus for a given data set and Assumption 1, a diffusion process model with drift and diffusivity parameterised by $\Psi_0$ is then said to be correctly specified for $b_0$ and $\sigma_0$ under the null hypothesis

$$\mathbb{H}_0 : (b_0, \sigma_0) \in \Psi_0,$$  

where the alternative hypothesis is

$$\mathbb{H}_1 : (b_0, \sigma_0) \notin \Psi_0.$$  

These hypotheses can also be written as

$$\mathbb{H}_0 : b(x) = b_0(x) \text{ and } \sigma = \sigma_0 \text{ almost everywhere for some } (b, \sigma) \in \Psi_0,$$  

and

$$\mathbb{H}_1 : b(x) \neq b_0(x) \text{ or } \sigma \neq \sigma_0 \text{ for some } x \text{ and for all } (b, \sigma) \in \Psi_0.$$  

Since the drift and diffusivity of a continuous-time diffusion process uniquely characterises its marginal and transition densities, the hypotheses (4.3) and (4.4) can be written in terms of these densities.

Aït-Sahalia [3] proposed a marginal density-based hypothesis test for univariate diffusion process models with parametric drift and diffusivity. In addition to Assumption 1, the diffusion process is also assumed to be stationary. This joint parameterisation of $b$ and $\sigma$ corresponds to a parameterisation of the marginal and transition densities as well. For dimension 1, the marginal density is explicitly avai-
lable (see Durrett [46]) as
\[
p(x;b,\sigma) = \frac{\xi(b,\sigma)}{\sigma^2} \exp\left\{ 2 \int_{x_0}^{x} \frac{b(u)}{\sigma^2} du \right\},
\]
where \(\xi\) is the normalising constant given as
\[
\xi(b,\sigma) = \int_{\Omega} \frac{1}{\sigma^2} \exp\left\{ 2 \int_{x_0}^{x} \frac{b(u)}{\sigma^2} du \right\} \, dx.
\]
The null and alternative hypotheses are then stated as
\[
H_0 : p(x;b,\sigma) = p(x;b_0,\sigma_0) \text{ almost everywhere for some } (b,\sigma) \in \Psi_0, \quad (4.7)
\]
versus
\[
H_1 : p(x;b,\sigma) \neq p(x;b_0,\sigma_0) \text{ for some } x \text{ and for all } (b,\sigma) \in \Psi_0. \quad (4.8)
\]
where \(p(x;b_0,\sigma_0)\) is the true stationary density of the process. The null hypothesis given by \((4.7)\) can then be tested using a distance measure between the marginal densities, which is given as
\[
M = \min_{(b,\sigma) \in \Psi_0} \int_{\Omega} \left( p(x;b,\sigma) - p(x;b_0,\sigma_0) \right)^2 p(x;b_0,\sigma_0) \, dx, \quad (4.9)
\]
where \(p(x;b_0,\sigma_0)\) is estimated non-parametrically, for example by using a kernel estimator.

The marginal density alone however provides insufficient information to fully characterise a diffusion process. This means that while it is necessary for \((4.7)\) to be true when \((4.3)\) is true, the converse is not always the case. In particular, diffusion processes with the same marginal density may not necessarily have the same drift and diffusivity coefficients. The stationary assumption can also be too restrictive, as Pritsker [92] showed that for data with strong persistence the test has poor finite sample performance, due to the estimators utilised in the distance measure having slow convergence even with a twentyfold increase in amount of data.
Therefore, Aït-Sahalia additionally considered the formulation of a test statistic based on the transition density of a diffusion process. The corresponding null hypothesis is

\[ H_0 : p(x,t|y,s;b,\sigma) = p(x,t|y,s;b_0,\sigma_0) \text{ almost everywhere for some } (b, \sigma) \in \Psi_0, \]  

versus the alternative

\[ H_1 : p(x,t|y,s;b,\sigma) \neq p(x,t|y,s;b_0,\sigma_0) \text{ for some } t > s \text{ and for all } (b, \sigma) \in \Psi_0, \]  

where \( p(x,t|y,s;b,\sigma) \) is the transition density of \( X \). Note that this null hypothesis (4.10) is equivalent to (4.3).

Appealing to stationarity results so that

\[ p(x,t|y,s;b,\sigma) = p(x,t-s|y,0;b,\sigma) \]

and the forward and backward Kolmogorov equations

\[ \frac{\partial p(x,t|y,s;b,\sigma)}{\partial t} = -\frac{\partial p(x,t|y,s;b,\sigma)}{\partial s}, \]

hold (see Karatzas and Shreve [71]), Aït-Sahalia then utilised the transition discrepancy

\[ m(x,y,t-s;b,\sigma) = \frac{\partial p(x,t-s|y,0;b,\sigma)}{\partial t} - \frac{\partial p(x,t-s|y,0;b,\sigma)}{\partial s}, \]

which should be identically zero for all time steps \( t-s \), in constructing the relevant test statistic, which is given as

\[ T = \min_{(b,\sigma) \in \Psi_0} \int_\mathbb{X} \int_\mathbb{X} (m(x,y,t-s;b,\sigma))^2 p(x,t-s,y,0) \, dx \, dy, \]  

where the joint density \( p(x,t-s,y,0) \) is estimated non-parametrically using a kernel estimator. As with the marginal density based test, this test statistic relies heavily on the assumption that the underlying process has achieved stationarity, which may be not be true generally and may lead to similar problems encountered by Pritsker.

In conclusion, the marginal density-based test (4.9) relies on the parameterisation of the drift to construct the marginal density \( p(x;b,\sigma) \). While an explicit dependence on the parameterisation of the drift was made in the original work by
Aït-Sahalia for the transition density-based test (4.12), we note that this parameterisation need not enter the stationary results required to construct the test statistic. In light of our non-parametric approach to drift estimation, we remove this parameterisation restriction for the drift, so that the null and alternate hypotheses can be generalised for models with drift $b$ arising from some arbitrary (possibly non-parametric) family. For our diffusion process model with non-parametric drift and constant diffusivity, we have

$$
\Psi_0 = \left\{ (b, \sigma); b \in H^s_{\text{per}}, \sigma \in \mathbb{R}^{d \times d} \right\},
$$

(4.13)

where $s$ is chosen as in Section 3.6.1.

However, in allowing drifts of non-parametric form, the corresponding transition density of the diffusion process will not necessarily be parameterised. Obtaining its non-parametric estimate and consequently that of the forward and backward Kolmogorov equations will either become computationally expensive or will result in an estimate of Aït-Sahalia’s transition density-based test statistic (4.12) that is difficult to compute and does not converge for finite samples.

Thus moving forward, in Section 4.2.2 onwards, we will focus on implementing an alternative transition density-based testing method by Hong and Li [66] where the removal of the parameterisation restriction for the drift will not negatively affect its performance. The null hypothesis under consideration will be (4.10) in conjunction with the family (4.13). In Sections 4.2.2 we will reconstruct their test statistic for the 1-dimensional case, and for the multi-dimensional case in Section 4.2.3. We will consider improvements made for our implementation of the test statistic and explore the choice of bandwidth in Section 4.2.4. In Section 4.2.5 we investigate the finite sample properties of the test statistic and show that reliance on asymptotic properties is misguided for standard sample sizes encountered in practice. We investigate the power of the test statistic in Section 4.2.6. In Section 4.2.7 alternatives to the test statistic are considered. Finally, we consider transitioning towards a Bayesian approach by attaching a prior to our drift and diffusivity parameters, in line with the Bayesian inference methodology outlined in the pre-
vious chapter, which leads to a revision of the null hypothesis (4.10) in Section 4.3.1.

4.2.2 Transition density-based test statistic

As with Aït-Sahalia [3], Hong and Li [66] considered a transition density-based approach to hypothesis testing for diffusion process models. We will first outline this transition density-based test as originally conceived for a 1-dimensional diffusion process.

The generalised residuals

Given the 1-dimensional form of Assumption 1 and a discretely observed data set \( \{X_i\}_{i=1}^n \) fitted with the 1-dimensional diffusion process model described by (4.1), a dynamic probability integral transform can be applied to \( \{X_i\}_{i=1}^n \) to obtain the generalised residuals of the model given by

\[
Z_i = \int_{-\infty}^{X_i} p(x, t_i | X_{i-1}, t_{i-1}) \, dx.
\]  

Under the null hypothesis (4.10), the realised generalised residuals \( \{Z_i\}_{i=1}^n \) are i.i.d. \( U[0,1] \), in the same vein as the usual probability integral transform. Intuitively, for the generalised residuals (4.14) the i.i.d. property determines the Markov property of the underlying process, while its distribution by \( U[0,1] \) ensures that the process is indeed correctly specified by the non-parametric drift and diffusivity of the model. Therefore, the null hypothesis (4.10) can be tested by testing whether these generalised residuals \( \{Z_i\}_{i=1}^n \) are in fact i.i.d. \( U[0,1] \).

We note that the transition density of non-parametric diffusion processes are typically unavailable in closed form, so numerical methods are used instead to obtain the generalised residuals, which we discuss in Section 4.2.4.

The test statistic

The generalised residuals can in turn be paired into \( \{Z_i, Z_{i-j}\}_{j \geq 1}^n \) for some lag \( j < n \), \( j \in \mathbb{N} \), and these paired generalised residuals are i.i.d. \( U[0,1] \times U[0,1] \) under the
null, which allows the construction of a distance measure by comparing the densities of \( \{Z_i, Z_{i-j}\}_{i>j}^n \) and \( U[0, 1] \times U[0, 1] \) in the form of

\[
M(j) = \int_0^1 \int_0^1 [g_j(z_1, z_2) - 1]^2 \, dz_1 \, dz_2
\]

where \( g_j \) is the density of \( \{Z_i, Z_{i-j}\}_{i>j}^n \), which can for example be estimated non-parametrically using kernel density estimators. The test for independence in this density comparison only checks for pairwise independence of the terms in \( \{Z_i, Z_{i-j}\}_{i>j}^n \), and ideally we would want to perform this comparison for every lag \( j \). However the first few lags are the most informative and practical to test for.

The kernel density estimator of \( g_j \) is given by

\[
\hat{g}_j(z_1, z_2) = (n - j)^{-1} \sum_{i=j+1}^{n} K_h(z_1, Z_i)K_h(z_2, Z_{i-j}),
\]

(4.15)

where the kernel is given as

\[
K_h(x, y) = h^{-1}k\left(\frac{x - y}{h}\right),
\]

(4.16)

with the choice of quartic kernel function

\[
k(u) = \frac{15}{16}(1 - u^2)^2 \chi_1(\|u\| \leq 1),
\]

and bandwidth \( h = S_Z n^{-1/6} \) where \( S_Z \) is the sample standard deviation of \( \{Z_i\}_{i=1}^n \).

We note here that in the original work by Hong and Li [66] their chosen kernel contains a boundary-modification to account for biased estimates near the boundaries of the data, which is not required here, since we apply a convolution during the implementation of the kernel density estimator. Additionally, the choice of kernel function is not as important as that of the choice of bandwidth, and other kernel functions may be used here instead. We will discuss this convolution method and choice of bandwidth later in Section 4.2.4.

The kernel density estimator \( \hat{g}_j(z_1, z_2) \) is compared with the density of
Hypothesis testing

The test statistic \( \hat{Q}(j) \) is then given as the centred and scaled version of (4.17) such that

\[
\hat{Q}(j) = h[(n - j)\hat{M}(j) - A_0^h]/V_0^{1/2},
\]

with centering and scale factors

\[
A_0^h = \left[ (h^{-1}) \int_{-1}^{1} k^2(u)du \right]^2 - 1,
\]

and

\[
V_0 = 2\left[ \int_{-1}^{1} \left( \int_{-1}^{1} k(u+v)k(v)dv \right)^2 du \right]^2.
\]

Hong and Li then establish the following asymptotic result for the test statistic \( \hat{Q}(j) \):

**Theorem 5.** [66, p. 72] Let Assumptions 1–5 [66, p. 72] hold and \( h = cn^{-\delta} \) for \( c \in (0, \infty) \) and \( \delta \in (0, \frac{1}{2}) \). Then for any integer \( j > 0 \) such that \( j = o(n^{1-\delta(5-2v)}) \) where \( v \) is as in Assumption 1 [66, p. 72], we have

\[
\hat{Q}(j) \overset{d}{\to} N(0, 1)
\]

under the null given by (4.10).

Thus the null (4.10) can be tested by comparing \( \hat{Q}(j) \) against the \( N(0, 1) \) distribution.

### 4.2.3 The multi-dimensional case

Let us return now to the multi-dimensional case, where we are given Assumption 1 and a \( d \)-dimensional data set \( \{X_t\}_{i=1}^n = \{X_i\}_{i=1}^n \) is fitted with the \( d \)-dimensional diffusion process model described by (4.1).

The dynamic probability integral transform for the 1-dimensional case given by (4.14) is not directly applicable here. Hong and Li [66] consider instead a fac-
4.2. Hypothesis testing

torisation of the joint transition density of the \( d \) components \( X_i = (X_{1,i}, \ldots, X_{d,i})^\top \)

such that

\[
p(X_{1,i}, \ldots, X_{d,i}, t_i | X_{1,i-1}, \ldots, X_{d,i-1}, t_{i-1}) = \prod_{k=1}^d p(X_{k,i}, t_i | X_{k-1,i}, \ldots, X_{1,i}, X_{1,i-1}, t_{i-1}). \tag{4.22}
\]

This leads to the component-wise generalised residuals \( \{Z_{k,i}\}_{i=1}^n \), for \( k \in \{1, \ldots, d\} \), such that

\[
Z_{k,i} = \int_{-\infty}^{X_{k,i}} p(x_k, t_i | X_{k-1,i}, \ldots, X_{1,i}, X_{1,i-1}, t_{i-1}) dx_k, \tag{4.23}
\]

which are constructed using the dynamic probability integral transform of the marginal transition densities \( p(X_{k,i}, t_i | X_{k-1,i}, \ldots, X_{1,i}, X_{1,i-1}, t_{i-1}) \). As with the the 1-dimensional case, these component-wise generalised residuals \( \{Z_{k,i}\}_{i=1}^n \) should each be i.i.d. \( U[0, 1] \) under the null hypothesis (4.10). While there are \( d! \) ways of factoring the joint transition density (4.22), and therefore \( d! \) possible arrangements of (4.23), we need only consider one such possibility for practical purposes, the simplest of which is the ordering given by (4.22).

To evaluate the joint transition density (4.22), the combined generalised residuals \( Z_{\text{comb}} \) can also be constructed via the combination of the component-wise generalised residuals (4.23), given as

\[
Z_{\text{comb}} = \{Z_{1,1}, \ldots, Z_{d,1}, \ldots, Z_{1,n}, \ldots, Z_{d,n}\}. \tag{4.24}
\]

Once again, these combined generalised residuals \( Z_{\text{comb}} \) are i.i.d. \( U[0, 1] \) under the null (4.10).

Therefore, to test the null hypothesis (4.10) for our \( d \)-dimensional diffusion process model, the test statistic (4.18) can be evaluated and compared against the \( N(0,1) \) distribution, once for each set of component-wise generalised residuals \( \{Z_{k,i}\}_{i=1}^n \), for \( k \in \{1, \ldots, d\} \), and once more for the combined generalised residuals \( Z_{\text{comb}} \).

For rest of this chapter, we will focus on implementing Hong and Li’s transition density based test for the 2-dimensional diffusion process model, and thus will
only require \( \{Z_{1,i}\}_{i=1}^n \), \( \{Z_{2,i}\}_{i=1}^n \) and \( Z_{\text{comb}} = \{Z_{1,1}, Z_{2,1}, \ldots, Z_{1,n}, Z_{2,n}\} \), and we will refer to them as the first component, second component and combined generalised residuals respectively. We note however, that this implementation can be extended to the multi-dimensional case without difficulty.

### 4.2.4 Kernel density estimation, transition density approximation, and bandwidth selection

Kernel density estimation

Kern et al. [72] described a method to evaluate the kernel density estimator (4.15) at a reduced computational cost over its direct implementation given by Hong and Li [66].

A discretised grid is constructed on the \([0, 1] \times [0, 1]\) space of generalised residuals. The kernel density estimator (4.15) can be approximated by

\[
\hat{g}_j(z_1, z_2) = \frac{1}{n - j} \sum_{k=1}^N \sum_{l=1}^N N(s_k, s_l) K_h(z_1, s_k) K_h(z_2, s_l),
\]

where \( s_{kl} = (s_k, s_l) \) are the centre points of the \((k, l)\)-th grid cell, \( N(s_{kl}) \) refers to the number of paired generalised residuals \((Z_i, Z_{i-j})\) located in the \((k, l)\)-th grid cell. It can be seen that as the cell widths of the grid tend toward zero, this approximation will converge towards the estimator (4.15). The approximation (4.25) can then be written as the convolution

\[
\hat{g}_j = \frac{1}{n - j} (N * (K_h K_h)) = \frac{1}{n - j} \left[ \mathcal{F}^{-1}(\mathcal{F}(N) \cdot \mathcal{F}(K_h K_h)) \right].
\]

where \( \mathcal{F} \) and \( \mathcal{F}^{-1} \) are the Fourier and inverse Fourier transformations respectively and \( * \) denotes the convolution operation.

The original implementation by Hong and Li [66] involved boundary correcti-
4.2. Hypothesis testing

ons for the kernel, which we reproduce below as

$$K_h(x,y) = \begin{cases} 
  h^{-1}k\left(\frac{x-y}{h}\right) / \int^{1}_{-(x/h)} k(u)du, & \text{if } x \in [0,h), \\
  h^{-1}k(x/h), & \text{if } x \in [h, 1-h], \\
  h^{-1}k\left(\frac{x-y}{h}\right) / \int^{(1-x)/h}_{1} k(u)du, & \text{if } x \in (1-h, 1] 
\end{cases}$$ (4.27)

and an additional integral term in the centering factor (4.19) of the test statistic given as

$$A_0^h = \left( (h^{-1} - 2) \int^{1}_{-1} k^2(u)du + 2 \int^{1}_{-1} \int^{b}_{-1} \kappa^2(u)du db \right)^2 - 1,$$

where $\kappa_b(\cdot) = k(\cdot) / \int_{-1}^{b} k(v)dv$.

With the new implementation (4.26), this boundary-correction is no longer needed for both the kernel and the centering factor. However, the estimator (4.26) is now taken over a periodic grid, with the bias of the estimates near the boundaries reduced instead by the inclusion of zero-padding at the boundaries.

Transition density approximation

Since the transition density of non-parametric diffusion processes are typically unavailable in closed form, numerical methods are used instead to obtain the generalised residuals in Section 4.2.2.

Suppose that we have fitted the 2-dimensional diffusion process model described by (4.1) to a 2-dimensional trajectory $\{X_i\}_{i=1}^n$, and have performed the composite Bayesian inference methodology in Section 3.7. To compute the test statistic $\hat{Q}(j)$, we will first need to obtain the first component generalised residuals

$$Z_{1,i} = \int_{-\infty}^{X_{1,i}} p(x_1,t_i|x_{i-1},t_{i-1})dx_1,$$

and the second component generalised residuals

$$Z_{2,i} = \int_{-\infty}^{X_i} p(x_2,t_i|x_{1,i},X_{i-1},t_{i-1})dx_2.$$

The transition density of our diffusion process model with non-parametric drift is
4.2. Hypothesis testing

unavailable in closed form, and therefore we are unable to compute the generalised residuals directly. However, we can estimate these transition densities by simulating the end-points of the transitions and then estimating the resultant empirical density function.

Let us first estimate the joint transition density \( p(x_1, x_2, t_i | X_{i-1}, t_{i-1}) \). Given the posterior mean of the drift and diffusivity, we simulate \( M \) paths from the diffusion process model starting at the point \( X_{i-1} \) for \( t_i - t_{i-1} \) time units, using the Euler method described in Definition [10]. The joint transition density can then be estimated by collecting the \( M \) end-points \( X_{\text{end},m}^{\text{end},m} = (X_{1,1}^{\text{end},m}, X_{2,1}^{\text{end},m})^T \) of the simulated paths as \( \{X_{i}^{\text{end},m}\}_{m=1}^{M} \), and calculating its kernel density estimate using the estimator

\[
\hat{p}(x_1, x_2, t_i | X_{i-1}, t_{i-1}) = \frac{1}{nh} \sum_{m=1}^{M} K \left( \frac{x_1 - X_{1,1}^{\text{end},m}}{h} \right) K \left( \frac{x_2 - X_{2,1}^{\text{end},m}}{h} \right).
\]

The first component transition density \( p(x_1, t_i | X_{i-1}, t_{i-1}) \) can be estimated using a marginal kernel density estimator of the form

\[
\hat{p}(x_1, t_i | X_{i-1}, t_{i-1}) = \frac{1}{nh} \sum_{m=1}^{M} K \left( \frac{x_1 - X_{1,1}^{\text{end},m}}{h} \right),
\]

and the kernel conditional density estimator for the second component transition density is given by Chen et al. [29] as

\[
\hat{p}(x_2, t_i | x_1, X_{i-1}, t_{i-1}) = \frac{\hat{p}(x_1, x_2, t_i | X_{i-1}, t_{i-1})}{\hat{p}(x_1, t_i | X_{i-1}, t_{i-1})}.
\]

As a result, we are now able to calculate the generalised residuals by simply integrating over the kernel density estimates.

4.2.5 Finite sample performance of the test statistic

One of the assumptions for Theorem 5 in the setting where the diffusion process model is assumed have parametric coefficients is the existence of a \( \sqrt{n} \)-consistent estimator \( \hat{\psi} \in \Psi \) for the true parameter vector \( \psi_0 \), such that \( \sqrt{n}(\hat{\psi} - \psi_0) = O_p(1) \). In our setting where the drift is inferred non-parametrically for a multi-dimensional
4.2. Hypothesis testing

For the diffusion process model, we do not have access to such consistency results. We have also made adjustments to the original implementation of the kernel (4.27) by Hong and Li, to allow for the use of the kernel density estimator (4.26). The scaling factor (4.19) has also been modified as a result. As such, we are interested in investigating the finite sample behaviour of $\hat{Q}(j)$ for practical data sample sizes given these modifications.

**Mean and variance**

Let us denote by $\hat{Q}_V$ the unscaled version of $\hat{Q}$ (4.18), given by

$$\hat{Q}_V(j) = h((n - j)M(j) - A_0^0).$$

Since Theorem 5 states that $\hat{Q}(j) \xrightarrow{d} N(0, 1)$, we expect to see the equivalent result where $\hat{Q}_V(j) \xrightarrow{d} N(0, V_0)$.

![Figure 4.1: Finite sample mean (left) and variance (right) of the unscaled test statistic $\hat{Q}_V(1)$ versus uniform generalised residuals of size $n = 100 \times 2^l$, $l \in \{0, 1, \ldots, 25\}$. The uniform generalised residuals were simulated 10000 times for $l = 0, 1, \ldots, 19$, 1000 times for $l = 20$, 500 times for $l = 21, 22$ and 100 times for $l = 23, 24$ due to computational constraints. This is reflected in the size of their respective confidence intervals.](image)

As a preliminary check, we first study the finite sample mean and variance of this unscaled $\hat{Q}_V$. We focus on lag $j = 1$, and since under correct model specification the generalised residuals $\{Z_i\}_{i=1}^n$ follow the $U[0, 1]$ distribution, we can
substitute them with \( n \) draws from the \( U[0, 1] \) distribution when calculating \( \hat{M}(1) \). This bypasses any effects that would arise from our use of a non-parametric drift estimator with dubious consistency. As shown by Figure 4.1, the resultant sample mean of \( \hat{Q}_V(1) \) stabilises immediately around the asymptotic mean of zero. However, the sample variance of \( \hat{Q}_V(1) \) only began to approach the asymptotic variance of \( V_0 = 0.5234 \) for sample sizes greater than \( n = 100 \times 2^{20} \). For practical purposes then, it is necessary to apply a finite sample correction to \( \hat{Q}_V(1) \) by using its variance according to sample size to standardise it rather than simply scaling by \( 1/V_0^{1/2} \).

**Density and cumulative distribution function**

Let us now denote

\[
\hat{Q}_C := \frac{\hat{Q}_V}{\text{empvar}(\hat{Q}_V)}
\]

the test statistic normalised according to its finite sample variance \( \text{empvar}(\hat{Q}_V) \). With this correction in mind, we now check whether the test statistic \( \hat{Q}_C \) still has its intended distribution, that is, \( \hat{Q}_C \overset{d}{\to} N(0, 1) \) even for non-asymptotical sample size \( n \).

We compare the empirical distribution of \( \hat{Q}_C(1) \) against the standard normal distribution using Q-Q plots in Figure 4.2, where the data sample sizes considered are \( n = 800, 6400, 51200 \) and 409600. We have also calculated the \( p \)-values arising from a Kolmogorov-Smirnov test for standard normality. It is clear that as \( n \) increases, the empirical distribution of \( \hat{Q}_C(1) \) approaches its intended asymptotic distribution of \( N(0, 1) \), as can be seen from the improvement of both the Q-Q plots and Kolmogorov-Smirnov test \( p \)-values. However, this improvement is quite gradual, as evidenced by the \( p \)-value of 0.057033 in Figure 4.2(d) for the case of sample size \( n = 409600 \), which is more than 100 times the length of our capuchin monkey movement data.

Therefore, instead of evaluating the test statistic \( \hat{Q}_C \) against the standard normal distribution, we will instead compare \( \hat{Q}_C \) against its empirical distribution tailored for the size of the data in question. We will refer to this distribution as the
Figure 4.2: Q-Q plots comparing the distribution of $\hat{Q}_C(1)$ against the standard normal distribution for data sample sizes $n = 800, 6400, 51200$ and $409600$ (a, b, c, d), by simulating $\hat{Q}_C(1)$ 1000 times each using draws from the uniform distribution as a substitute for generalised residuals.

reference finite sample distribution of $\hat{Q}_C$.

**Distribution of the $p$-value**

In order to investigate the finite sample properties of $\hat{Q}_C$, we took advantage of the fact that under correct model specification, the generalised residuals of the data will follow the uniform distribution. However, this behaviour cannot be guaranteed to hold when the generalised residuals are computed for actual data of small sample sizes arising from the null model, which in turn may negatively impact the finite sample properties of $\hat{Q}_C$.

Thus we examine the empirical distributions of $\hat{Q}_C$ and its corresponding plug-in $p$-values when the generalised residuals have been computed using data obtained under correct model specification. This is carried out using Algorithm 11.
Algorithm 11. Target: Distribution of $p_{plug}$

1. Fix $b, \sigma$

2. Iterate from $i = 1$ to $m$:
   
   (a) Sample $x_i$ from $f(x_i|b, \sigma)$
   
   (b) Calculate $\hat{Q}_C(1)$ using $b, \sigma$ as the plug-in parameters
   
   (c) Calculate $p_{plug,i}$ by comparing $\hat{Q}_C(1)$ with its reference empirical distribution.

We set $b, \sigma$ to the drift and diffusivity of the toy diffusion process described in Section 2.1. Algorithm 11 is then run for 400 iterations, where trajectories are sampled from the toy diffusion process with size $n = 200$ and total time $T = 20$. The resultant empirical distributions of $\hat{Q}_C$ for the first, second component and combined generalised residuals are compared against their reference finite sample distributions (obtained using 10000 samples of $\hat{Q}_C$ using generalised residuals drawn from the uniform distribution) in Figure 4.3 (left column). Some deviations away from the reference distributions are observed, but the shape still fits relatively well. This is also reflected in the empirical distributions of $p_{plug}$ in Figure 4.3 (right column), which display slight deviations from the uniform distribution since the computation of $p_{plug}$ here involves comparing $\hat{Q}_C$ against its reference empirical distribution.

If so desired, instead of relying on the reference finite sample distribution for $\hat{Q}_C$ obtained by substituting generalised residuals under the null for samples from the uniform distribution, we can use the empirical distribution of $\hat{Q}_C$ obtained via Algorithm 11 as a reference distribution instead. Of course, this results in a higher computational cost. In later sections, when we refer to evaluations of $p_{plug}$ for $\hat{Q}_C$, we will in fact be using the reference finite sample distribution of $\hat{Q}_C$ based on uniform samples.
Figure 4.3: Left column: the kernel density estimates of $\hat{Q}_C$ evaluated using generalised residuals arising from simulated data (orange) and its reference finite sample distribution (blue). Right column: the histograms and kernel density estimates of the plug-in $p$-values obtained by comparing $\hat{Q}_C$ arising from simulated data against its reference finite sample distribution. These plots are arranged according to the (a)-(b) first component, (c)-(d) second component and (e)-(f) combined generalised residuals.
Choice of bandwidth

We take a further look at the choice of bandwidth \( h = S_n Z n^{-1/6} \) used by Hong and Li [66] in the kernel density estimator (4.15), which they claim attains the optimal rate for bivariate density estimation. We initially suspect that this may be in error, since according to Scott [103, p. 152], this bandwidth \( h \) was shown by to be the optimal rate for the estimation of bivariate densities only when using Gaussian kernels. Therefore we consider instead the equivalent bandwidth

\[
h_{op} = \sqrt{7} S_n n^{-1/6}
\]

which once again according to Scott is the optimal bandwidth for the quartic kernel function used in our current implementation. Bandwidths are generally considered optimal if they minimise the mean integrated squared error

\[
\text{MISE}(h) = \mathbb{E} \left[ \int (\hat{f}_h(x) - f(x))^2 \, dx \right]
\]

where \( f \) is the target, unknown density and \( \hat{f}_h \) the kernel density estimator with bandwidth \( h \). Thus to compare the two bandwidths, we look at the integrated square error of our testing framework given by (4.17) as

\[
\hat{M}(j) = \int_0^1 \int_0^1 [\hat{g}_j(z_1, z_2) - 1]^2 \, dz_1 \, dz_2
\]

However, this comparison is meaningless, since minimising \( \hat{M}(j) \) for “flat” distributions like the uniform distribution requires large values for the bandwidth \( j \), which in turn results in useless kernel density estimates \( \hat{g}_j \) due to \( \hat{g}_j \) taking values equal to one regardless of the information provided by the residuals \( \{Z_i, Z_{i-j}\}_{i>j} \).

Thus, to examine how the choice of bandwidth affects the behaviour of \( \hat{Q}_C \), we calculate their reference finite sample distributions for sample sizes \( n = 800, 6400, 51200 \) and \( 409600 \). The bandwidth in each case is chosen as \( h = \kappa S_n n^{-1/6} \), with scaling factor \( \kappa = \{0.5, 1, 1.5, 2, 2.5, \sqrt{7}, 3\} \). The resultant distributions are displayed in Figure 4.4. As long as the choice of bandwidth is not too extreme, the
4.2. Hypothesis testing

Figure 4.4: Kernel density estimates of \( \hat{Q}_V \) for varying bandwidth scaling factors 0.5, 1, 1.5, 2, 2.5, 2.6458, 3, with the standard normal density function as reference. \( \hat{Q}_V \) was simulated 10000 times each. Data sizes are \( n = 800, 6400, 51200 \) and 409600.

The distribution of \( Q(1) \) should eventually converge to the standard normal distribution. However, a bad choice as in the case of \( h = 0.5S_Zn^{-1/6} \) in Figure 4.4(d), actually leads to a further deviation in distribution. It is also clear that bandwidths \( h = \kappa S_Zn^{-1/6} \) with scaling factor \( \kappa > 1 \), including the optimal bandwidth given as \( h = \sqrt{7}S_Zn^{-1/6} \) by Scott, appears to perform especially poorly for small sample sizes.

Therefore, the choice of bandwidth \( h = S_Zn^{-1/6} \) by Hong and Li appears to be an appropriate one.

4.2.6 Power of the test

Parametric models

Hong and Li [66] investigated the power of the test statistic \( \hat{Q}(j) \) versus time lags \( j \) for the 1- and 3-dimensional cases by comparing parametric models. In the 1-
4.2. Hypothesis testing

dimensional case, the null hypothesis that the data comes from the Vasicek \cite{120} model was tested using data simulated from the Cox-Ingersoll-Ross (CIR), Chan, Karolyi, Longstaff, and Sanders’ \cite{28} (CKLS), Ahn and Gao’s \cite{2} Inverse-Feller, and A"ıt-Sahalia’s \cite{3} non-linear drift models. The test was shown to have good power in rejecting the Vasicek model across all four alternative models. However, this performance was dependent on the data sample sizes considered \((n = 1000, 2500\) and \(5500\)), with near unit power at \(n = 5500\). Additionally, we note that varying \(j\) did not have a discernible effect on the power of \(\hat{Q}(j)\).

For the 3-dimensional case, the null hypothesis that the data comes from the Vasicek \cite{120} model was tested using data simulated from three canonical affine diffusions as used by A"ıt-Sahalia and Kimmel \cite{6}. The power of the test was evaluated based on the generalised residuals of the three individual dimensions and the combined generalised residuals. The test was observed to have good power in detecting model misspecification whenever a dimensional component was misspecified. The power of the test also showed a clear decrease as lag \(j\) increased when the models have been misspecified. They reported a periodic pattern for the power of \(\hat{Q}(j)\), such that when the lag \(j\) was a multiple of the dimension of the process, the power was significantly higher. At such lags, each pair in the combined generalised residuals (due to the way these are constructed, see Section 4.2.3) correspond to the pairs present in the generalised residuals for each component. Thus a misspecification at any dimensional component is correspondingly reflected.

For both the 1- and 3-dimensional power investigations, the first lag at \(j = 1\) seems the most informative, as the power of the test remained stable or decreased as \(j\) increased. Only for the combined generalised residuals, when the lag \(j\) corresponds to a multiple of the dimension of the process, do the powers increase. However, this is simply a reflection of a detection of model misspecification using the individual generalised residuals. Thus in the section to follow, we will focus our own power investigation for the first lag \(j = 1\).
4.2. Hypothesis testing

Non-parametric models

Here we investigate the power performance of the test statistic $\hat{Q}(1)$ for four 2-dimensional diffusion process models. In each of the cases we simulate trajectories from the model described by the alternative hypotheses and test for their respective null hypotheses.

We once again choose to use the 2-dimensional toy diffusion process described in Section 2.1 and denote by $b_0$ and $\sigma_0$ its drift and diffusivity coefficients. For the first three hypotheses, the data is simulated from this toy diffusion process. We then denote by $\hat{b}$, $\hat{\sigma}$ their posterior means obtained upon performing the composite Bayesian inference methodology on the data as described in Section 3.7, and $\gamma$ a multiplicative scaling constant taking 0.1 increments in $[0, 1]$.

We then consider the null hypotheses that the data arises from the same toy diffusion process but with multiplicatively scaled coefficients. These nulls hypotheses are respectively (in accordance with (4.10) but with the dependence of the transition densities of the model on the drift and diffusivity coefficients made explicit):

1. $H_0^0: p(x,t|y,s; \gamma b_0, \sigma_0) = p(x,t|y,s; b_0, \sigma_0)$ a.e.,

2. $H_0^1: p(x,t|y,s; \gamma b, \sigma_0) = p(x,t|y,s; b, \sigma_0)$ a.e. for some $(b, \sigma_0) \in \Psi_0$, and

3. $H_0^2: p(x,t|y,s; \gamma b, \sigma) = p(x,t|y,s; b_0, \sigma_0)$ a.e. for some $(b, \sigma) \in \Psi_0$,

wherein $(\gamma b_0, \sigma_0)$, $(\gamma \hat{b}, \sigma_0)$ and $(\gamma \hat{b}, \hat{\sigma})$ are used as plug-in parameters to evaluate the plug-in $p$-value $\hat{p}_{plug}$.

For the final hypothesis, the data is also simulated from the toy diffusion model but with a spatially dependent diffusivity coefficient $\sigma_0(\zeta + (1 - \zeta) \frac{b_{0,x}}{|b_{0,x}|})$ where $b_{0,x}$ is the first-dimensional component of $b_0$ and $\zeta$ a constant taking 0.1 increments on $[0, 1]$. Here we test the null hypothesis that the diffusivity is constant:

4. $H_0^3: p(x,t|y,s; b, \sigma) = p(x,t|y,s;b_0, \sigma_0(\zeta + (1 - \zeta) \frac{b_{0,x}}{|b_{0,x}|}))$ a.e. for some $(b, \sigma) \in \Psi_0$,

where $(\hat{b}, \hat{\sigma})$ is used to evaluate $\hat{p}_{plug}$. 
4.2. Hypothesis testing

Figure 4.5: The empirical power, $P(\text{reject } H_0|H_1 \text{ is true})$, for the null models (a, b, c) with drift and diffusivity $(\gamma b_0, \sigma_0)$, $(\gamma b, \sigma_0)$ and $(\gamma b, \sigma)$ versus the alternative $(b_0, \sigma_0)$, and the null model (d) with drift and diffusivity $(b, \sigma)$ versus the alternative $(b_0, \sigma_0(\zeta + (1 - \zeta)b_0)\sigma)$. The simulated data consists of trajectories with sample size $n = 400$. The three superimposed lines correspond to the first (blue) and second (red) component generalised residuals, and the combined (orange) generalised residuals.

As Figure 4.5 a-c shows, the $\hat{Q}(1)$ test has poor power when testing for scaled versions of the true drift coefficient. However, its power does improve in detecting deviations in the posterior mean of the drift coefficient from that of the true drift. This suggests that the test may be more sensitive to the specification of the direction of the drift, which may not be correctly identified by the posterior mean estimate, rather than its magnitude.

We also note that the power of the test for the first component generalised residuals is significantly lower than the second component and combined generalised residuals. As shown in Section 4.2.3, the construction of the second component generalised residuals relies on the transition density $p(x_{2,i}, t_i | x_{1,i}, x_{i-1}, t_{i-1})$. Due to
4.2. Hypothesis testing

the rotational aspect of the drift $b_0$, the additional information provided by the first component $x_{1,i}$ allows a better prediction of the second component $x_{2,i}$, which in turn manifests as increased power.

Figure 4.5(d) suggests that the $\hat{Q}(1)$ test is slightly more prone to reject models with constant diffusivity when the true diffusivity is in fact spatially dependent, which suggests that a correct specification of the diffusivity is as important as a correct specification of the drift in determining goodness-of-fit of a diffusion process model, if not more so.

In terms of identifying and quantifying model misspecification, this power performance investigation may be useful in measuring how different the null model is from the true model, once the differing areas have been correctly identified. Figure 4.5 serves as a case in point with the magnitude of the drift being of interest.

4.2.7 The portmanteau statistic

Let us consider the following theorem from Hong and Li [66].

**Theorem 6.** [66, p. 72] Set $\hat{Q} \equiv [\hat{Q}(j_1), \ldots, \hat{Q}(j_L)]^\top$, where $j_1, \ldots, j_L$ are $L$ distinct positive integers, and $L$ is a fixed integer. Then, under the same conditions of Theorem 5

$$\hat{Q} \overset{d}{\to} N(0, I)$$

(4.30)

under the null (4.10), where $I$ is an $L \times L$ identity matrix. Consequently, $\hat{Q}(j_{c_1})$ and $\hat{Q}(j_{c_2})$ are asymptotically independent whenever $j_{c_1} \neq j_{c_2}$, $\forall c_1, c_2 \in \{1, \ldots, L\}$.

As noted in Section 4.2.6, the power of the test statistic $\hat{Q}(j)$ for the individual component-wise generalised residuals was shown by Hong and Li to be fairly robust to the choice of lag $j$. However, the power of the test may still potentially be sensitive to the choice of lag $j$ as a result of spatial dependency between the combined generalised residuals. Thus to avoid any further complications that might arise from the choice of lag $j$, one could instead use a portmanteau statistic arising from Theorem 6, which is given as

$$\hat{W}(p) = p^{-1/2} \sum_{j=1}^{p} \hat{Q}(j)$$
which, for any positive integer \( p < n \), converges to \( N(0,1) \) under correct model specification. While the power of \( \hat{W}(p) \) will still depend on \( p \), this dependence should be less noticeable than that of \( \hat{Q}(j) \) on \( j \).

We are not overly concerned with using this portmanteau statistic within our model fitting framework, since the identification of where possible model misspecification can occur is immediately possible at lag \( j = 1 \). However, we have shown that the asymptotic variance and distribution of \( \hat{Q}(j) \) were not achieved in finite-sample experiments in Section 4.2.5, and thus have reason to probe further the finite sample properties of the portmanteau statistic \( \hat{W}(p) \).

As in Section 4.2.5, the generalised residuals are substituted with draws from the \( U[0,1] \) distribution, which are then used to calculate \( \hat{W}(p) \), for data sample size of \( n = 800 \). Figure 4.6 shows the empirical kernel density estimates obtained for 10000 samples of the portmanteau statistic \( \hat{W}(p) \) for the first component and combined generalised residuals at cumulative lag \( p = 10 \). It is plain to see that the empirical densities of \( \hat{W}(10) \) deviates greatly from the desired \( N(0,1) \) distribution.

To understand this deviation from the \( N(0,1) \) distribution, we refer back to Theorem 6 which implies that

\[
\text{cov}[\hat{Q}(i), \hat{Q}(j)] \to 0 \text{ in probability for } i \neq j
\]

as \( n \to \infty \), under correct model specification. We are, however, once again in a
4.2. Hypothesis testing

finite sample setting. Looking at the covariance matrix of $\text{cov}[^\hat{Q}\,(i), \hat{Q}\,(j)]$ across lags $i, j \in \{1, \ldots, 10\}$ for one draw of $\hat{W}(10)$ in Figure 4.7, it is clear that none of the covariance terms are close to zero. The covariance term $\text{cov}[^\hat{Q}\,(1), \hat{Q}\,(2)]$ for lags 1 and 2 is plotted in Figure 4.8 for increasing sample sizes, and it can be seen that even for sample sizes of up to $n = 100 \times 2^{15}$, this covariance term $\text{cov}[^\hat{Q}\,(1), \hat{Q}\,(2)]$ does not go down to zero.

Regardless, as with the test statistic $\hat{Q}_C\,(j)$ in Section 4.2.5, the portmanteau statistic $W(p)$ can still be utilised once a finite-sample correction has been applied.

**Figure 4.7**: The covariance matrix of $\text{cov}[^\hat{Q}\,(i), \hat{Q}\,(j)]$ represented as a 2-dimensional coloured grid with each cell corresponding to a covariance matrix term, across lags $i, j \in \{1, \ldots, 10\}$, where $i \neq j$ for the first component (left) and combined (right) generalised residuals.

**Figure 4.8**: The covariance terms $\text{cov}[^\hat{Q}\,(i), \hat{Q}\,(j)]$, for lags 1 and 2, evaluated against sample size (or equivalently, number of uniform samples used as a substitute for the generalised residuals under correct model specification).
4.3 Bayesian posterior $p$-values

4.3.1 Hypothesis testing: a Bayesian point of view

When considering goodness-of-fit testing from a Bayesian point of view, the statement of the null hypothesis has to be suitably modified to include information provided through the prior and posterior distributions of the relevant parameters.

In our interpretation of the frequentist null hypothesis (4.10) to include the prior densities of the drift and diffusivity coefficients, we state the null hypothesis as follows:

\[ H_0 : \text{The underlying distribution of the data has a transition density which coincides almost everywhere with that of a diffusion process with drift } b \in H_{\text{per}}^s \text{ and diffusivity } \sigma \in \mathbb{R}^{d \times d}, \text{ with measures governed by the prior densities } \pi(b) \text{ and } \pi(\sigma) \text{ (see Sections 3.6.1 and 3.4.1).} \]

(4.31)

The prior densities of the drift and diffusivity make it clear that we will be considering their posterior distributions in formulating the test statistic and therefore the resultant Bayesian posterior $p$-value. We do not concern ourselves with a specific statement of the alternative hypothesis, except to elucidate that it is the complement of the null hypothesis, and to assume the existence of a transition density for the underlying distribution.

We will later on in Section 4.4 evaluate the goodness-of-fit of a diffusion process model with conservative drift. The null hypothesis is simply a re-statement of 4.31, but with the inclusion of the prior density given by $\pi(b|b_\perp = 0)$.

4.3.2 $p$-values: frequentist to Bayesian

As a stepping stone towards a thoroughly Bayesian framework, let us first consider the Bayesian posterior predictive $p$-value as prescribed by Guttman [62] and Rubin [99], which generalises the frequentist $p$-value from Section 4.1.2 and offers an intuitive answer as to how nuisance parameters should be handled. The Bayesian posterior predictive $p$-value is defined as

\[ p_{\text{post}} = P\{T(X) > T(x)|H_0\}, \]

(4.32)
4.3. Bayesian posterior $p$-values

which is taken over

$$f(X|x, H_0) = \int_\psi f(X|\psi) \pi(\psi|x) d\psi,$$

where the dependence on the data $x$ is made explicit and $\pi(\psi|x)$ is the posterior density of $\psi$ conditional on $H_0$ given as

$$\pi(\psi|x) \propto f(x|\psi) \pi(\psi),$$

where $\pi(\psi)$ is the prior density on $\Psi_0$.

Next, Meng [76] extended the posterior predictive $p$-value (4.32) by replacing the test statistic $T(x)$ with a discrepancy variable $D(x, \psi)$, which is a function of both the data and parameters, resulting in the discrepancy $p$-value

$$p_{\text{disc}} = P\{D(X, \psi) > D(x, \psi)|x, H_0\}, \quad (4.33)$$

which is taken over

$$f(X, \psi|x, H_0) = f(X|\psi) \pi(\psi|x).$$

This generalisation of the test statistic to be parameter-dependent provides a direct measure of the discrepancy between the sample and population quantities. As in the case of the plug-in $p$-value, the classical test statistic $T(x)$ can then be viewed as a special case of the discrepancy variable whereby a point-wise, efficient estimate $\hat{\psi}_0$ of the parameters under $H_0$ is used so that $T(x) = D(x, \hat{\psi}_0)$.

We will be using this discrepancy $p$-value $p_{\text{disc}} (4.33)$ in conjunction with the null hypothesis (4.31) as a tool to assess model fit by providing a measure of compatibility between the observed data and a specified model, in the manner suggested by Gelman [57] p.146. For a further discussion of the discrepancy $p$-value (4.33), we refer to Gelman [58], [57] p.150.
4.3. Bayesian posterior \( p \)-values

4.3.3 Estimating the discrepancy \( p \)-value

Now we are faced with the problem of estimating the discrepancy \( p \)-value (4.33) within the context of hypothesis testing for diffusion process models. This can be solved by combining the composite Bayesian inference methodology for diffusion process models outlined in Section 3.7 and the goodness-of-fit testing methodology described in Section 4.2, such that we have as our discrepancy variable the test statistic (4.28), that is \( D(X, \psi) := \hat{Q}_C(1) \).

We begin with the naive approach by re-expressing (4.33) as

\[
p_{\text{disc}} = \mathbb{E} \left[ \mathbb{1}(D(X, \psi) \geq D(x, \psi)) \mid x, H_0 \right],
\]

which leads to the straightforward Monte Carlo estimator

\[
\hat{p}_\text{disc}^{\text{MC}} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(D(X_i, \psi_i) \geq D(x, \psi_i)),
\]

(4.34)

where \( x \) is the observed data, \( X_i \) is a sample from \( f(X \mid x, \psi) \) and \( \psi_i \) a sample from \( \pi(\psi \mid x) \). We have described how to evaluate the discrepancy variable \( D(x, \psi) \) in Section 4.2.2 and are left with the problem of sampling from \( f(X \mid x, \psi) \) and \( \pi(\psi \mid x) \).

We denote \( \psi = (b, \sigma) \) for the drift \( b \) and diffusivity \( \sigma \). Given a trajectory \( X \) the nuisance parameter \( \psi \) can thus be sampled according to Sections 3.4 and 3.6. Given \( \psi \), the trajectory \( X \) can be sampled by an application of the Euler method given as Definition 10.

All together, we can outline this procedure similarly to the Metropolis-within-Gibbs sampler loop constructed in Section 3.7 with the additional iterative step of computing the indicator function \( \mathbb{1}[D(X_i, \psi_i) \geq D(x, \psi_i)] \) and the final step of calculating the discrepancy \( p \)-value estimator \( \hat{p}_\text{disc}^{\text{MC}} \) once the Metropolis-within-Gibbs sampler loop has completed. This is given below as:

**Algorithm 12.** Target: Calculate \( \hat{p}_\text{disc}^{\text{MC}} \) given a trajectory \( x \)

1. Iterate from \( i = 1 \) to \( n \):
   
   (a) Sample \( \psi_i = (b_i, \sigma_i) \) from \( \pi(\psi \mid x) \)
4.3. Bayesian posterior \( p \)-values

(b) Sample \( X_i \) from \( f(X|\psi) \)

(c) Calculate \( 1 \left[ D(X_i, \psi_i) \geq D(x, \psi_i) \right] \)

2. Calculate \( \hat{p}_{\text{disc}}^{\text{MC}} = \frac{1}{n} \sum_{i=1}^{n} 1 \left( D(X_i, \psi_i) \geq D(x, \psi_i) \right) \)

However, the calculation of \( D(X_i, \psi_i) \) in Algorithm[12] is relatively computationally intensive. Thus we desire a computationally cheaper alternative, which can be obtained by exploiting the distribution of \( D(X_i, \psi_i) \). Under the null hypothesis, the asymptotic distribution of \( D(X_i, \psi_i) \) is known as per Theorem[5], which allows us to denote

\[
D(X_i, \psi_i) = G \sim N(0, V),
\]

for a known variance \( V \). We now consider an alternative representation of the discrepancy \( p \)-value

\[
p_{\text{disc}} = \int_{\psi} P_{x^{\text{rep}}}(D(x^{\text{rep}}, \psi) \geq D(x, \psi) | \psi, x, H_0) dP(\psi),
\]

\[
= \int_{\psi} P_G(G \geq D(x, \psi) | \psi, x, H_0) dP(\psi),
\]

\[
= \int_{\psi} [1 - F_G(D(x, \psi))] dP(\psi),
\]

where \( x^{\text{rep}} \) denotes the replication or future observation of \( x \) and \( F_G \) is the c.d.f. of \( G \). This representation leads to a simpler estimator of the form

\[
\hat{p}_{\text{disc}} = \frac{1}{n} \sum_{i=1}^{n} (1 - F_G(D(x, \psi_i))),
\]

whereby the evaluation of \( D(X_i, \psi_i) \) and therefore the sampling of \( X_i \) for every iteration as originally required for the naive estimator \( \hat{p}_{\text{disc}}^{\text{MC}} \) are no longer necessary. Now we need only compare \( D(x, \psi_i) \) against the c.d.f. \( F_G \) of the \( G \sim N(0, V) \) distribution. Thus we have effectively reduced the computation costs of evaluating the estimator \( \hat{p}_{\text{disc}} \) \( n \)-fold when compared to that of \( \hat{p}_{\text{disc}}^{\text{MC}} \).

While both estimators \( \hat{p}_{\text{disc}} \) and \( \hat{p}_{\text{disc}}^{\text{MC}} \) are unbiased, we can show that the variance of \( \hat{p}_{\text{disc}} \) is lower than that of \( \hat{p}_{\text{disc}}^{\text{MC}} \), per the following calculations. The variance
4.3. Bayesian posterior \(p\)-values

The variance of \(\hat{p}_{\text{disc}}\) is

\[
\text{Var}(\hat{p}_{\text{disc}}) = \text{Var}_{\{\psi_i\}} \left( \mathbb{E}_G[\hat{p}_{\text{disc}} | \{\psi_i\}] + \mathbb{E}_{\{\psi_i\}}[\text{Var}_G(\hat{p}_{\text{disc}} | \{\psi_i\})] \right),
\]

\[
= \text{Var}_{\{\psi_i\}} \left( \frac{1}{n} \sum_{i=1}^{n} (1 - F_G(D(x, \psi_i))) \right) + 0,
\]

while the variance of \(\hat{p}_{\text{disc}}^{MC}\) is given by

\[
\text{Var}(\hat{p}_{\text{disc}}^{MC}) = \text{Var}_{\{\psi_i\}} \left( \mathbb{E}_{\{X_i\}}[\hat{p}_{\text{disc}}^{MC} | \{\psi_i\}] + \mathbb{E}_{\{\psi_i\}}[\text{Var}_{\{X_i\}}(\hat{p}_{\text{disc}}^{MC} | \{\psi_i\})] \right),
\]

\[
= \text{Var}_{\{\psi_i\}} \left( \frac{1}{n} \sum_{i=1}^{n} P_X(D(X_i, \psi_i) \geq D(x, \psi_i) | \psi_i) \right),
\]

\[
+ \mathbb{E}_{\{\psi_i\}} \left[ \frac{1}{n^2} \sum_{i=1}^{n} w(\psi_i)(1 - w(\psi_i)) \right],
\]

with

\[
\mathbb{E}_{\{X_i\}}[\hat{p}_{\text{disc}}^{MC} | \{\psi_i\}] = \mathbb{E}_{\{X_i\}} \left[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(D(X_i, \psi_i) \geq D(x, \psi_i)) \right],
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\{X_i\}} [\mathbb{1}(D(X_i, \psi_i) \geq D(x, \psi_i)) | \psi_i],
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} P_X(D(X_i, \psi_i) \geq D(x, \psi_i) | \psi_i),
\]

and

\[
\text{Var}_{\{X_i\}}(\hat{p}_{\text{disc}}^{MC} | \{\psi_i\}) = \text{Var}_{\{X_i\}} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(D(X_i, \psi_i) \geq D(x, \psi_i)) \right),
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}_{\{X_i\}}(\mathbb{1}(D(X_i, \psi_i) \geq D(x, \psi_i)) | \psi_i),
\]

\[
= \frac{1}{n^2} \sum_{i=1}^{n} w(\psi_i)(1 - w(\psi_i)),
\]

where

\[
w(\psi_i) = P_X(D(x_{\text{rep}}, \psi_i) \geq D(x, \psi_i) | \psi_i),
\]

since \(\text{Var}_{x}(\mathbb{1}(D(x_{\text{rep}}, \psi_i) \geq D(x, \psi_i)) | \psi_i)\) can be represented as the variance of a
4.3. **Bayesian posterior \( p \)-values**

Bernoulli distribution with probability of success \( w(\psi_i) \). Since

\[
\mathbb{E}_{(\psi_i)} \left[ \frac{1}{n^2} \sum_{i}^{n} w(\psi_i)(1 - w(\psi_i)) \right] \neq 0,
\]

this shows that \( \text{Var}(\hat{p}_{\text{disc}}) < \text{Var}(\hat{p}^{\text{MC}}_{\text{disc}}) \).

While the estimator \( \hat{p}_{\text{disc}} \) is superior to the naive estimator \( \hat{p}^{\text{MC}}_{\text{disc}} \) in terms of variance and computational cost, our finite-sample experiments in Section [4.2.5](#) have shown that for small sample size \( n \), \( D(X, \psi) \) does not follow a \( N(0, V) \) distribution. We resolve this by obtaining an empirical estimate of its c.d.f., which we denote by \( \hat{F}_{G} \) and use it instead of \( F_{G} \) in (4.35) to estimate \( \hat{p}_{\text{disc}} \). This empirical c.d.f. can be obtained as shown in Section [4.2.5](#) for a relatively low computational cost at an arbitrarily high precision.

Thus to evaluate the estimator \( \hat{p}_{\text{disc}} \) we have the following procedure:

**Algorithm 13.** Target: Calculate \( \hat{p}_{\text{disc}} \) given a trajectory \( x \)

1. Calculate the empirical c.d.f. \( \hat{F}_{G} \)
2. Iterate from \( i = 1 \) to \( n \):
   
   (a) Sample \( \psi_i = (b_i, \sigma_i) \) from \( \pi(\psi|x) \)
   
   (b) Evaluate \( \hat{F}_{G}(D(x, \psi_i)) \)
3. Calculate \( \hat{p}_{\text{disc}} = \frac{1}{n} \sum_{i=1}^{n} (1 - \hat{F}_{G}(D(x, \psi_i))) \)

In the following Section [4.3.4](#) we will investigate the distribution of \( \hat{p}_{\text{disc}} \) for the toy diffusion process model, and in Section [4.4](#) we will evaluate the goodness-of-fit of diffusion process models for the capuchin monkey movement data using \( \hat{p}_{\text{disc}} \).

### 4.3.4 Distribution of the discrepancy \( p \)-value

In frequentist testing frameworks, to assess the compatibility of the null hypothesis with the data, the \( p \)-values in question should ideally be frequentist, that is they follow the uniform distribution, to allow a straightforward interpretation. However
for general cases or for variants on the frequentist $p$-values, this is often unachievable. Unfortunately, discrepancy $p$-values do not necessarily follow the uniform distribution, as shown in two examples by Gelman [56].

While such comparisons are not necessarily required for Bayesian analysis, Rubin [99] argued that:

“frequency calculations that investigate the operating characteristics of Bayesian procedures are (Bayesianly) relevant and justifiable when investigating or recommending procedures for general consumption.”

Gelman et al. [57] additionally advocate treating the discrepancy $p$-value as a tool to quantify departures from the probability model of interest rather than as an indicator of rejection.

With this in mind, we would like to have an understanding of the distribution of the discrepancy $p$-value under our diffusion process inference and testing framework. We can do so by running Algorithm 14 below to collect samples from the distribution of $\hat{p}_{\text{disc}}$.

**Algorithm 14.** Target: The distribution of $\hat{p}_{\text{disc}}$

1. Fix $\psi = (b, \sigma)$

2. Iterate for $j = 1, 2, 3, \ldots$:
   
   (a) Sample $x_j$ from $\mathbb{P}(X|\psi)$

   (b) Iterate from $i = 1$ to $n$:
      
      i. Sample $\psi_i = (b_i, \sigma_i)$ from $\pi(\psi|x_j)$

      ii. Calculate $\hat{F}_G(D(x_j, \psi_i))$

   (c) Calculate $\hat{p}_{\text{disc}, j} = \frac{1}{n} \sum_{i=1}^{n} (1 - \hat{F}_G(D(x_j, \psi_i)))$

**Toy example**

Let us consider again the toy diffusion process from Section 2.1 with diffusivity set to $\sigma = (1.5, 0.4; 0.4, 1.2)$. Each sample path from the toy diffusion process will have
4.3. Bayesian posterior $p$-values

A size of $n = 401$ and a total time of $T = 20$. We fit to each sample path the diffusion process model specified by \((4.1)\). For the Gaussian measure prior of the drift, we set the hyper-parameters of the precision operator to $k = 4$, $\eta = 5 \times 10^{-7}$ and $\epsilon = 5 \times 10^{-3}$. The hyper-parameters of the Inverse Wishart prior for the diffusivity are set to $\nu = 8$ and $\Psi = 6I$.

We run Algorithm 14 for 200 outer iterations, while the inner loop in steps 2.(b)i.-iii. corresponding to the composite Bayesian inference methodology is run for 40 iterations (excluding burn-in) each time.

We include a snapshot of one outer iteration of Algorithm 14 in Figure 4.9 which displays the posterior mean of the drift and the posterior densities of the diffusivity. The discrepancy $p$-values $\hat{p}_{\text{disc}}$ associated with the generalised residuals of the sample path from this iteration are 0.375717 (first component), 0.237132 (second component) and 0.255960 (combined).

The results of Algorithm 14 are shown in Figure 4.10 which displays the empirical distribution of $\hat{p}_{\text{disc}}$ corresponding to the first component, second component and combined generalised residuals using histograms and kernel density estimates. As expected, unlike frequentist $p$-values, the distribution of $\hat{p}_{\text{disc}}$ do not approach the uniform distribution. However, since we are not constrained under the frequentist framework, these empirical distributions of $\hat{p}_{\text{disc}}$ can in fact be quite useful, by
4.3. Bayesian posterior $p$-values

Figure 4.10: The empirical distributions of the discrepancy $p$-values $\hat{p}_{\text{disc}}$ based on the (a) first, (b) second component and (c) combined generalised residuals of the data. The discrepancy $p$-values $\hat{p}_{\text{disc}}$ were sampled 200 times and used to construct the histograms and kernel density estimates here.

treating them as reference distributions when performing goodness-of-fit hypothesis testing for alternative models.

For example, we perform a slightly different experiment by running Algorithm 14 once more, simulating trajectories from the toy diffusion process, but instead modify step 2.(b)i. to sample $\psi_i = (b_i, \sigma_i)$ from $\mathbb{P}(\psi| x_j, b_{\perp} = 0)$ the posterior distribution conditioned on the drift being conservative, as described in Section 3.6.6. Figure 4.11 shows the posterior mean for the conservative drift and the posterior densities for the diffusivity, for one iteration of Algorithm 14. The discrepancy $p$-values $\hat{p}_{\text{disc}}$ associated with the generalised residuals of the sample path from this iteration are 0.000000 (first component), 0.105438 (second component) and 0.045869 (combined).
The results of Algorithm 14 are shown in Figure 4.12. The choice of conservative drift leads to empirical distributions of \( \hat{p}_{\text{disc}} \), which when compared to those in Figure 4.10, are even further from the uniform distribution.

![Figure 4.11: Drift (conditioned to be conservative) and diffusivity inference results after 40 MCMC iterations excluding burn-in from one iteration of Algorithm 14. Left: The posterior mean drift. Right: Posterior densities of the diffusivity. The discrepancy \( p \)-values \( \hat{p}_{\text{disc}} \) for this data are 0.000000 (first component), 0.105438 (second component) and 0.045869 (combined).](image)

The empirical distribution of \( \hat{p}_{\text{disc}} \) for the second component in Figure 4.12(b) is of interest, since it provides \( p \)-values that are more acceptable than that of the first component. To investigate further, we plot the \( x \) and \( y \)-components of the posterior mean drift separately in Figure 4.13. The posterior mean estimates for the unconditioned drift appears to be a decent approximation to the original drift of the toy diffusion process. The posterior mean of the conservative drift in the \( x \)-component in Figure 4.13(e) however bears no resemblance to \( x \)-component of the original drift in Figure 4.13(a), which explains why the empirical distribution of \( \hat{p}_{\text{disc}} \) in Figure 4.12 takes extremely small values. On the other hand, the \( y \)-component in Figure 4.13(f) while being a poor approximation, does bear some resemblance to the original drift in Figure 4.13(b) and to the \( y \)-component posterior mean of the unconditioned drift in Figure 4.13(d). This is in turn reflected in the empirical distribution of \( \hat{p}_{\text{disc}} \) for the second component in Figure 4.12(b) as an improvement over its first component counterpart in 4.12(a).
4.3. Bayesian posterior $p$-values

First component

Second component

Combined

Figure 4.12: The empirical distributions of the discrepancy $p$-values $\hat{p}_{\text{disc}}$ based on the (a) first, (b) second component and (c) combined generalised residuals (for drift conditioned to be conservative) of the data. The discrepancy $p$-values $\hat{p}_{\text{disc}}$ were sampled 200 times and used to construct the histograms and kernel density estimates here.
Figure 4.13: The $x$- (left column) and $y$-components (right column) of (a)-(b) the true drift, (c)-(d) the posterior mean of the unconditioned drift, and (e)-(f) the posterior mean of the drift conditioned to be conservative.
4.3. Bayesian posterior \( p \)-values

4.3.5 Bayesian \( p \)-values and alternatives

For a further overview of Bayesian \( p \)-values including the Bayesian posterior predictive (4.32) and discrepancy (4.33) \( p \)-values, and alternatives such as the conditional predictive \( p \)-value and the partial posterior predictive \( p \)-value proposed by Bayarri and Berger [12] [14], we refer to Robins et al. [96] who studied their asymptotic distribution when used in composite hypothesis models. Robins et al. showed that the conditional predictive and partial posterior predictive \( p \)-values are asymptotically frequentist, which may be a desirable property. The conditional predictive \( p \)-value requires the use of a conditional MLE however, which is unsuitable for our context. The partial posterior predictive \( p \)-value is taken over the distribution

\[
\int_{\psi} f(X|\psi) \pi_{\text{ppost}}(\psi|x) d\psi,
\]

with partial posterior denoted by

\[
\pi_{\text{ppost}}(\Psi|x) \propto f(x|T(x), \psi) \pi(\psi) \propto \frac{f(x|\psi)\pi(\psi)}{f(T(x)|\psi)},
\]

where \( T \) is the test statistic. While this \( p \)-value is asymptotically frequentist and has more power than the other \( p \)-values considered, it does not provide for the use of a discrepancy variable and cannot be modified to allow so without significant difficulty.

Robins et al. additionally proposed a modification of Meng’s discrepancy variable as follows

\[
\bar{D}(X, \psi) := D(X, \psi) - \hat{\nu}_\psi(\psi) i^{-1}_{\psi\psi}(\psi)n^{-1}S_\psi(\psi),
\]

where \( \hat{\nu}_\psi \) is the partial derivative of the asymptotic mean of \( D(X, \psi) \), \( S_\psi(\psi) \) the score for \( \psi \) and \( i^{-1}_{\psi\psi}(\psi) \) the Fisher information. The discrepancy \( p \)-value based on this modified discrepancy statistic \( \bar{D}(X, \psi) \) is then shown to be asymptotically frequentist. However, the estimation of \( \hat{\nu}_\psi \), \( S_\psi(\psi) \) and \( i^{-1}_{\psi\psi}(\psi) \) adds a layer of complexity and may be computationally expensive.
4.3. Bayesian posterior $p$-values

In our context, we do not in any way approach the asymptotic regime which allows for such niceties. As Section 4.2.5 shows, our discrepancy statistic already requires a significant finite-sample correction. The discrepancy $p$-value that we utilise here provides an improvement on the plug-in $p$-value while being of moderate implementational and computational complexity, and also allows for the relaxation away from the asymptotic regime, which is desirable here due to the relatively small sample sizes under consideration in Sections 3.9, 4.4 and 4.5, and the lack of need of a frequentist $p$-value in our model fit setting.

However, the discrepancy $p$-value, as with the alternative Bayesian $p$-values, remains an amalgamation of frequentist and Bayesian concepts. We would ideally continue with an investigation of Bayes factors to coincide with a fully Bayesian framework but defer this for future ventures.
4.4 Application and interpretation

A criticism of model evaluation for animal movement modelling

Notably absent in state-of-the-art approaches to modelling animal movement using diffusion process models are quantifiable measures of model fit, for example see Preisler et al. [70], Parton et al. [85] and Gurarie et al. [61]. If discussion of model fit is present, it is often delegated to reliance on prior knowledge or qualitative analysis of the data at hand due to the lack of classical goodness-of-fit tests, which McClintock et al. [111] acknowledged in an attempt to construct a general modelling framework for animal movement. This behaviour reflects the lack of model evaluation in the general literature for diffusion process models, as previously outlined in Section 4.1.3. In an overview of challenges prevalent in animal movement modelling, Patterson et al. [86] imply that the ability to identify appropriate models and improve upon them as critical in building a cohesive rationale for studying animal movement.

This is not to say that attempts at model evaluation has not been made, as in a recent exception, Breed et al. [1] utilised a modified Kullback-Leibler method, first proposed by Potts et al. [89], where the kernel density estimates of the marginal distributions of data and model are compared to evaluate model fit. While relatively simple to implement, in ignoring the information provided by the transition densities, this method lacks power and identifiability of potential model misfit compared to the goodness-of-fit tests catered specifically to diffusion process models as summarised in Section 4.2.1.

4.4.1 Capuchin monkey movement data

With the null hypothesis (4.31) in mind, let us once again consider the capuchin monkey movement data introduced in Section 3.9. We sub-sample this data, such that there is a minimum of 40 minutes separating any two adjacent time-wise data points, for reasons which will be discussed later in Section 4.5.3. We fit to this sub-sampled data the diffusion process model described by (4.1), with the hyper-parameters of the Gaussian prior precision operator for the drift as $k = 4$, $\eta =$
$5 \times 10^{-7}$ and $\varepsilon = 5 \times 10^{-3}$. The hyper-parameters of the Inverse Wishart prior are set to $\nu = 8$ and $\Psi = 6I$. We then run Algorithm 13 for 100 iterations.

With this, we can then evaluate the test statistic $\hat{Q}_C(1)$ from Section 4.2.2 using the posterior mean of the drift and therefore the plug-in $p$-value $\hat{p}_{\text{plug}}$ associated with $\hat{Q}_C(1)$. These alongside the Bayesian discrepancy $p$-value $\hat{p}_{\text{disc}}$ from Section 4.3.3 are shown below in Table 4.1.

<table>
<thead>
<tr>
<th></th>
<th>First</th>
<th>Second</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{Q}_C(1)$</td>
<td>2.9403</td>
<td>0.6964</td>
<td>2.0130</td>
</tr>
<tr>
<td>$\hat{p}_{\text{plug}}$</td>
<td>0.0022</td>
<td>0.1852</td>
<td>0.0163</td>
</tr>
<tr>
<td>$\hat{p}_{\text{disc}}$</td>
<td>0.0066</td>
<td>0.0614</td>
<td>0.0900</td>
</tr>
</tbody>
</table>

**Table 4.1:** From left to right: the discrepancy statistic (top), plug-in $p$-value (middle) and discrepancy $p$-value (bottom) evaluated for the first, second components and the combined generalised residuals for the capuchin monkey movement data.

When evaluating both the plug-in $p$-value $\hat{p}_{\text{plug}}$ and the Bayesian discrepancy $p$-value $\hat{p}_{\text{disc}}$ under the common choice of significance level at 0.05, the diffusion process model for the capuchin monkey movement data is rejected according to the first component generalised residuals, but not so for the second component and combined generalised residuals, if just barely. With these results, we have reason to doubt that the our diffusion process model is an appropriate fit for the capuchin monkey movement data.

Given our initial assessment of the results from our composite Bayesian inference methodology in Section 3.9, we can be fairly confident that the posterior mean of the drift (used in the evaluation of $\hat{p}_{\text{plug}}$) and the overall posterior distribution of the drift (used to evaluate $\hat{p}_{\text{disc}}$) reflect the true drift of the capuchin monkey movement data. Assuming that this is indeed the case, we can turn instead to other potential sources of model discrepancy for example the choice of constant diffusivity, the choice of sub-sampling, or even the choice of diffusion process model as a whole.
Conservative drift

As a possible scientific motivation in comparing diffusion process models with conservative drift against its non-conservative counterpart, we follow the example of McClintock et al. who studied the movement data of a grey seal. They contend that it is more valuable to extract quantitative information about centres of attraction in the geography as the result of statistical inference rather than as a confirmation of prior knowledge typical in analyses where models are constructed with certain properties in mind. For example Preisler et al.’s and Brillinger’s use of conservative drift fields assume the existence of at least one centre of attraction, which makes sense when foraging and resting areas are taken into consideration. However, other unknown interactions with the environment, for example a cliff edge which guides the animal towards a circular path, cannot be accounted for by the use of a conservative drift field.

Let us now replicate the set up described in the beginning of Section but with the null hypothesis modified so that the drift is conditioned to be conservative, that is \( b = b_\parallel \). Algorithm then run once more for the capuchin monkey movement data. The results are displayed in Table 4.2 below.

<table>
<thead>
<tr>
<th></th>
<th>First</th>
<th>Second</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Q_C(1) )</td>
<td>3.9368</td>
<td>4.9486</td>
<td>4.0376</td>
</tr>
<tr>
<td>( \hat{p}_{\text{plug}} )</td>
<td>0.0020</td>
<td>0.0000</td>
<td>0.0010</td>
</tr>
<tr>
<td>( \hat{p}_{\text{disc}} )</td>
<td>0.0023</td>
<td>0.0001</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

Table 4.2: From left to right: the discrepancy statistic (top), plug-in \( p \)-value (middle) and discrepancy \( p \)-value (bottom) evaluated for the first, second components and the combined generalised residuals for the capuchin monkey movement data fitted with a diffusion process model with conservative drift.

With a significance level of 0.05, we find ourselves rejecting the diffusion process model with conservative drift for both \( \hat{p}_{\text{plug}} \) and \( \hat{p}_{\text{disc}} \) across all types of generalised residuals. In comparison with Table the \( \hat{p}_{\text{plug}} \) and \( \hat{p}_{\text{disc}} \) values for the second component and combined generalised residuals of the conservative drift model are orders of magnitudes smaller than that of its counterpart with unconditioned drift. The \( \hat{p}_{\text{plug}} \) and \( \hat{p}_{\text{disc}} \) values for the first component generalised residuals...
of both models are fairly similar however.

![Figure 4.14: The posterior mean drifts of the model with unconditioned drift (left column) and the model with drift conditioned to be conservative (right column), given as (a)-(b) the full posterior mean, and (c)-(d) the x- and (e)-(f) y-components of the full posterior mean](image)

To investigate this phenomenon, we display the posterior mean of the drifts and their x- and y-components for both conservative and unconditioned drift models in Figure 4.14. The close values of $\hat{\rho}_{\text{plug}}$ and $\hat{\rho}_{\text{disc}}$ for the first component generalised
residuals for both models may be explained by the similarities in the $x$-component of the posterior mean drifts in terms of both scale and shape as shown by Figures 4.14(c)-(d). Interestingly, the $y$-component of the drift in both Figures 4.14(e)-(f) display a similar shape, but the scales are significantly different. This discrepancy between the posterior mean drifts of the two models may suggest why the fit for the model with conservative drift is significantly worse.

A brief conclusion

Given the results in Tables 4.1 and 4.2, a strict interpretation via the plug-in $p$-value $\hat{p}_{\text{plug}}$ should lead to a wholesale rejection of the model. On the other hand, we can conclude that while our diffusion process model with unconditioned drift is not exactly a good fit for the capuchin monkey movement data, we do not reject it on the basis of the discrepancy $p$-values $\hat{p}_{\text{disc}}$ for the second component and combined generalised residuals.

We additionally note that with the exception of the second-component generalised residuals for the unconditioned drift model, the discrepancy $p$-value $\hat{p}_{\text{disc}}$ gives higher values than that of the plug-in $p$-value $\hat{p}_{\text{plug}}$. This is unsurprising since the $\hat{p}_{\text{disc}}$ is allowed to explore the posterior of the drift and diffusivity as opposed to the single estimate used to calculate $\hat{p}_{\text{plug}}$.

As for the model with conservative drift, we have a reason other than the extremely low $p$-values which lead us to believe that this model is a poor fit for the capuchin monkey movement data. In particular, the posterior mean of the conservative drift is unable to capture the rotational aspects of the data.

Therefore, it must be emphasized again that the rejection of a probability model should not be the be-all and end-all of $p$-value interpretation. As Gelman et al. [57] put it:

The relevant goal is not to answer the question, ’Do the data come from the assumed model?’ (to which the answer is almost always no), but to quantify the discrepancies between data and model, and assess whether they could have arisen by chance, under the model’s own assumptions.
What is of concern though is the magnitude of the discrepancy variable, or rather, the magnitude of the Bayesian discrepancy $p$-values: just how poor is our model fit? Identification of sources of model discrepancy however can be difficult from looking at the discrepancy variable or the Bayesian discrepancy $p$-value alone. With this in mind, we point forward to Section 4.5, where we will attempt to improve model fit by identifying potential areas where the discrepancies between data and the diffusion process model present themselves and address these discrepancies. In turn, improvements in model fit would be reflected in an increase of the discrepancy $p$-values.
4.5 Improving model fit

Introduction

The majority of this chapter has been focussed on formal goodness-of-fit hypothesis testing for our choice of diffusion process model, which we fit to a set of capuchin monkey movement data. However, it can be difficult to determine whether a particular model should be used in the first place, so one might first consider some form of informal model validation. These informal model checks can also be performed after fitting the model, especially in the absence of formal goodness-of-fit tests.

A popular choice would be to perform graphical model checking as suggested by Gelman et al. [57] in the form of a display of the observed data, a display of data summaries or parameter estimates, graphs of residuals or other measures of discrepancy between the fitted model and observed data. These initial checks alongside the application of goodness-of-fit hypothesis testing not only allows a quantification of model discrepancy, but also possible identification of sources of said model discrepancy, which once rectified may lead to potential alternative models. Another common check would be to examine the data for potential outliers. This can be performed graphically, or by comparing the data with prior information.

In any case, these model checks when taken together can help in improving a fitted model. In Section 4.5.1 we will consider graphical checks by examining the quadratic variation of the capuchin monkey movement data, and by comparing naive estimates of the drift and diffusivity of the data under the assumption that it comes from a diffusion process model. We then simulate data using the posterior means of the drift and diffusivity and compare them against the actual movement data. In Section 4.5.2 we will construct an outlier detection method using the generalised residuals as described in Section 4.2.2. In Section 4.5.3 we will attempt to identify the correct time scales in which the data behaves like a diffusion process model.
4.5. Improving model fit

4.5.1 Informal model validation

Quadratic variation

Prior to fitting diffusion process models, one can investigate the quadratic variation estimate of the data in question. The quadratic variation of a stochastic process $X$ with time index on $t = [0, T]$ is defined as

$$[X, X]_T := \mathbb{E} - \lim_{\|\Pi_n\| \to 0} \sum_{i=1}^{n} (X_{t_i} - X_{t_{i-1}})(X_{t_i} - X_{t_{i-1}})^T,$$

(4.36)

with partition $\Pi_n := \{t_0 = 0 < \ldots < t_n = T\}$. For a diffusion process model to be considered a possible fit to the observed data, we must expect the data to have regularity comparable to that of a diffusion process, that is its quadratic variation should not equal to zero. The quadratic variation should also stabilise to its limit as the partition $\Pi_n$ tends to zero.

![Figure 4.15: Empirical quadratic variations of the capuchin monkey movement data when sub-sampled along the time index.](image)

In Figure 4.15 we plot the quadratic variation estimates for the capuchin monkey movement data in the presence of zero drift, at different levels of sub-sampling on the time index. In this sub-sampling scheme, data points are removed from the trajectory so that a minimum of $\Delta t$ minutes is achieved for any two time-wise adjacent points. We note that the capuchin monkey movement data does not make for a good example here, since the minimum $\Delta t$ between any two points is 8 minutes, which disallows a direct comparison with (4.36) as to whether its empirical quadratic variation converges or not. However, what we do observe is that its empirical
4.5. Improving model fit

quadratic variation is observed to stabilise for higher levels of sub-sampling, especially for $\Delta t > 35$. which tentatively suggests the plausibility of fitting diffusion process models. In fact, we will show in Section 4.5.3 that at higher levels of sub-sampling, a better model fit can be achieved for the capuchin monkey movement data.

**Graphical comparisons: parameter estimates**

In a recent paper, García-Portugués et al. [55] evaluated the goodness-of-fit of their parametric diffusion process model by graphically comparing their drift and diffusivity estimates with a non-parametric counterpart. For the two-dimensional case, they constructed non-parametric estimates using product kernels and show that the parametric estimates of their models do indeed match their non-parametric counterparts for regions where data is present.

We do not exhibit a similar confidence in using graphical comparisons of the drift estimate of our diffusion process model as a check for goodness-of-fit, since we already estimate our drift non-parametrically.

Graphical comparisons for the diffusivity coefficient on the other hand should be more illuminating since diffusion processes are sensitive to small changes to their diffusivity coefficients, as evidenced by the availability or lack thereof of the Radon-Nikodym derivative. So if the true diffusivity of the data is spatially dependent, our model where the diffusivity is assumed to be constant will fit poorly. In that case, the diffusivity should be estimated using non-constant parametric or even non-parametric estimators. This however leads to a breakdown in our composite Bayesian inference methodology, since the lack of a general multi-dimensional Lamperti transform complicates the construction of a diffusion bridge sampler and conjugate prior for the drift.

To check whether it is plausible for the data to have constant diffusivity or not, we require only a simple estimator. We construct here a very naive piecewise constant estimator, using a similar discretisation set up as in Section 3.6.4 such that
its value when evaluated on the \((j,k)\)-th cell of a grid partition is given by

\[
\hat{\sigma}(j,k) = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{j,k}(X_{t_i})} \sum_{i=1}^{n} \mathbb{1}_{j,k}(X_{t_i}) \frac{|X_{t_i} - X_{t_{i-1}}|^2}{t_i - t_{i-1}},
\]  
(4.37)

with indicator function

\[
\mathbb{1}_{j,k}(x) = \begin{cases} 
1, & \text{if } x \text{ is in the } (j,k)\text{-th partition of the grid.} \\
0, & \text{otherwise.} 
\end{cases}
\]

Figure [4.16] illustrates the estimated \(\hat{\sigma}_{1,1}, \hat{\sigma}_{1,2}\) and \(\hat{\sigma}_{2,2}\) terms of \(\hat{\sigma}\) for the capuchin monkey movement data. Apart from a few spikes the diffusivity terms \(\hat{\sigma}_{1,1}\) and \(\hat{\sigma}_{1,2}\) remain relatively constant, whereas the \(\hat{\sigma}_{2,2}\) term exhibits a centred cone-like shape. This suggests that the choice of constant diffusivity for our diffusion process model may be not be entirely appropriate.

Due to the difficulty in constructing a suitable parametric drift estimator for the capuchin monkey movement data, and due to the non-parametric nature of our posterior mean estimator of the drift, we consider instead a very naive non-parametric estimator, in the same vein as the above naive diffusivity estimator (4.37). This is given by \(\hat{b}\), whose value when evaluated on the \((j,k)\)-th partition of the grid is given by

\[
\hat{b}(j,k) = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{j,k}(X_{t_i})} \sum_{i=1}^{n} \mathbb{1}_{j,k}(X_{t_i}) \frac{X_{t_i} - X_{t_{i-1}}}{t_i - t_{i-1}}.
\]

This estimator assumes unit diffusivity, which relies on the previously established notion that the data may have constant diffusivity.

In Figure [4.17] we compare the naive piece-wise constant estimate for the drift of the capuchin monkey movement data with a non-parametric drift estimate as obtained using the drift inference methodology in Section [3.6] At first glance, this naive estimate enforces the idea that a parametric estimator would be inadequate. Barring the stronger boundary forces present in the non-parametric estimate, both estimates bear similarities in areas with strong directing forces. However, the regularity enforced by the precision operator of the Gaussian measure drift prior
4.5. Improving model fit

Figure 4.16: The piece-wise constant estimates of the $\hat{\sigma}_{1,1}$, $\hat{\sigma}_{1,2}$ and $\hat{\sigma}_{2,2}$ terms (a,b,c) of the diffusivity coefficient for the capuchin monkey movement data.

Figure 4.17: A non-parametric estimate (left) and naive piece-wise constant estimate (right) of the drift coefficient for the capuchin monkey movement data.

results in a loss of detail within the interior as opposed to that of the naive drift estimate. In attempting to improve model fit, we could re-evaluate our choice of hyper-parameters for the precision operator, to allow for a less regular estimate of the drift, for example.
Graphical comparisons: simulated trajectory

As a posterior predictive check, we compare the observed capuchin monkey movement data with a simulated trajectory from our fitted diffusion process model using the posterior means for the drift and diffusivity coefficients as obtained in Section 3.9. Both the original data and the simulated path are displayed in Figure 4.18. The range of the data mostly agree in both cases, since the simulated data is constrained by the strong boundary forces present in the posterior mean estimate of the drift, see for example Figure 3.23 (left). There are no obvious outliers but we note that while the areas of concentration observed in the movement data are also present in the simulated data, the simulated data are more evenly spread out. This suggests a failure of the drift estimate to account for these attraction points or perhaps the existence of an underlying diffusivity that is spatially dependent.

![Figure 4.18: Trajectories of the observed capuchin monkey movement data (left), and simulated data (right) from our diffusion process model using the posterior means for its drift and diffusivity.](image)

A brief conclusion

We have made some fairly general remarks concerning our choice of diffusion process model for the capuchin monkey data, having examined the quadratic variation of the data, compared the estimates of the drift and diffusivity, and compared the observed data against simulated data. Taken individually, these informal checks enable the identification of some discrepancies between the data and our diffusion process model, but cannot provide a definitive answer as to the goodness-of-fit of
the model.

However, when these model checks are considered together, we should begin to suspect that our model might not be a good fit after all. This is further evidenced by the evaluation of our diffusion process model as fitted to the capuchin monkey movement data in Section 4.4 using both the plug-in and discrepancy $p$-values, where there is in fact reason to consider the model to be a poor fit.

### 4.5.2 Outlier detection

“How to bring order and yet to permit diversity is a problem for all men and not least for the statistician.”

– Box and Tiao, *A Bayesian approach to some outlier problems.*

We use the term “outlier” loosely to refer to an observation that is suspected to deviate from the generation mechanism of a fitted stochastic model, see for example Box [21] and Anscombe [8]. Thus as an initial attempt at improving model fit, one might analyse the data for such outliers and consider removing them if it makes sense to do so.

We can construct an outlier detection method along the lines of a conventional outlier rejection rule. As described by Anscombe [8] Rule 0, p. 129], an observation $x_i$, which has generalised residual $Z_i$ with standard deviation $s_Z$, is considered an outlier if

$$|Z_i| > C s_Z,$$

for some constant $C > 0$.

For lag $j = 1$, the pairwise generalised residuals $\{Z_{i-1}, Z_i\}_{i=2}^n$ for the trajectory can be obtained as in Section 4.2.2. Under the null hypothesis these generalised residuals follow a 2-dimensional uniform distribution. However, the uniform distribution is platykurtic and “flat-topped” with bounded support, which disallows the use of the outlier detection rule (4.38) as it does not produce outliers in the sense defined here. To counteract this, an inverse probability transformation is applied as follows

$$Z'_{i-1} = \sqrt{2}\text{erf}^{-1}(2Z_{i-1} - 1),$$

(4.39)
where \( \text{erf}^{-1} \) is the inverse error function. These transformed generalised residuals, denoted as \( \{Z_{i-1}', Z_i'\}_{i=1}^n \), now follow a 2-dimensional standard normal distribution.

Applying the outlier detection rule (4.39) for \( C = 3 \) and \( s_Z = 1 \) here corresponds to considering generalised residuals lying three standard deviations away from the zero mean as outliers. Additionally, according to (4.14) each pair of generalised residuals \( (Z_{i-1}, Z_i) \) is computed using the data points \( X_{i-2}, X_{i-1} \) and \( X_i \), so all three can in turn be considered outliers. However, it is the middle data point \( X_{i-1} \) which is most suspect, since if the surrounding points \( X_{i-2} \) and \( X_i \) are indeed outliers, the adjacent generalised residuals \( (Z_{i-2}, Z_{i-1}) \) and \( (Z_i, Z_{i+1}) \) are likely to be flagged as outliers as well.

**Toy example**

To illustrate this outlier detection method, we simulate a trajectory from the toy diffusion process as described in Section 2.1. A single point from this trajectory is chosen at random and incremented in both dimensions arbitrarily by 0.2. Our diffusion process model as described by (4.1) is fitted to the trajectory and the composite Bayesian inference methodology as outlined in Section 3.7 is performed. The outlier detector method above is then applied, where the generalised residuals are computed using the posterior mean estimates of the drift and diffusivity.

This results in the identification of an outlier, which can be seen lying without the red perforated ellipse in Figure 4.19 (left). The three corresponding suspect data points are circled accordingly in Figure 4.19 (right). The data point circled in red (with adjacent points circled in magenta) is exactly the point that was arbitrarily incremented earlier.

Once outliers have been detected, one should additionally determine why said observations are suspect. Anscombe [8] described three sources of variability in the observations as inherent variability, measurement error and execution error. The outlier found here is simply the result of artificial incrementation, and therefore may be classified as execution error.
4.5. Improving model fit

Figure 4.19: Left: the transformed generalised residuals \( \{ Z_i^{N+1}, Z_i^{N} \}_{i=2}^n \) for the \( x \)-component of the simulated trajectory, with a red eclipse identifying any points lying more than three standard deviations away from the mean. Right: simulated trajectory from the toy diffusion process where the incremented point has been identified and circled in red, with adjacent suspect points circled in magenta.

Capuchin monkey

We now apply the same outlier detection method for the capuchin monkey movement data. 33 trajectories are found to contain outliers in Figure 4.20 (left), and we have chosen to display one with a seemingly obvious outlier in Figure 4.20 (right).

Figure 4.20: Left: capuchin monkey movement trajectories which contain outliers. Right: a closer look at one of the trajectories.

If the highlighted observation in Figure 4.20 (right) is indeed correct and not an outlier, this corresponds to the capuchin monkey having travelled a total of approximately 250 metres over a period of 19 minutes. This is not unreasonable as Janson and Bitetti [69] found the average movement speed of a group of capuchin monkeys during rapid travel to be 16.03 m/min, whereas Parr et al. [84] recorded
sustained travel speeds equivalent to 4.50 m/min over half-hour intervals.

What makes this data point suspect however, is the sudden and sharp deviation from the trajectory and back again that is inconsistent with the rest of the trajectory. If the adjacent points are assumed to be correct (or rather, more correct than the suspected outlier), then the point in question could be the result of a failure of the GPS recording device and therefore a measurement error.

After detecting an outlier and having sufficient reasons to remove it, as is the case here, one would hope to observe a subsequent improvement in model fit. Unfortunately for the capuchin monkey movement data, this does not appear to be the case. We detect no change in the $\hat{Q}_C(1)$, $\hat{p}_{\text{plug}}$ and $\hat{p}_{\text{disc}}$ values for the capuchin monkey movement data whether the outliers are removed or not. It might be the case that the posterior distributions arising from our choice of model is outlier-prone as defined by O’Hagan [79], in which case we can rely on the Bayesian inference procedure to automatically account for and weigh all observations including any outliers accordingly.

In conclusion, we have provided an outlier detection method which should help to improve model fit, if the outliers as defined above are indeed responsible for discrepancies between the fitted model and the data. For the capuchin monkey movement data however, we suspect it may be other sources of discrepancies rather than outliers with suspect residuals that are at fault here, which leads us to our next step: sub-sampling of data.

### 4.5.3 Multiscale data and subsampling

We now turn to investigating the possible existence of inliers within the data and a method to overcome this potential source of model misspecification. According to Winkler [123], an “inlier” refers to an observation that is suspected to be in error of a fitted stochastic model while lying within the interior of its distribution, as opposed to an outlier which lies in its tails.

In particular, we have shown in Section 4.4 that according to the plug-in and Bayesian discrepancy $p$-values, our diffusion process model is a poor fit. In striving for an improvement of model fit, we contend that there may be inliers present in
4.5. Improving model fit

the data to be investigated and removed as an alternative to outright rejection of the model. Therefore we consider the possibility of the capuchin monkey movement data of having multiscale properties, wherein the data exhibits different properties at different timescales, and should fitted with multiple models simultaneously at these timescales. These multiscale properties would then be reflected as systematic inliers which cannot be explained under our diffusion process model.

Homogenisation

Pavliotis et al. [87] and Papavasiliou et al. [83] analysed issues arising from working with multiscale data and showed how appropriate levels of sub-sampling allowed the use of diffusion process models. Components of certain scale-separated models with fast/slow systems of known form have been shown to converge weakly to stochastic differential equations under certain scaling limits, while exhibiting different properties at other scales. For example, a scale-separated system of ordinary differential equations can give rise to a stochastic differential equation [87, p. 3] and a scale-separated system of stochastic differential equations can give rise to a homogenised stochastic differential equation [87, p. 4]. However, fitting a diffusion process model at the wrong time scale can result in incorrect inferences for its parameters. In particular, we are interested in the homogenisation situation, whereby observing data at small time-scales results in incorrect inference.

Papavasiliou et al. [83] considered a coupled system of stochastic differential equations for \((x,y)\) in \(\mathcal{X} \times \mathcal{Y}\). We simplify their set up by using the 1-dimensional unit torus \(\mathcal{X} = \mathcal{Y} = \tau\). This coupled system is given by

\[
\begin{align*}
\frac{dx}{dt} &= \frac{1}{\varepsilon} f_0(x,y) + f_1(x,y) + \alpha_0(x,y) \frac{dU}{dt} + \alpha_1(x,y) \frac{dV}{dt} \\
\frac{dy}{dt} &= \frac{1}{\varepsilon} g_0(x,y) + \frac{1}{\varepsilon} g_1(x,y) + \frac{1}{\varepsilon} \beta(x,y) \frac{dV}{dt}
\end{align*}
\] (4.40a)

which projects onto the x-coordinate, for some \(f_0, f_1, \alpha_0, \alpha_1, g_0, g_1, \beta : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}, \ t \in [0,T] \) and \(U,V\) the standard 1-dimensional Wiener process. As \(\varepsilon \rightarrow \infty\), this projection is then shown to converge weakly to a diffusion process.
4.5. Improving model fit

Theorem 7. [83, p. 3181] Let Assumptions 2.1, 2.4 [83, p. 3177, 3181] hold. Then \( x \Rightarrow X \) (in the sense of weak convergence of probability measures) in \( C([0,T], \mathcal{X}) \) as \( \varepsilon \to 0 \) where \( X \) solves the SDE

\[
\frac{dX}{dt} = F(X; \theta) + A(X) \frac{dW}{dt} \tag{4.41}
\]

with \( W \) a standard \( l \)-dimensional Wiener process. The \( F \) and \( A \) terms are obtained by integrating over \( f_0, f_1, \alpha_0, \alpha_1, g_1 \) and \( \beta \) as shown in Assumptions 2.4 [83, p. 3177, 3181].

If one were to fit a diffusion model described by (4.41) onto the projected trajectory observed at small time-steps and perform inference for the drift parameter \( \theta \in \Theta \) using a maximum likelihood estimator (MLE) however, the estimates would not converge to the true parameter. This is due to the likelihood of the projection arising from (4.40) not converging to the likelihood of the homogenised stochastic differential equation (4.41). This is given below by Theorem 8.

Theorem 8. [83, p. 3191] Let Assumptions 2.1, 2.4, 3.1, 3.5, 3.7 and (3.8) [83, p. 3177, 3181, 3183, 3185, 3186 and 3185] hold. Let \( \{x(t)\}_{t \in [0,T]} \) be a sample path of (4.40) and \( \{X(t)\}_{t \in [0,T]} \) a sample path of (4.41) at \( \theta = \theta_0 \). Then the following limits, to be interpreted in \( L^2(\Omega) \) and \( L^2(\Omega_0) \) respectively (where \( \Omega \) is the probability space for \((U,V)\) and \( \Omega_0 \) the probability space for \( W \)), and almost surely with respect to \( x(0), y(0), X(0) \), are identical:

\[
\lim_{\varepsilon \to 0} \lim_{T \to \infty} \frac{1}{T} \mathbb{L}(\theta; x) = \lim_{T \to \infty} \frac{1}{T} \mathbb{L}(\theta; X) + E_\infty(\theta),
\]

where \( E_\infty(\theta) \) is the asymptotic error in the likelihood [83, p. 3190].

Theorem 8 shows that the correct limit of the likelihood is not obtained unless \( E_\infty \) is constant in \( \theta \), which will not be true in general. A MLE \( \hat{\theta} \) for the drift parameter will be asymptotically biased, and will in effect maximise for the incorrect likelihood.

The sample path \( \{x(t)\}_{t \in [0,T]} \) can be sub-sampled by selecting \( \{x(n\delta)\}_{n=0}^{N-1} \) for
some $\delta$. It can be shown, see [83, p. 3193], that by appropriately sub-sampling, or by choosing $\delta$ at an appropriate $\epsilon$-dependent rate, the $E_{\infty}$ term disappears. Theorem 9 then shows that the MLE for the drift parameter converges correctly to the true parameter.

**Theorem 9.** [83, p. 3196] Let Assumptions 2.1, 2.4, 3.1, 3.5, 3.7 and 4.4 [83, p. 3177, 3181, 3183, 3185, 3186 and 3195] hold and assume that $\theta \in \Theta$, a compact set. Let $\{x(t)\}_{t \in [0,T]}$ be a sample path of (4.40) at $\theta = \theta_0$. Define the maximum likelihood estimator

$$\hat{\theta}(x; \epsilon) := \arg\max_{\theta} \tilde{Z}_{N,\delta}(\theta; x)$$

with $\delta = \epsilon^\alpha$ with $\alpha \in (0,1)$, $N = \lfloor \epsilon^{-\gamma} \rfloor$ with $\gamma > \alpha$ and $\tilde{Z}_{N,\delta}(\theta; x)$ defined as a discretised form of the log-likelihood function [83, p. 3195] for $\theta$ of $X$ (4.41). Then,

$$\lim_{\epsilon \to 0} \hat{\theta}(x; \epsilon) = \theta_0, \quad \text{in probability.}$$

The above Theorems 7, 8 and 9 assumed a diffusion process model with parametric drift, and inference performed via a maximum likelihood estimator. We suspect that for our composite Bayesian inference scenario, a similar phenomena occurs, whereby our non-parametric inference for the drift will perform no better or even worse than its parametric counterpart if multi-scale properties exist within the data and no sub-sampling is performed.

**Sub-sampling**

We perform systematic sub-sampling on the capuchin monkey movement data $\{x_{t_i}\}_{i=0}^N$ whereby data points are removed until any two adjacent points in the subsampled trajectory $\{x_{t_i}\}_{i=0}^N$ have a minimum of $\Delta t$ minutes between them, such that $t_{i+1}^s - t_i^s > \Delta t$.

Table 4.3 displays the amount and percentage of capuchin monkey movement data retained after systematic sub-sampling has been performed. It appears at first
4.5. Improving model fit

<table>
<thead>
<tr>
<th>Sub-sampling min($\Delta t$)</th>
<th># of points after sub-sampling</th>
<th>% of original data points</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3033</td>
<td>100.00</td>
</tr>
<tr>
<td>9</td>
<td>2473</td>
<td>81.53</td>
</tr>
<tr>
<td>10</td>
<td>1739</td>
<td>57.33</td>
</tr>
<tr>
<td>16</td>
<td>1436</td>
<td>47.34</td>
</tr>
<tr>
<td>20</td>
<td>951</td>
<td>31.35</td>
</tr>
<tr>
<td>30</td>
<td>715</td>
<td>23.57</td>
</tr>
<tr>
<td>40</td>
<td>508</td>
<td>16.75</td>
</tr>
</tbody>
</table>

Table 4.3: Number and percentage of data points remaining after systematic time-wise sub-sampling of the capuchin monkey movement data.

glance that after sub-sampling, a significant amount of data, and therefore information about the drift and diffusivity, has been discarded. When taking into account the data remaining after sub-sampling however, as displayed in Figure 4.21 (left column), the observed spatial dependency manifesting as a rotation anti-clockwise combined with the concentration points of the data, indicate little if no loss of information required for inference.

Figure 4.21 (right column) also displays histograms of the remaining time differences for the sub-sampled capuchin monkey movement data.

The test statistics $\hat{Q}_C(1)$, plug-in $p$-values $\hat{p}_{\text{plug}}$ and discrepancy $p$-values $\hat{p}_{\text{disc}}$ for the sub-sampled capuchin monkey movement data are displayed in Figure 4.22 for both the unconditioned drift and conservative drift diffusion process models.

Let us first consider only the unconditioned drift model. We observe that as the level of sub-sampling increases, the model fit improves. Interestingly, there appears to be a reversal as to which generalised residual components indicate better model fit. For minimum time differences of $\Delta t = 9, 10, 16$ and 20 minutes, the $\hat{Q}_C(1)$ values associated with the first component generalised residuals indicate better model fit when compared to the second component and combined generalised residuals. This situation is reversed for minimum time differences of $\Delta t = 30, 40$ and 50 minutes. These are also the only time differences where we would not reject the model when compared against a frequentist 5% significance level, for both $\hat{p}_{\text{plug}}$ and $\hat{p}_{\text{disc}}$.

A possible explanation is that the capuchin monkey movement data does in-
4.5. Improving model fit

deed exhibit multi-scale properties, whereby a diffusion process model may be an appropriate fit, but only when the data is sub-sampled such that there is a minimum time difference of 30 minutes between any two observed points. In particular, the model is able to more correctly identify the true $y$-component of the drift at this level of sub-sampling.

As for the model with drift conditioned to be conservative, we note that while an improvement in model fit can be observed in terms of $\hat{Q}(1)$, this improvement is not reflected in the $\hat{p}_{plug}$ and $\hat{p}_{disc}$ values, regardless of level of sub-sampling. Given the rotational aspect still present in the data, which cannot be captured by the conservative drift, we have reason enough to believe that the conservative drift model is in fact not an appropriate model for the capuchin monkey data.
Figure 4.21: Left column: movement data of the capuchin monkey with colour-coded data points to represent their time-indexing, with the start-points of the trajectory in yellow and the end-points in blue. Right column: histogram of minimum time differences $\Delta t$ after sub-sampling. These plots are obtained for the capuchin monkey movement data after sub-sampling so that there is a minimum time difference of (a) 10 minutes, (b) 20 minutes and (c) 30 minutes between data points.
4.5. Improving model fit

Figure 4.22: The (a) $\hat{Q}_C(1)$ values, (b) plug-in $p$-values $\hat{p}_{\text{plug}}$ and (c) Bayesian discrepancy $p$-values $\hat{p}_{\text{disc}}$ for the capuchin monkey movement data at various levels of sub-sampling, for the diffusion process models with unconditioned drift and drift conditioned to be conservative.
4.5. Improving model fit

Some asides

To determine if systematic sub-sampling at an appropriate rate results in good estimators as described by Theorems 8 and 9 as opposed to random sub-sampling, we randomly sub-sample without replacement the equivalent number of data points from the capuchin monkey movement data. As can be shown in Figure 4.23, systematic sub-sampling does provide better $\hat{Q}_C(1)$ values when compared to random sub-sampling. This improvement is not as obvious for the case where $t_{i+1}^s - t_i^s > 9$ minutes, but since the model fit is markedly poor regardless at this level of sub-sampling, we are not overly concerned with this result.

![Figure 4.23: The $\hat{Q}_C(1)$ values for the capuchin monkey movement data at various levels of sub-sampling, for the diffusion process models with unconditioned drift. The sub-sampling was performed in two ways: (1) data was removed sequentially to ensure a minimum time difference between any two points and (2) data was removed randomly while preserving the same number of points as obtained using the first sub-sampling method.](image)

It may be the case that some data cannot be modelled well by diffusion processes regardless of sub-sampling. As a very brief example, we look at the movement data of an olive baboon, *Papio anubis*, at the Mpala Research Center in Kenya [108] [37]. The trajectory of the baboon and the posterior mean of the drift is given in Figures 4.24(a) and (b). Unfortunately, we do not see much improvement from sub-sampling the data, as evidenced by the values of $\hat{Q}_C(1)$ in Figures 4.24(c) and (d). This may simply be because the diffusion process model is not appropriate for the olive baboon movement data, and other models should be considered instead.
Figure 4.24: (a) Movement data of the olive baboon. (b) Posterior mean of the drift. Test statistic $\hat{Q}_C(1)$ values (c) with a close-up view (d) after systematic subsampling of the baboon movement data.
Chapter 5

Conclusions and future work

5.1 Inference methodology

We have described a comprehensive Bayesian inference methodology for diffusion process models where the drift is estimated non-parametrically and the diffusivity is assumed to be constant. This should help reduce the complexity involved for researchers who wish to fit diffusion process models to time series data, especially for data sets with dimensions greater than one, since it can be difficult to construct a suitable parametric drift without expert knowledge. We also assert here that our drift inference methodology has several advantages over existing non-parametric estimators for the drift. Unlike the naive piece-wise constant estimator as in Section 4.5.1, we are able to increase the resolution of the domain arbitrarily high (subject to computational storage requirements) without loss of information. Combined with the diffusivity inference methodology, the availability of the posterior distributions of both the drift and the diffusivity also allows us to have an intuitive understanding of the uncertainty involved in their estimates, and in turn can be incorporated into a Bayesian hypothesis testing framework, as we have done with Bayesian discrepancy $p$-values. This is not possible with more advanced smoothed estimators (see for example the Florens-Zmirou estimator [39] and alternatives [68 p. 204]).

Our main contribution is the generalisation of Papaspiliopoulos et al.’s [81] existing non-parametric drift inference methodology for 1-dimensional diffusion process models to the multi-dimensional case. The local time process used to factorise
the likelihood for the drift does not extend beyond the 1-dimensional case and no longer forms a sufficient statistic alongside the start and end-points of the observed data. We utilise the empirical measure of the diffusion process as a replacement, which in turn requires a different factorisation of the likelihood while retaining a quadratic form with respect to the drift. This allows a conjugate Gaussian prior for the drift, with a posterior distribution that is formally shown to be a Gaussian measure. We show how to obtain the posterior mean using the conjugate gradient algorithm which treats the combination of empirical measure, observed data and prior precision operator as a system of equations. We show how the pre-conditioner for the conjugate gradient solver can be constructed from the precision operator of the posterior distribution. We also outline an implementation of the pseudo-spectral method to take advantage of the diagonal property of the empirical measure and the precision operator as a result of our choice of basis.

We also introduce a competing diffusion process model whose drift is conditioned to be a conservative vector field, which is a common choice in literature. Finding a solution for its posterior mean requires conditioning on a subspace of half the dimensionality of the unconditioned drift of the first model, which arises from the Helmholtz decomposition of vector fields. We show that for a fixed observation, this solution is not equivalent to the Helmholtz projection of the solution to the unconditioned drift inference problem unless for the specific case where the diffusivity coefficient is a multiple of the identity.

As for the contributing components of the inference methodology, we adapt an existing random walk Metropolis-Hastings algorithm for sampling diffusion bridge processes, such that we do not require knowledge of the drift of the target diffusion bridge as typically required of Langevin MCMC methods when constructing proposals arising as solutions to stochastic partial differential equations. We also describe a straightforward implementation of the random walk Metropolis-Hastings algorithm to sample from the posterior of the diffusivity coefficient given an Inverse Wishart prior.

Finally, we apply our new inference methodology to a data example from ani-
5.2 Model fit

With the lack of consideration for model validation when diffusion process models are fitted in the literature in mind, we have outlined a holistic approach to evaluating the goodness-of-fit of models, where informal checks are considered alongside formal tests, to allow for identification and rectification of sources of model discrepancy.

The few model validation tests that are available for diffusion process models are frequentist goodness-of-fit hypothesis tests, where the null and alternative models are set to those with parametric drift and diffusivity. Our contribution includes the novelty of having a null model with drift that has been estimated non-parametrically, which allows us to focus on other potential sources of model discrepancy. We implement Hong and Li’s [66] transition density-based test and show that a not insignificant finite sample correction is required for the test statistic for data with practical sample sizes. Our experiments also reveal for example that the test is sensitive to small deviations in the diffusivity coefficient, but has lower power when the shape of the drift is correctly identified, despite deviations in terms of scale.

To transition into a Bayesian setting, we adopt the transition density-based test statistic as a discrepancy variable for use with Meng’s [76] Bayesian discrepancy $p$-value and construct an estimator for it which takes advantage of the distribution of the discrepancy variable under the null. We describe how the distribution for the discrepancy $p$-value evaluated under the null can be used as a reference in evaluating
model fit as the discrepancy p-value no longer follows the uniform distribution.

As for identifying model discrepancies, we return to our capuchin monkey movement data example. Contrary to the increasingly prevalent choice of conservative drift models in the study of animal movement ecology and other sciences, we are able to determine that our conservative drift model is unsuitable for the capuchin monkey movement data when compared to our model with unconditioned drift, since the data exhibits a rotational behaviour. From a modelling standpoint, we contend that data should not be analysed in the vacuum, and that available prior information should be made use of by adopting a Bayesian framework. In the case of the capuchin monkey, possible meta information include the maximum movement speed of its species and pattern of behaviour over sustained periods of time. We describe how an outlier detection method can be constructed using the generalised residuals of the transition density-based test statistic, which identifies errors in the observed data corresponding to atypical movement speeds. We also include the possibility of the data having multi-scale properties, which unsurprisingly can be accounted for by the monkey behaving unlike a Brownian particle at shorter time-scales. We show that systematic sub-sampling of the data alleviates this problem, allowing for better model fit.

5.3 Future work

For our diffusion bridge sampler, the random walk proposals do not involve the drift of the target diffusion bridge, leading to a poorer performance when compared to Langevin MCMC-based proposals which take advantage of the drift being either linear or a gradient of a potential. We suspect that the guided proposals method by Schauer et al. [101] might serve as an improvement, where the guiding drift of the proposal process can possibly be constructed in our case by manipulation of the generalised residuals to approximate the gradient of the transition density of the target diffusion bridge process.

Our diffusion process model assumes that the diffusivity coefficient is constant, which we acknowledge as a design choice which allows for a convenient set up for
our drift inference framework since the Lamperti transformation is unavailable for multi-dimensional diffusion processes when the diffusivity is spatially dependent. This assumption of constant diffusivity, when used in conjunction with the transition density-based test which is sensitive to discrepancies in the diffusivity and with data that exhibit non-constant diffusivity, will lead to a poor model fit. Ideally a non-constant diffusivity should be assumed, and our overall inference methodology involving diffusion bridges, reparameterisation of the trajectory and drift estimation should be re-examined to allow for this.

Our choice of Gaussian prior for the drift has a precision operator with hyper-parameters which affects the regularity, mean reversion speed and scaling of the samples. Our choice of values for these hyper-parameters reflects our prior belief, and should ideally be explored by the setting of hyper-priors instead. van der Meulen et al. [116] provide an example in their construction of a reversible jump MCMC method for the 1-dimensional non-parametric drift inference setting, which could potentially be extended for our multi-dimensional setting.

Our use of Bayesian discrepancy p-value can be considered a stop-gap solution in comparison to a fully Bayesian approach to model validation including the use of Bayes factors, especially when the goal is to compare two available models directly.
Appendix A

MCMC diagnostics
A.1 Drift: conservative case

Evaluation of the drift inference procedure as in Section 3.8.2 for the case where the inferred drift is conditioned to be conservative.

Trace and autocorrelation plots

Figure A.1: The trace plots (left) of the x- (top) and y-components (bottom) of the drift evaluated at (0.28125, 1.6875), where the traces were started at different initial states of the drift. The autocorrelation plots (right) correspond to the trace with the worst autocorrelation.
Prior reproduction

Figure A.2: The x- (left) and y-components (right) of a sample from the prior (top) and posterior (bottom) of the drift.
Figure A.3: Comparison of the prior and posterior distributions for the x- (left) and y-components (right) of the drift using scatter plots (top) and QQ-plots (bottom) for one random point on the grid, and histograms (bottom) of p-values of KS2-tests for 200 randomly sampled points on the grid. 10000 samples were drawn from the prior for the KS2-tests, 500 of which were coupled with the 500 posterior samples to generate the scatter plots.
A.2 Metropolis-within-Gibbs: conservative case

Evaluation of the Metropolis-within-Gibbs sampler as in Section 3.8.4 for the case where the inferred drift is conditioned to be conservative.

Trace and autocorrelation plots

![Trace plots for the middle point of a randomly selected diffusion bridge.](image1)

**Figure A.4:** Trace (left) and autocorrelation (right) plots for the middle point of a randomly selected diffusion bridge.

![Trace plots for the drift evaluated at a randomly selected point on the grid.](image2)

**Figure A.5:** Trace (left) and autocorrelation (right) plots for the drift evaluated at a randomly selected point on the grid.
A.2. Metropolis-within-Gibbs: conservative case

Figure A.6: Trace (left) and autocorrelation (right) plots for the diffusivity.
A.3 Metropolis-within-Gibbs: capuchin monkey, conservative case

Evaluation of the Metropolis-within-Gibbs sampler applied to the capuchin monkey movement data as in Section 3.9.2 for the case where the inferred drift is conditioned to be conservative.

Trace and autocorrelation plots

Figure A.7: Trace (left) and autocorrelation (right) plots for the middle point of a randomly selected diffusion bridge.

Figure A.8: Trace (left) and autocorrelation (right) plots for the drift evaluated at a randomly selected point on the grid.
Figure A.9: Trace (left) and autocorrelation (right) plots for the diffusivity.
Bibliography


[87] Grigoris A. Pavliotis, Yvo Pokern, and Andrew M. Stuart. Parameter es-


