

Imperial College London
Department of Mathematics

Bayesian approaches to modelling physical and socio-economic systems

Louis Ellam

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy at Imperial College London, November 2018

Declaration

I confirm that the work presented in this thesis is my own, and that where information or ideas have been derived from other sources it has been fully acknowledged according to the standard referencing practices of the discipline.

Copyright

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Abstract

Modelling, simulation and experimentation are each important to understand the world that we live in. The world itself is a complex system, although the physical and socio-economic systems within are usually studied separately with a particular objective in mind. Owing to economic, practical and ethical considerations, experimentation and direct measurement are not always possible for these systems. Instead, a mathematical model can be used to retrieve the unknown quantities of interest from indirect measurement data, and to study how the system will behave in a number of different scenarios. Both tasks require solving the so-called parameter estimation problem.

There are a number of uncertainties encountered in the modelling process, including those introduced by the mathematical model and its numerical simulation, and those present in the measurement data. A statistical approach to parameter estimation allows these uncertainties to be formally accounted for. There are a number advantages for using the Bayesian framework, although there are usually non-trivial statistical and computational challenges that must first be overcome.

In this thesis, some new Bayesian approaches are developed to provide improved modelling capabilities for physical and socio-economic systems. The systems studied in this thesis involve spatially-distributed data, which includes the estimation of a spatially varying parameter via a complex function, parameter estimation for high-dimensional Gaussian models and uncertainty quantification for urban simulations. Monte Carlo methods are used throughout to obtain accurate summaries of the complex probability distributions involved. The new approaches are demonstrated with empirical simulation studies. Whilst the new approaches are shown to provide improvements on the existing ones, Bayesian inference remains a challenging and computationally intensive task. Further work is suggested to accelerate Bayesian inference, so that inferences can be made on more practical time scales.

Acknowledgements

I am grateful to my advisors Prof. Mark Girolami and Prof. Grigorios Pavliotis for their help, guidance, patience and support. I have benefited immensely from their expertise and enthusiasm, and am very fortunate to have been introduced to a number of interdisciplinary topics that interested me greatly.

I have also been very fortunate to collaborate with Sir. Alan Wilson, Dr. Iain Murray and Prof. Nicholas Zabaras, who have each been generous with their advice. I would like to thank all listed above for the support I have received, which has allowed me to develop as a researcher over the course of my PhD.

Some of the research in this thesis was carried out at the Alan Turing Institute, London. The Alan Turing Institute has provided an outstanding environment to carry out research. I would also like to express my gratitude to the Engineering and Physical Sciences Research Council (EPSRC) for funding my research.

Finally, I would like to thank friends and family for their support and, in particular, I would like to thank Vicki her continued encouragement.

Contents

Abstract	7
1 Introduction	21
1.1 Modelling physical and socio-economic systems	21
1.2 Probability notation	23
1.3 The Bayesian framework	24
1.3.1 The posterior distribution	25
1.3.2 The likelihood function	26
1.3.3 The prior distribution	27
1.3.4 The marginal likelihood	27
1.4 Thesis contributions	28
1.4.1 Chapter 3: Bayesian estimation of a multiscale spatially varying parameter	29
1.4.2 Chapter 4: Bayesian parameter estimation for high-dimensional Gaussian models	29
1.4.3 Chapter 5: Uncertainty quantification in urban simulations	30
2 Preliminaries	32
2.1 Exploring the posterior distribution	32
2.1.1 Monte Carlo integration	33
2.1.2 Importance sampling	34
2.1.3 Markov chain Monte Carlo	36
2.2 The Metropolis-Hastings algorithm	39
2.2.1 Gibbs sampler	41
2.2.2 Hamiltonian Monte Carlo	43
2.2.3 Parallel tempering	45
2.3 Sequential Monte Carlo	47
2.3.1 Sequential importance sampling	48

2.3.2	Sequential importance resampling	50
2.4	Metropolis-Hastings methods for doubly-intractable distributions	51
2.4.1	Auxiliary variable methods	52
2.4.2	Pseudo-marginal Markov chain Monte Carlo	53
2.5	Discussion	54
3	Bayesian estimation of a multiscale spatially varying parameter	55
3.1	Context and literature review	55
3.2	Mathematical modelling	58
3.2.1	Forward model	59
3.2.2	Existence, uniqueness and boundedness	60
3.2.3	Spatially varying parameter	62
3.2.4	Numerical approximation	62
3.3	Multiscale modelling	63
3.3.1	Multi-resolution analysis	63
3.3.2	Multiscale modelling of a spatially varying parameter	66
3.4	Bayesian inverse problem	66
3.4.1	Prior modelling	68
3.4.2	Likelihood specification	69
3.5	Sequential Monte Carlo samplers	70
3.5.1	Sequential Monte Carlo	71
3.5.2	Multiscale approaches	73
3.5.3	Computational considerations	79
3.5.4	Summary of proposed approach	80
3.6	Numerical results	81
3.6.1	Implementation details	81
3.6.2	Preliminary studies	82
3.6.3	Presentation of results	82
3.7	Discussion and conclusions	99

4	Bayesian parameter estimation for high-dimensional Gaussian models	101
4.1	Context and literature review	101
4.2	Linear Gaussian models	104
4.2.1	Covariance functions	105
4.2.2	Bayesian inference for linear Gaussian models	107
4.3	Computational aspects	109
4.3.1	Direct methods	109
4.3.2	Iterative methods and rational approximations	110
4.4	Methodology	113
4.4.1	Model parameter updates	114
4.4.2	Auxiliary variable updates	115
4.5	Numerical results	116
4.5.1	Random pattern precision matrices	116
4.5.2	Sparse covariances for spatial modelling of anti-social crime data	117
4.5.3	Gaussian Markov random field models specified by a whitening matrix	122
4.6	Discussion and conclusions	126
5	Uncertainty quantification in urban simulations	128
5.1	Context and literature review	129
5.2	The Boltzmann Lotka-Volterra methodology	131
5.2.1	Spatial interaction models (Boltzmann)	131
5.2.2	Structural evolution (Lotka Volterra)	134
5.2.3	Case study: the London retail system	135
5.3	Stochastic modelling of urban structure	137
5.3.1	A stochastic reformulation of the Harris and Wilson model	137
5.3.2	The principle of maximum entropy	142
5.3.3	The thermodynamics of cities and regions	145
5.4	Bayesian parameter estimation	146
5.4.1	Existing approaches	146
5.4.2	Proposed Bayesian hierarchical model	147
5.4.3	Parameter estimation in the noiseless limit	148
5.4.4	Markov chain Monte Carlo with Laplace approximations	149
5.4.5	Markov chain Monte Carlo with unbiased estimates	150

5.4.6	Implementation details	151
5.5	Case study	152
5.5.1	Preliminary studies	153
5.5.2	Prior modelling	154
5.5.3	Bayesian parameter estimation	155
5.5.4	Forward evolution of urban structure	160
5.6	Discussion and conclusions	160
6	Summary and further work	163
6.1	Summary of thesis	163
6.2	Limitations and further work	165
6.3	Closing remarks	166
	Bibliography	167
A	Diffusion processes and the Fokker-Planck equation	188
A.1	Formal definition	189
A.2	Stochastic integration	189
A.3	Existence and uniqueness	190
A.4	The Itô formula and the Fokker-Planck equation	191
A.5	The Smoluchowski equation	192
A.5.1	Convergence to equilibrium	193
A.5.2	Connection with entropy maximisation	194
A.6	Other interpretations of the stochastic integral	195

List of Tables

3.1	Table of key parameter values for numerical examples.	82
4.1	Comparison of MCMC efficiency for the random pattern precision matrices example	118
4.2	Comparison of MCMC efficiency for the anti-social crime dataset	119
4.3	Comparison of MCMC efficiency for a synthetic dataset	121
4.4	Comparison of different MCMC schemes for the Gaussian Markov random field example	125

List of Figures

2.1	Illustration of the simple Monte Carlo estimator.	34
2.2	Illustration of the importance sampling estimator.	36
2.3	Illustration of the MCMC estimator for different Metropolis-Hastings transition kernels	42
2.4	Illustration of the MCMC estimator for a Gibbs sampler	43
2.5	Illustration of the MCMC estimator for HMC	46
2.6	Illustration of the MCMC estimator for parallel tempering	47
2.7	Illustration of the SMC estimator.	51
3.1	Illustration of one-dimensional Daubechies wavelets with 4 vanishing moments. .	67
3.2	Illustration of the three quadrees for the wavelet coefficients of a two-dimensional periodic function	77
3.3	Draws from the prior for each model for $S = 0, \dots, 5$	83
3.4	The true log-permeability and velocity fields for example 1	84
3.5	Inferred posterior means for example 1 for the 10×10 and 5×5 dataset	85
3.6	Inferred posterior quantiles for example 1 for the 10×10 dataset	85
3.7	Inferred posterior quantiles for example 1 for the 5×5 dataset	86

3.8	Average acceptance rates against iteration number for the random walk Metropolis transition kernels	86
3.9	Inferred posterior means of the log-permeability for example 1 using the multiscale approach for $S = 0, \dots, 5$	87
3.10	Plot of estimated log-Bayes factors for example 1 using the multiscale approach for $S = 0, \dots, 5$	88
3.11	The inferred posterior mean and quantiles against the true log-permeability along the line $y = x$ for example 1 using the multiscale approach for $S = 0, \dots, 5$	88
3.12	The true log-permeability and velocity fields for example 2	89
3.13	Inferred posterior means for example 2 for fixed $S = 0, \dots, 3$	90
3.14	Inferred posterior means of the log-permeability for example 2 using the multiscale approach for $S = 0, \dots, 3$	91
3.15	Plot of estimated log-Bayes factors for example 2 using the multiscale approach for $S = 0, \dots, 3$	91
3.16	The inferred posterior mean and quantiles against the true log-permeability along the line $y = x$ for example 2 using the multiscale approach for $S = 0, \dots, 3$	92
3.17	The true log-permeability and velocity fields for example 3	93
3.18	Inferred posterior means of the log-permeability for example 3 using the multiscale approach for $S = 0, \dots, 5$ for the low-noise 10×10 dataset.	94
3.19	Plot of estimated log-Bayes factors for example 3 using the multiscale approach for $S = 0, \dots, 5$ for the low-noise 10×10 dataset.	94
3.20	Approximate draws from the posterior density for $S = 4$ for example 3 for the low-noise 10×10 dataset.	95
3.21	Inferred posterior quantiles for $S = 5$ for example 3 for the low-noise 10×10 dataset	95

3.22	The inferred posterior mean and quantiles against the true log-permeability along the line $y = x$ for example 3 using the multiscale approach for $S = 0, \dots, 5$ for the low-noise 10×10 dataset	96
3.23	Approximate draws from the posterior density for $S = 1$ for example 2 for the high-noise 5×5 dataset.	96
3.24	Plot of estimated log-Bayes factors for example 3 using the multiscale approach for $S = 0, \dots, 2$ for the high-noise 5×5 dataset.	97
3.25	Approximate draws from the posterior density for $S = 1$ for example 3 for the high-noise 5×5 dataset	97
3.26	Inferred posterior quantiles for example 3 for the high-noise 5×5 dataset	98
3.27	The inferred posterior mean and quantiles against the true log-permeability along the line $y = x$ for example 3 using the multiscale approach for $S = 0, \dots, 2$ for the high-noise 5×5 dataset	98
3.28	Estimated posterior marginal density plots for selected wavelet coefficients for the low-noise 10×10 dataset	98
3.29	Estimated posterior marginal density plots for selected wavelet coefficients for the high-noise 5×5 dataset	99
4.1	An illustration of the ‘fill-in’ effect for the Cholesky decomposition	110
4.2	Sparsity pattern of a well-conditioned randomly generated matrix	118
4.3	Datasets for the spatial modelling of anti-social crime example	120
4.4	Data generated from a function plus centred Gaussian noise for the Gaussian Markov random field example	124
4.5	Matrix size against the number of conjugate gradient iterations until convergence with no preconditioning and using an incomplete Cholesky preconditioner	126
5.1	Illustration of a single flow in a urban or regional system	132

5.2	Visualisation of the observation data for the London retail system over a map of London	136
5.3	The potential function for a 2-dimensional system	141
5.4	Evaluations of the log-density of the Boltzmann-Gibbs measure over a grid of 100×100 values of α and β	154
5.5	Evaluations of the R^2 value over a 100×100 grid of α and β values for the deterministic model	154
5.6	Approximate draws of the latent variables from the Boltzmann-Gibbs measure for a high-noise regime	155
5.7	Global minima of the density of the Boltzmann-Gibbs measure	156
5.8	Model evaluation for the deterministic model	157
5.9	Posterior marginal density estimates for α and β for the low-noise regime	157
5.10	Visualisation of the posterior latent variables X for the low-noise regime	158
5.11	Posterior marginal density estimates for α and β for the high-noise regime . . .	158
5.12	Visualisation of the posterior latent variables X for the high-noise regime	159
5.13	Forward evolution of the London retail system for two illustrative noise regimes	161

Chapter 1

Introduction

This thesis is concerned with the modelling of physical and socio-economic systems, for which there is a need to calibrate a mathematical model to measurement data for the purposes of simulation or prediction. There are many reasons for adopting the Bayesian framework, although there are significant statistical and computational challenges that must be overcome. Due to an improved ability to collect data and an increase in computer processing power, the development and application of new Bayesian approaches for physical and socio-economic systems is required and is the subject of this thesis.

1.1 Modelling physical and socio-economic systems

Modelling, simulation and experimentation are each important to understand the world we live in. Together they form the pillars of predictive science, and a common objective is to predict the system behaviour under a set of conditions so that informed decisions can be made. There is a long history in mathematical and statistical modelling together with numerical simulation, however, uncertainty quantification has only recently been recognised as an important topic [CS07, Smi13, Sul15].

The development of mathematical models of physical and socio-economic systems has been motivated by the need for informed decision making. Whilst experimentation and direct measurement are possible in the laboratory sciences, model building and simulation can be the best alternative in light of the economic, practical and ethical considerations for large-scale systems. For example, mathematical models have been used for simulation or prediction in

atmospheric science, geology, hydrology, urban geography and criminology [AKLR07, CJ08, SDP⁺08, Smi13, DW15, MHCC18].

A mathematical model may represent phenomena of interest $x(s)$ by a system of (coupled) equations:

$$\mathcal{A}(x(s); \theta(s)) = 0, \quad s \in \mathcal{X}. \quad (1.1)$$

Here, $\mathcal{A}$ denotes the operator acting on the domain $\mathcal{X} \subseteq \mathbb{R}^d$ that is specified together with any boundary and/or initial conditions, and θ is a model parameter that may vary over $\mathcal{X}$. For example, in Chapter 3, $\mathcal{A}$ is a differential operator that models the relationship between the permeability field of an aquifer and the pressure head for a steady-state flow, as given by the mass conservation equation combined with Darcy's law. In this case, there is a spatially varying parameter that represents the unknown permeability field. Another example is that, in Chapter 5, $\mathcal{A}$ is the differential operator of a dynamical system that describes the evolution of the sizes for a number competing retail centres, as given by the Harris and Wilson model. In this case, there are unknown parameters in the differential equation model.

Despite the availability of mathematical models, physical and socio-economic systems are generally difficult to predict. This is because they involve complex processes that act over a range of scales, and their behaviour is typically non-linear when represented mathematically. It is recognised that models provide the necessary simplifications of reality, which is usually done to permit their numerical simulation. A fairly recent paradigm is to study the joint effect of a large number of components, as opposed to the less tractable reductionist approach, in which systems are explained from their constituent parts. It has also become increasingly common for models to incorporate an element of probability to account for model misspecification, for example, this could be due to a lack of knowledge about the parameters values and/or flaws in the model itself. Whilst certain systems are thought to be inherently stochastic – for example, in quantum mechanics – this is not materially the case for the systems studied here. For example, most would agree that a Bernoulli random variable appropriately models the outcome of a well-conducted coin flip. This is even though the underlying equations of motion are thought to be deterministic and do not even appear in the probability model. In this case, the dynamics are chaotic and randomness provides a model for an unpredictable outcome.

Geophysical and socio-economic systems involve spatial dependency and are characterised

by similar length-scales. These systems are typically observed indirectly by geographically referenced data. According to Tobler's first law of geography "everything is related to everything else, but near things are more related than distant things" [Tob70]. If this were not the case, spatial analysis would not be possible, nor interesting, as the full range of conditions would be present in any arbitrarily sized region. Although there is no unified approach to modelling spatial phenomena, it is common to distinguish between discrete and continuous models of spatial variation [GDGF10]. Discrete models of spatial variation characterise the system by a finite number of regions, for example, by dividing the spatial domain into a number of sub-domains that are characterised by their centroid locations. In urban geography, gravity models have been used to estimate the flow of quantities between a discrete number of locations [Wil67]. Another example is discrete random field models, which have been used to represent spatially varying phenomena such as permeability fields in geology [FL07]. Continuous models of spatial variation, on the other hand, do not assume a discretisation of the spatial domain. Continuous models of spatial variation can be constructed from a series of basis functions over the spatial domain, for example, radial basis functions, Fourier basis functions and wavelets to obtain a sparse representation [FL07, DS17].

1.2 Probability notation

It is first necessary to briefly define the notion of a real-valued finite-dimensional random variable and its probability distribution. For an implicitly defined probability space $(\Omega, \mathcal{F}, P)^1$, the probability distribution of a $\mathcal{X}$ -valued random variable $X : \Omega \rightarrow \mathcal{X}$ is the assignment of probabilities

$$P_X(E) = P(X \in E) = P(X(\omega) \in E : \omega \in \Omega), \quad (1.2)$$

for each event $E \subseteq \mathcal{X}$. Moreover, P_X is a probability measure and when it admits a Lebesgue density $\pi_X \in L^1(\mathcal{X})$, the assignment of probabilities can be expressed via the Lebesgue integral²

$$P_X(E) = \int_E \pi_X(x) dx, \quad (1.3)$$

¹ $\mathcal{F}$ is a σ -algebra over Ω and P is a probability measure from $\mathcal{F}$ to $[0, 1]$. This means that $0 \leq P(E) \leq P(\Omega) = 1$ for all $E \in \mathcal{F}$, and $P(\cup_i E_i) = \sum_i P(E_i)$ for any countable disjoint sets $\{E_i\}$ with each belonging to $\mathcal{F}$.

² $L^1(\mathbb{R}^d)$ is defined as the space of the space of Lebesgue integrable functions on $\mathbb{R}^d$.

The probability density function π_X must be positive and integrate to one over $\mathcal{X}$. The expectation of $f(X)$ is defined by its centre of mass:

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x)\pi_X(x)dx, \quad (1.4)$$

provided that the integral is finite.

For simplicity, unless stated otherwise, both the probability distribution and probability density function of a random variable x are referred to by $\pi(x)$. The joint distribution of two random variables x and y is written as $\pi(x, y)$, and the conditional distribution of x given y , denoted $x|y$, is written as $\pi(x|y)$.

1.3 The Bayesian framework

The Bayesian framework is used throughout this thesis, and the arguments are similar to those presented in [Lin00]. The underpinning ideas of inverse probability can be traced back to the work of Thomas Bayes and Pierre-Simon Laplace in the 18th century, although it was not until the second half of the 20th century that Bayesian methods gained popularity [Lin72, Fie06]. The applicability of Bayesian methods was limited during this time due to the required computation. Since then, Bayesian methods have gained popularity with the advent of simulation tools, most notably Markov chain Monte Carlo [GLPR15]. Due to an improved ability to collect data and an increase in computer processing power, new Bayesian methods are required by modern-day science and engineering to study a range of challenging parameter estimation problems [KS06, DS17]. The Bayesian framework provides a principled way of calibrating complex mathematical and statistical models to real-world datasets. With the Bayesian framework, the practitioner's uncertainty can be fully quantified whilst making inferences or model predictions.

In the Bayesian framework, all unknown quantities are modelled as random variables, and the uncertainty in the unknown quantities is represented by the assignment of probabilities to their possible values. The probability distributions are conditioned on any available information and any assumptions that have been made. There is inherently a degree of subjectivity involved, as uncertainty is perceived differently from person to person. From a philosophical perspective, the subjective nature associated with using probability distributions to represent model or

parameter uncertainty is aligned with the Bayesian way of thinking.

There are also practical advantages for using a Bayesian approach to parameter estimation. For example, the Bayesian approach addresses issues over existence, uniqueness and data-sensitivity that one might encounter when attempting to obtain a point estimate of an unknown quantity. Another issue is that point estimates do not always adequately represent probability distributions, for example, a single draw from a high-dimensional Gaussian rarely resembles its mean. With the Bayesian approach, the most probable values of any unknown quantities can be summarised by Bayesian credible intervals, which have a simple interpretation.

The key components of the Bayesian framework are the specification of the likelihood function and the prior distribution; their product is proportional to the posterior distribution. When the likelihood function is a probability density function, the normalising constant is the marginal likelihood, which is important for model comparison. Each term is described in the proceeding subsections.

1.3.1 The posterior distribution

Let $x \in \mathbb{R}^N$ denote the unknown model parameters and let $y \in \mathbb{R}^M$ denote the observed or measured data. The prior distribution is denoted by $\pi(x)$, which describes the belief about x before y is observed, and the likelihood is given by the conditional distribution $\pi(y|x)$, which describes how likely y is to have arisen for a given parameter regime x . The solution to the parameter estimation problem is given by the posterior distribution, which is the conditional distribution $\pi(x|y)$ obtained from Bayes' rule:

$$\pi(x|y) = \frac{\pi(y|x)\pi(x)}{\pi(y)}. \quad (1.5)$$

In other words, the posterior is proportional to the prior multiplied by the likelihood and the normalising constant given by the marginal likelihood

$$\pi(y) = \int \pi(y|x)\pi(x)dx. \quad (1.6)$$

The estimation of normalising constants such as (1.6) is a notoriously challenging problem, because it involves a weighted sum over the entire state space. The problem is challenging because the weights tend to take very high values in a very small subset of the state space.

Nonetheless, the estimation of normalising constants is important for model comparison, and their estimation plays a significant role throughout this thesis.

The dominant task of Bayesian inference is the estimation of integrals of the form

$$\mathbb{E}[\phi(x)|y] = \int \phi(x)\pi(x|y)dx, \quad (1.7)$$

where ϕ is an integrable function with respect to $\pi(x|y)$. Equation (1.7) can be used to represent the posterior mean, posterior variance and smoothed density estimates of the posterior marginals. It is also possible to predict an outcome $\tilde{y}$ given the observations y via the posterior predictive distribution:

$$\pi(\tilde{y}|y) = \int \pi(\tilde{y}|y, x)\pi(x|y)dx. \quad (1.8)$$

In this way, any parameter uncertainty is fully accounted for when making predictions.

1.3.2 The likelihood function

The specification of the likelihood is often a straight-forward task. The interpretation of data requires a model of the underlying process. The likelihood function then describes any discrepancy between the model and the data, which is modelled as a random variable to represent any uncertainty. The two most common sources for any discrepancy are measurement noise and model error [K001]. The measurement noise describes the dependency between the truth and the measured value, and the model error captures the dependency between the model prediction and the truth, and includes any shortcomings of the model and its numerical approximation.

It is very common to use an additive noise model [KS06]. Let $y \in \mathbb{R}^M$ be the observed quantity, $x \in \mathbb{R}^N$ be the unknown parameter of interest and $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ be a function that is a model of the underlying process. The additive noise model is

$$y = f(x) + e, \quad (1.9)$$

where $e \sim \rho_{\text{noise}}$ is a random variable modelling the discrepancy between the model and the data. By conditioning on x , a change of variables yields the likelihood function

$$\pi(y|x) = \rho_{\text{noise}}(y - f(x)). \quad (1.10)$$

The noise distribution ρ_{noise} is typically chosen so that it can be evaluated pointwise. It is common to use a Gaussian likelihood function.

1.3.3 The prior distribution

The prior modelling expresses what is known about the unknown variable of interest, prior to collecting any measurement data. It is therefore a subjective part of the modelling process. The prior specification is particularly important when the likelihood function is insensitive to different parameter values, for example, when there is a limited number of observations.

One of the modelling issues encountered is that prior knowledge tends to be qualitative in nature. For the aquifer example that was introduced after (1.1), the practitioner may believe that the permeability field is relatively smooth but with an occasional discontinuity. For the retail example, the practitioner may believe that there are a variety of retail centres, but with a larger centre in a region that is easily accessible. It can be said that the prior biases inferences towards the possibilities that conform with our expectations.

1.3.4 The marginal likelihood

Occam's razor is the principle that states a preference towards simpler explanations, and is known for its empirical success [Mac03a, MG05]. The requirement for probability densities to be normalised automatically leads to an Occam's razor effect, as the unit of total probability must be spread over all possible outcomes. An arbitrarily complex probability model can always generate the data, however, the chance of this happening is low. Conversely, a simpler model that explains the data is more plausible as less probability is assigned to the alternative explanations.

The Bayesian approach to model selection is as follows. A set of competing models $\{\mathcal{M}_i\}_{i \in I}$ is considered, and each is assigned a prior probability $p(\mathcal{M}_i)$. Bayes' theorem conditioned on a chosen model $\mathcal{M}_i$ is then

$$p(x|y, \mathcal{M}_i) = \frac{p(y|x, \mathcal{M}_i)p(x|\mathcal{M}_i)}{p(y|\mathcal{M}_i)}, \quad (1.11)$$

and the marginal likelihood is given by

$$p(y|\mathcal{M}_i) = \int p(y|x, \mathcal{M}_i)p(x|\mathcal{M}_i)dx. \quad (1.12)$$

This is a weighted-sum over the data-likelihood $p(y|x, \mathcal{M}_i)$, where the weights are given by the model-likelihood $p(x|\mathcal{M}_i)$. A simpler model that frequently provides a good fit will have a higher marginal likelihood than a more complicated model that occasionally provides a good fit.

Lastly, another application of Bayes theorem gives

$$p(\mathcal{M}_i|y) \propto p(y|\mathcal{M}_i)p(\mathcal{M}_i), \quad (1.13)$$

which is used for the purposes of model averaging or model comparison. Any two models can be compared using Bayes factors, which are given by the ratio of posterior odds:

$$BF_{ij} = \frac{p(\mathcal{M}_i|y)}{p(\mathcal{M}_j|y)} = \frac{p(y|\mathcal{M}_i)}{p(y|\mathcal{M}_j)} \times \frac{p(\mathcal{M}_i)}{p(\mathcal{M}_j)}. \quad (1.14)$$

If $BF_{ij} > 1$ then model i should be preferred to model j , and if $BF_{ij} < 1$ then model j should be preferred to model i . The estimation of Bayes factors require that the marginal likelihoods can be estimated up to a multiplicative factor that does not depend on the choice of model.

1.4 Thesis contributions

The work in this thesis is concerned with the development of new approaches and methodologies for parameter estimation problems. The justification for the Bayesian approach has been provided. A review of simulation methods is provided in Chapter 2. The core of the thesis is the three topics forming Chapters 3, 4 and 5. An overview of these chapters is provided in the following subsections. Lastly, conclusions and suggestions for further work are given in Chapter 6.

1.4.1 Chapter 3: Bayesian estimation of a multiscale spatially varying parameter

An important inverse problem is the identification of a multiscale spatially varying parameter using indirect measurement data. A prototypical example is the estimation of a heterogeneous permeability field, which is an ill-posed inverse problem. The Bayesian framework is suitable for this problem because one can impose smoothness constraints on the permeability field via the prior modelling, however, exploring the associated posterior distribution is a non-trivial task. In Chapter 3, a new multiscale approach is proposed for multiscale statistical inversion.

The key contributions are as follows:

- A multi-resolution analysis is used to obtain a hierarchical, wavelet-based representation of a spatially varying parameter. A new multiscale approach is proposed to efficiently move between posterior densities defined on different scales using SMC.
- An adaptive approach to Bayesian model selection is proposed for multiscale statistical inversion, which is addressed without presuming the number of scales in the multiscale representation. Bayes' factors are estimated and compared for model comparison using the output from the SMC sampler.
- A computational study is performed for several realistic examples to demonstrate the capabilities of the new approach. The new approach is demonstrated for permeability estimation in groundwater flow using limited and noisy pressure data.

1.4.2 Chapter 4: Bayesian parameter estimation for high-dimensional Gaussian models

Linear Gaussian models are one of the most fundamental tools for statistical modelling. An unfortunate aspect of working with Gaussian models is that a single evaluation of a Gaussian probability density function requires solving a linear system and the evaluation of a determinant. For large Gaussian models whose covariance matrix has a sparse representation, it may be possible to solve the linear system via iterative methods, but impossible to evaluate the determinant term due to the matrix factorisations required. Whilst the determinant term can be estimated

using iterative methods, the estimates are biased and cannot be used directly in the Metropolis-Hastings algorithm. In Chapter 4, a determinant-free approach to asymptotically exact Bayesian inference is proposed instead.

The key contributions are as follows:

- A new Markov chain Monte Carlo (MCMC) method is proposed for asymptotically exact Bayesian inference to infer the hyperparameters of large-scale Gaussian models. Unlike the previously proposed schemes, the new scheme avoids the need to estimate the determinant term altogether.
- An empirical study is performed for Gaussian process models and Gaussian random Markov field models, using both synthetic and real data to compare the efficiency of the new method with an existing MCMC scheme that requires Cholesky factorisations.
- It is found that the proposed method can outperform the existing scheme, however, it is a requirement that the covariance matrix of the latent process is well-conditioned.

1.4.3 Chapter 5: Uncertainty quantification in urban simulations

The task of understanding the inner workings of cities and regions is a major challenge for contemporary science. The Harris and Wilson model is a system of coupled ordinary differential equations that describes the evolution of urban structure as a smooth function that converges to one of possibly multiple equilibrium states. A shortcoming of the Harris and Wilson model is that it does not account for any fluctuations that may arise from less predictable events. To address this shortcoming, Chapter 5 provides a stochastic reformulation of the Harris and Wilson model and an associated Bayesian inference methodology for parameter estimation and model calibration.

The key contributions are as follows:

- Chapter 5 provides the first stochastic reformulation of the Harris and Wilson model as a system of ergodic SDEs. It is shown that the structural variables can be represented by a single potential function.

- For the new model, it is shown that the entropy of the attractiveness variables increases in time. With a maximum entropy argument, the potential function can be interpreted in terms of economic constraints on the stationary distribution.
- It is shown that the parameters of the spatial interaction model can be estimated from urban structure via the Bayesian inferential framework, however, one must approach a doubly-intractable inference problem. Numerical examples for the London retail system are provided.

Chapter 2

Preliminaries

In Chapter 1 it was noted that stochastic simulation methods have made Bayesian inference possible for a large number of intractable problems. In most cases, it is necessary to construct a Markov chain to explore the posterior as the generation of exact and independent samples is a near impossible task. The construction of a suitable Markov chain is especially challenging when the target distribution is high-dimensional, anisotropic or multimodal, or when the target density contains terms that are unknown or computationally expensive to evaluate. This chapter is a review of Markov chain Monte Carlo methods, and is necessary background material for the following chapters that approach a number of challenging inference problems.

2.1 Exploring the posterior distribution

Following on from (1.7), the dominant task in Bayesian inference is the estimation of integrals of the form¹

$$\mathbb{E}_\pi[\phi(x)] = \int_{\mathcal{X}} \phi(x)\pi(x)dx, \quad \phi \in L^1(\pi), \quad (2.1)$$

where the expectation is taken with respect to a probability distribution π defined on $\mathcal{X} \subseteq \mathbb{R}^d$. It may also be useful to simulate from π , for example, to check the quality of posterior samples or to reaffirm the suitability of the prior modelling.

In all but special cases, integrals such as (2.1) are intractable and must be estimated. In a Bayesian setting, quadrature methods are unsuitable as they do not scale well with the dimension

¹ $L^1(\pi)$ is the space of functions that have finite expectation with respect to the probability distribution π .

number – at least for an assumed degree of smoothness in the integrand. $\pi(x)$ can be replaced by an approximate distribution that is easier to work with, for example, a Laplace approximation is the Gaussian obtained from a second-order Taylor expansion of the log-target density [Mac03a]. The key drawback with approximate methods is that it can be difficult to assess whether an acceptable level of accuracy has been achieved. On the other hand, Monte Carlo estimators are consistent and converge to the truth at a rate that does not depend on the dimension number. For this reason, Monte Carlo methods are the method of choice. This section provides a review of the Monte Carlo methods that are important to the subsequent chapters in the thesis.

2.1.1 Monte Carlo integration

Monte Carlo methods make use of (pseudo) random number generation, usually on a computer, and can be used to estimate integrals of the form in (2.1) [Rob04, Liu08]. A simple Monte Carlo estimator is

$$\mathbb{E}_\pi[\phi(x)] \approx \frac{1}{N} \sum_{n=1}^N \phi(X_n), \quad (2.2)$$

where each X_n is an independent and identically distributed (iid) random variable that has been drawn from the target distribution π . The estimator is illustrated for a bivariate Gaussian in Fig. 2.1. The estimator is unbiased and, as a result of the strong law of large numbers (SLLN), the estimator is also consistent in that

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \phi(X_n) = \mathbb{E}_\pi[\phi(x)]. \quad (2.3)$$

Moreover, the rate of convergence can be assessed via the central limit theorem (CLT), which states that as $N \rightarrow \infty$ the following quantity converges in distribution to a Gaussian

$$\sqrt{N} \left(\frac{1}{N} \sum_{n=1}^N \phi(X_n) - \mathbb{E}_\pi[\phi(x)] \right) \xrightarrow{d} N(0, \sigma_{\text{MC}}^2), \quad (2.4)$$

where the asymptotic variance is given by $\sigma_{\text{MC}}^2 = \text{Var}_\pi(\phi(x))$. The CLT shows that the standard deviation of the simple Monte Carlo estimator decreases at a rate proportional to $N^{-1/2}$. Whilst this rate may be described as slow, it does not depend on the dimension number and the smoothness of the test function ϕ . In practice, there may be multiple test functions that are of interest. The Monte Carlo estimator outperforms quadrature methods when there are 2 or more dimensions. The caveat is that (2.2) assumes the availability of iid exact samples, which

is typically not the case.

It is possible to use the inverse transform method to sample from univariate distributions when the inverse cumulative density function is known. It is also possible to use rejection sampling techniques to sample from a target distribution π whenever it is possible to find a proposal distribution q such that $\sup_{x \in \mathcal{X}} \pi(x)/q(x) < \infty$. Unfortunately, rejection sampling does not scale well to the higher dimension numbers encountered in a Bayesian setting.

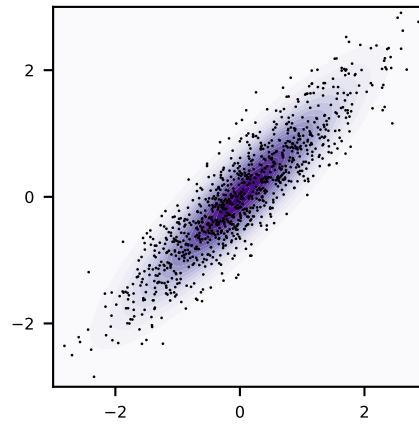


Figure 2.1: An illustration of the simple Monte Carlo estimator in (2.2) for a bivariate Gaussian (purple). 1,000 samples have been drawn from the target distribution and plotted at their drawn locations (black). The samples are equally weighted.

2.1.2 Importance sampling

Importance sampling avoids the need to sample from the target distribution and provides an alternative estimator that can be more precise than the simple Monte Carlo estimator (2.2). The idea is to rewrite the integral of interest in (2.1) as an expectation with respect to a proposal distribution $\rho(x)$ that can be readily sampled from:

$$\mathbb{E}_{\pi}[\phi(x)] = \int_{\mathcal{X}} \phi(x) \pi(x) dx = \int_{\mathcal{X}} \phi(x) \frac{\pi(x)}{\rho(x)} \rho(x) dx = \int_{\mathcal{X}} \phi(x) w(x) \rho(x) dx. \quad (2.5)$$

Here, the importance weights are defined by

$$w(x) = \frac{\pi(x)}{\rho(x)}, \quad (2.6)$$

and it is assumed that $\rho(x) > 0$ whenever $\pi(x) > 0$. As (2.5) has been rewritten as a expectation with respect to ρ , it can be approached by the previous Monte Carlo estimator to obtain:

$$\mathbb{E}_\pi[\phi(x)] \approx \frac{1}{N} \sum_{n=1}^N \phi(X_n) w(X_n), \quad (2.7)$$

where each X_n is an iid random variable that has been drawn from the proposal distribution ρ . The estimator in (2.7) is unbiased and consistent, as before, and the CLT in (2.4) gives an asymptotic variance of $\sigma_{\text{IS}}^2 = \text{Var}_\rho(w(x)\phi(x))$. As a result, importance sampling only works well if the proposal density ρ is close to the target density π .

Quite often, the target or proposal density is only known up to a normalising constant. Here the following notation is used:

$$\pi(x) = \frac{\pi_*(x)}{Z_\pi}, \quad Z_\pi = \int_{\mathcal{X}} \pi_*(x) dx < \infty, \quad (2.8)$$

where it is assumed that the function π_* is known but the constant Z_π is unknown. This is typically the case in Bayesian statistics, as the marginal likelihood in (1.5) is intractable. An application of the importance sampling estimator for $\phi(x) = Z_\pi/Z_\rho$ yields an unbiased and consistent estimator of the ratio of normalising constants

$$\frac{Z_\pi}{Z_\rho} \approx \frac{1}{N} \sum_{n=1}^N w_*(X_n), \quad (2.9)$$

where

$$w_*(x) = \frac{\pi_*(x)}{\rho_*(x)}. \quad (2.10)$$

The self-normalised importance sampling estimator is obtained by combining (2.7) and (2.9) to obtain a biased but consistent estimator:

$$\mathbb{E}_\pi[\phi(x)] \approx \frac{\sum_{n=1}^N \phi(X_n) w_*(X_n)}{\sum_{m=1}^N w_*(X_m)} = \frac{\sum_{n=1}^N \phi(X_n) w(X_n)}{\sum_{m=1}^N w(X_m)}, \quad (2.11)$$

where the X_n are as before. The estimator is biased because it is the ratio of two correlated estimators. A CLT similar to (2.4) holds in that

$$\sqrt{N} \left(\frac{\sum_{n=1}^N \phi(X_n) w(X_n)}{\sum_{m=1}^N w(X_m)} - \mathbb{E}_\pi[\phi(x)] \right) \rightarrow N(0, \sigma_{\text{SNIS}}^2), \quad (2.12)$$

where the asymptotic variance is given by $\sigma_{\text{SNIS}}^2 = \mathbb{E}_\rho[w(x)^2(\phi(x) - \mathbb{E}_\pi[\phi(x)])^2]$, as shown

in [Rob04, Liu08].

Two importance sampling estimators for the bivariate Gaussian are shown in Fig. 2.2. It is shown that a poor choice of proposal can lead to just a few samples carrying most of the weight. The resulting estimator will have a high asymptotic variance. The task of identifying a suitable proposal distribution can make importance sampling techniques unreliable in practice, which is particularly the case for higher dimensions.

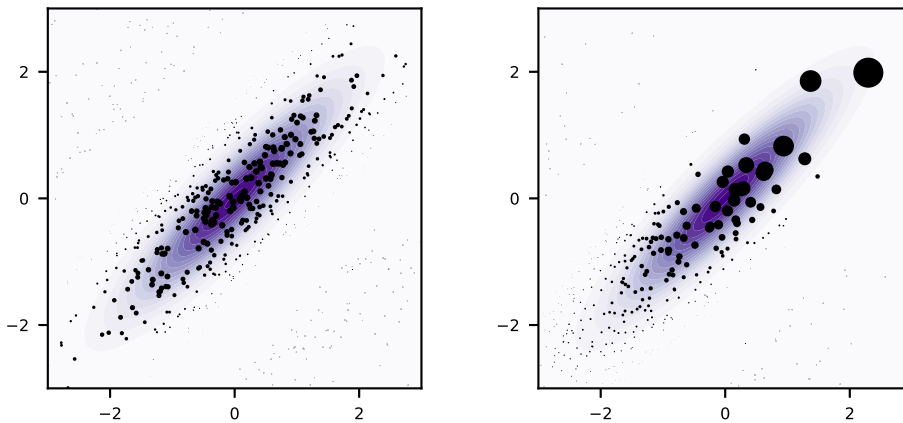


Figure 2.2: An illustration of the importance sampling Monte Carlo estimator in (2.11) for a bivariate Gaussian (purple). 1,000 samples have been drawn from a spherical Gaussian located at $(0, 0)^T$ (left) and $(-2, -2)^T$ (right) and plotted at their drawn locations (black). The marker sizes are proportional to the respective importance weights.

2.1.3 Markov chain Monte Carlo

Markov chain Monte Carlo (MCMC) methods can be relied upon for a surprisingly broad class of complex distributions by relaxing the requirement that samples are drawn independently. The advent of MCMC methods, which has its origins in computational physics, has substantially improved the applicability and popularity of Bayesian methods.

MCMC is based on the construction of a Markov chain that is usually discrete in time. A discrete-time Markov chain is a stochastic process $\{X_n \in \mathcal{X} : n \in I\}$, with $I \subseteq \mathbb{N}$ and $\mathcal{X} \subseteq \mathbb{R}^d$, that satisfies the Markov property

$$P(X_n \in A | X_{n-1}, \dots, X_0) = P(X_n \in A | X_{n-1}), \quad (2.13)$$

for each $n \in I$ and for all $A \subseteq \mathcal{X}$. Each state is only conditionally dependent on the previous

one, however, all states are typically correlated to an extent. The transition probabilities for a time-homogeneous Markov chain can be expressed in terms of a transition kernel $K(x, y)$ by

$$P(X_n \in A | X_{n-1} = x) = \int_A K(x, y) dy, \quad (2.14)$$

A requirement of MCMC is that the Markov chain is ergodic with the target density π as its unique stationary distribution. More precisely,

$$\lim_{n \rightarrow \infty} \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} = 0, \quad (2.15)$$

for π -almost all² initial conditions $x \in \mathcal{X}$, where the n -step transition probability is defined by $P^n(x, A) = P(X_n \in A | X_0 = x)$ and the total variation norm is defined by

$$\|\pi_1(\cdot) - \pi_2(\cdot)\|_{TV} = \sup_A |\pi_1(A) - \pi_2(A)|. \quad (2.16)$$

In order for π to be a stationary distribution of the Markov chain, the transition kernel K must leave π invariant:

$$\pi(y) = \int_{\mathcal{X}} K(x, y) \pi(x) dx. \quad (2.17)$$

Therefore, if a state has π as its marginal distribution, then so does any subsequent state. A sufficient but not necessary condition is the detailed balance condition for π , which is that

$$K(x, y) \pi(x) = K(y, x) \pi(y), \quad (2.18)$$

for all $x, y \in \mathcal{X}$. A transition kernel satisfying the detailed balance condition for π also leaves π invariant, since

$$\int_{\mathcal{X}} \pi(x) K(x, y) dx = \pi(y) \int_{\mathcal{X}} K(y, x) dx = \pi(y). \quad (2.19)$$

A Markov chain satisfying the detailed balance condition is also said to be reversible.

For a Markov chain with a stationary distribution π , the following two conditions are sufficient to ensure that (2.15) holds [RR04]:

1. The Markov chain is φ -irreducible: given the distribution φ on $\mathcal{X}$, the following holds for

²All except those belonging to a set of measure zero.

all $A \subseteq \mathcal{X}$ and $x \in \mathcal{X}$:

$$\varphi(A) > 0 \implies P(\tau_A < \infty | X_0 = x) > 0, \quad (2.20)$$

where $\tau_A = \inf\{n \geq 1 : X_n \in A\}$ is the first entry time to the set A .

2. The Markov chain is aperiodic: there do not exist two or more disjoint subsets $\mathcal{X}_1, \dots, \mathcal{X}_k \subseteq \mathcal{X}$ such that $P(X_{n+1} \in \mathcal{X}_{i+1} | X_n \in \mathcal{X}_i) = 1$ for $1 \leq i \leq k-1$ and $P(X_{n+1} \in \mathcal{X}_1 | X_n \in \mathcal{X}_k) = 1$.

For an ergodic Markov chain, which satisfies (2.15), the time-average is equal to the configuration average so that a SLLN similar to (2.3) applies (with probability one). This gives rise to the following MCMC estimator,

$$E_\pi[\phi(x)] \approx \frac{1}{N} \sum_{n=1}^N \phi(X_n). \quad (2.21)$$

The efficiency of a MCMC method can be assessed via a CLT result. Under certain conditions, for example, reversibility, it is possible to show that the MCMC estimator converges in distribution to a Gaussian

$$\sqrt{N} \left(\frac{1}{N} \sum_{n=1}^N \phi(X_n) - E_\pi[\phi(x)] \right) \xrightarrow{d} N(0, \sigma_{\text{MCMC}}^2), \quad (2.22)$$

provided that the asymptotic variance given by

$$\sigma_{\text{MCMC}}^2 = \text{Var}_\pi(\phi(x)) + 2 \sum_{k=1}^{\infty} \text{Cov}_\pi(\phi(X_0)\phi(X_k)), \quad (2.23)$$

is finite [KV86, RR04]. The quantity $\tau_\phi = \sigma_{\text{MCMC}}^2 / \text{Var}_\pi(\phi(x))$ is known as the integrated autocorrelation time. It is necessary to construct a Markov chain with a fast decay of autocorrelation in order for MCMC to perform well. For example, an MCMC estimator satisfying (2.22) asymptotically has the same variance as a Monte Carlo estimator with N/τ_ϕ iid samples. The quantity

$$N_{\text{effective}} = N/\tau_\phi, \quad (2.24)$$

is known as the effective sample size.

Under stronger conditions, the rate of convergence of an ergodic Markov chain can be quantified. A Markov chain is geometrically ergodic if there exists a function $V : \mathcal{X} \rightarrow [1, \infty)$ and constants $\rho < 1$ and $R < \infty$ such that for all $x \in \mathcal{X}$ the following holds:

$$\|P^n(x, \cdot) - \pi(\cdot)\|_V \leq RV(x)\rho^n, \quad (2.25)$$

where $\|\mu\|_V = \sup_{\|g\| \leq V} |\int \mu(dy)g(y)|$. An even stronger form of ergodicity applies when the function V does not depend on $x \in \mathcal{X}$, so that $V(x) = V$, which is known as uniform ergodicity. Much effort has been devoted to establish conditions under which geometric or uniform ergodicity applies [MT96, RT96a, RT96b, BBLG17], however, the required conditions are not always satisfied nor do these results guarantee a fast decay of autocorrelation over finite simulation times.

2.2 The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm provides a way to construct a reversible Markov chain that has π as its stationary distribution [MRR⁺53, Has70]. The Metropolis-Hastings algorithm iterates the following two steps for $n = 1, 2, \dots, N$ with the initial condition $X_0 = x$:

1. Draw $X'|X_{n-1} \sim q(X'|X_{n-1})$.
2. Set $X_n = X'$ with probability $\alpha(X'|X_{n-1})$, where

$$\alpha(x|y) = \min \left\{ 1, \frac{\pi(x) q(y|x)}{\pi(y) q(x|y)} \right\}. \quad (2.26)$$

Otherwise set $X_n = X_{n-1}$.

Here $q(x|y)$ is the proposal distribution and $\alpha(x|y)$ is the acceptance probability. The Metropolis-Hastings transition kernel is given by

$$K(x, y) = \alpha(y|x)q(y|x) + (1 - \alpha(y|x))\delta_x(y) = \alpha(y|x)q(y|x). \quad (2.27)$$

The detailed balance condition in (2.18) is satisfied for π , since

$$K(x, y)\pi(x) = \alpha(y|x)q(y|x)\pi(x) = \min \left\{ \pi(x)q(y|x), \pi(y)q(x|y) \right\}, \quad (2.28)$$

is symmetric in both of its arguments.

Three remarks are in order. First, the acceptance probability in (2.26) shows that the target density only needs to be evaluated up to a normalising constant that does not depend on the sampling variable. This is because the normalising constants cancel. Second, there is usually

a trade off between the acceptance rate and the distance between successive but different states. Larger moves can be rejected more frequently, but smaller moves also lead to higher autocorrelation. Third, a number of Metropolis-Hastings transition kernels can be applied in succession or at random, and these do not need to update all variables at the same time so long as the resulting Markov chain is ergodic with respect to the target distribution.

Random walk Metropolis A common choice of proposal is given by a random walk:

$$q(x'|x) = q(x' - x). \quad (2.29)$$

A usual choice is to let $q(x'|x)$ be a Gaussian distribution centred at x . As with most MCMC methods, improvements can be made by applying a linear transform to the variables being sampled. The random walk Metropolis algorithm is popular because it can be tuned easily and does not require any additional inputs, such as the gradient calculations required for the Metropolis-adjusted Langevin algorithm (MALA) and Hamiltonian Monte Carlo (HMC). It is suggested in [RG97] that the proposal variance should scale like $O(d^{-1})$ with an optimal acceptance rate of 0.234, where d is the number of dimensions.

Metropolis-adjusted Langevin algorithm MALA is motivated by a discrete approximation to a Langevin diffusion process that is ergodic with respect to π [GM94]. New states are proposed from the following Gaussian distribution that uses evaluations of the gradient of the log target density:

$$q(x'|x) = N(x'|x + \tau \nabla \ln \pi(x), 2\tau I), \quad (2.30)$$

It is suggested in [RR98] that the proposal variance should scale like $O(d^{-1/3})$ with an optimal acceptance rate of 0.574, where d is the number of dimensions.

Independence sampler Another common choice of proposal is a distribution that does not depend on the previously accepted state:

$$q(x'|x) = q(x'). \quad (2.31)$$

When it is possible to draw exact samples from the target distributions, all draws are accepted with probability one and there is no need to evaluate the target density. The independence

sampler can also be used to avoid the computation of intractable terms in the target density. Unfortunately, there are no tuning parameters to control and the acceptance rate can be unacceptably low.

An illustration of the MCMC estimator for the bivariate Gaussian is shown in Fig. 2.3, for random walk Metropolis, MALA and an independence sampler. The step sizes for the random walk Metropolis and MALA transition kernels were tuned to give an acceptance rate close to their optimal values. It is seen that the MALA algorithm provides improved mixing over random walk Metropolis. A Gaussian proposal located at $(-2, -2)^T$ is used for the independence sampler, which illustrates that high rejection rates can be expected from a poor choice of proposal.

2.2.1 Gibbs sampler

There can be two or more groups of variables that are difficult to update together. It may be possible to design better proposals by isolating a group of variables at a time. This is possible with the Gibbs sampler, in which groups of variables are updated either systematically or at random [GG84, RS94, RS97].

Each step in a Gibbs sampler can be viewed as a Metropolis-Hastings transition kernel, in which the proposal mechanism only updates a group of variables $x_{i \in I}$ at a time whilst the remaining variables $x_{i \notin I}$ are held fixed:

$$q(x'|x) = q(x'_{i \in I} | x_{i \notin I}) \delta_{x_{i \notin I}}(x'_{i \notin I}). \quad (2.32)$$

Here, $\delta_{x_0}(x) = \delta(x - x_0)$ is the Dirac delta function. The acceptance probability in the Metropolis-Hastings algorithm is therefore

$$\alpha(x'|x) = \min \left\{ 1, \frac{\pi(x')}{\pi(x)} \frac{q(x_{j \notin I} | x'_{i \in I})}{q(x'_{i \in I} | x_{j \notin I})} \right\}. \quad (2.33)$$

In the special case of being able to draw from the full-conditionals, an independence sampler can be used with $q(x'_{i \in I} | x_{i \notin I}) = \pi(x'_{i \in I} | x_{i \notin I})$, and then the proposals are always accepted:

$$\alpha(x'|x) = \min \left\{ 1, \frac{\pi(x'_{i \in I} | x'_{i \notin I}) \pi(x'_{i \notin I})}{\pi(x_{i \in I} | x_{i \notin I}) \pi(x_{i \notin I})} \frac{\pi(x_{j \notin I} | x'_{i \in I})}{\pi(x'_{i \in I} | x_{j \notin I})} \right\} = 1. \quad (2.34)$$

Nonetheless, Gibbs sampling is not always efficient as there can be slow exploration of the state

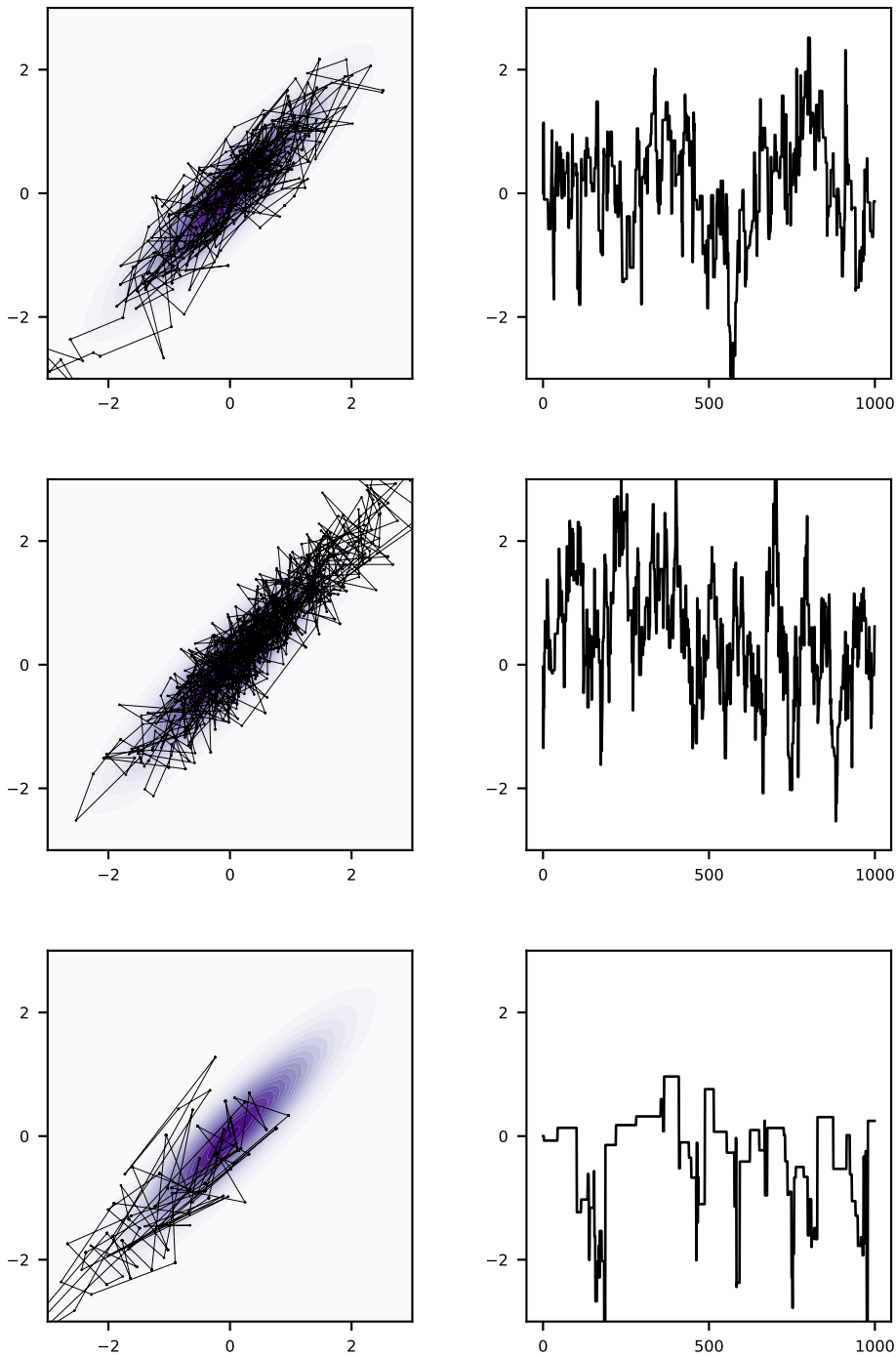


Figure 2.3: An illustration of the MCMC estimator in (2.21) for different Metropolis-Hastings transition kernels for a bivariate Gaussian (purple): random walk Metropolis (top), MALA (middle) and independence sampler (bottom). For each, a Markov chain of length 1,000 has been simulated and plotted at the drawn locations (left). Trace plots are shown for the first component (right).

space due to strong correlations between the groups of variables. This point is illustrated for a bivariate Gaussian in Fig. 2.4. It is also illustrated that a Gibbs sampler is better able to explore the state space for a parameterisation with independent components. For a Gaussian

target density, such a parameterisation can be achieved by applying a linear transform $y = Lx$ where $\Sigma = LL^T$ is the covariance matrix.

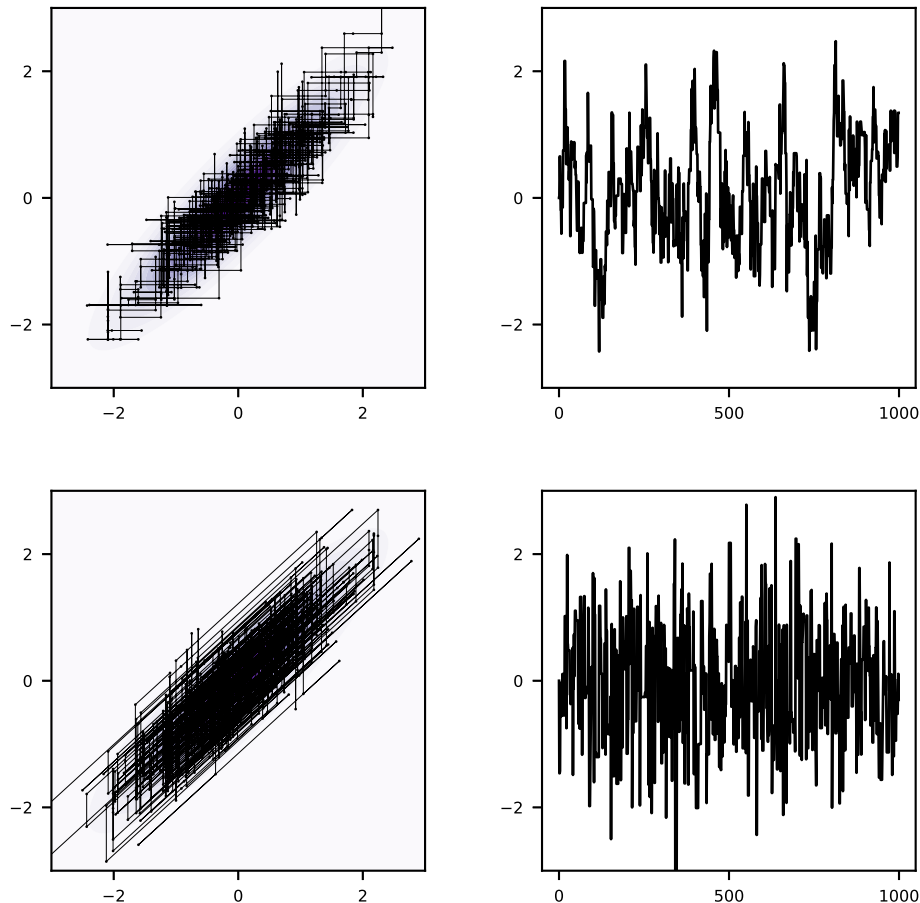


Figure 2.4: An illustration of the MCMC estimator in (2.21) for Gibbs samplers for a bivariate Gaussian (purple): a Gibbs sampler (top), a Gibbs sampler after applying a linear transformation so that the transformed components are independent (bottom). For each, a Markov chain of length 1,000 has been simulated and plotted at the drawn locations (left). Trace plots are shown for the first component (right).

2.2.2 Hamiltonian Monte Carlo

One of the most efficient ways to incorporate the gradient of the log target density into the proposal is via Hamiltonian Monte Carlo (HMC) [DKPR87, Nea11]. The state space x is augmented with the auxiliary variables p . MCMC is performed on the joint distribution and the extra variables are discarded. The joint distribution is given by

$$\pi(x, p) \propto e^{-H(x, p)}, \quad (2.35)$$

where the function $H(x, p)$ is the Hamiltonian given by

$$H(x, p) = -\ln \pi(x) + |p|^2/2. \quad (2.36)$$

HMC is an iterative procedure comprising of two steps. First, a Gibbs step is performed on the auxiliary variables by drawing from the full-conditional:

$$p \sim N(0, I). \quad (2.37)$$

Second, a Metropolis-Hastings step is performed where new values of x and p are proposed according to Hamiltonian dynamics, as given by the phase flow $g^t : (x(0), p(0)) \mapsto (x(t), p(t))$ described by

$$\frac{dx}{dt} = \frac{\partial H}{\partial p}, \quad \frac{dp}{dt} = -\frac{\partial H}{\partial x}. \quad (2.38)$$

There are three important properties of Hamiltonian dynamics that motivate the second step.

First, Hamiltonian dynamics conserves the Hamiltonian:

$$\frac{dH}{dt} = \frac{dx}{dt} \frac{\partial H}{\partial x} + \frac{dp}{dt} \frac{\partial H}{\partial p} = \frac{\partial H}{\partial p} \frac{\partial H}{\partial x} - \frac{\partial H}{\partial x} \frac{\partial H}{\partial p} = 0. \quad (2.39)$$

Second, as a result of Liouville's theorem, Hamiltonian dynamics preserves volume in phase space

$$\frac{d}{dt} \int_{V(t)} dx dp = \int_{V(t)} \left(\frac{dx}{dt}, \frac{dp}{dt} \right) \cdot \hat{n} ds = \int_{V(t)} \nabla \cdot \left(\frac{dx}{dt}, \frac{dp}{dt} \right) dx dp = 0, \quad (2.40)$$

which is the case because the phase flow is divergence free:

$$\nabla \cdot \left(\frac{dx}{dt}, \frac{dp}{dt} \right) = \frac{\partial^2 H}{\partial x \partial p} - \frac{\partial^2 H}{\partial p \partial x} = 0. \quad (2.41)$$

Third, Hamiltonian dynamics can be reversed by changing the direction of time. The same can be achieved by changing the sign of the momentum variables, since if $g^t(x, p) = (x', p')$ then $g^t(x', -p') = (x, -p)$. Changing the sign of the momentum variables also conserves the Hamiltonian, since $H(x, p) = H(x, -p)$. These three properties suggest a Metropolis-Hastings transition kernel that will always be accepted. Reversibility and volume preservation mean that it is possible to construct a symmetric proposal by simulating Hamiltonian dynamics and then changing the sign of the momentum variables; and conservation of the Hamiltonian means that the acceptance probability will always be one.

In practice, Hamiltonian dynamics must be approximated numerically via a leapfrog integrator for a length of time ϵL , where ϵ is the step size and L is the number of leapfrog steps. The approximate phase flow is denoted $\hat{g}^T : (\hat{x}(0), \hat{p}(0)) \mapsto (\hat{x}(t), \hat{p}(t))$, in which a single leapfrog step is given by:

$$\hat{p}(t' + \epsilon/2) = \hat{p}(t') + \epsilon/2 \nabla \ln \pi(\hat{x}(t')), \quad (2.42)$$

$$\hat{x}(t' + \epsilon) = \hat{x}(t') + \epsilon \hat{p}(t' + \epsilon/2), \quad (2.43)$$

$$\hat{p}(t' + \epsilon) = \hat{p}(t' + \epsilon/2) + \epsilon/2 \nabla \ln \pi(\hat{x}(t' + \epsilon)). \quad (2.44)$$

Unlike Hamiltonian dynamics, the leapfrog integrator does not conserve the Hamiltonian due to numerical error. It is, however, volume preserving because each step in (2.42)-(2.44) is a shear transformation with an identity Jacobian. It is also reversible, as before, since if $\hat{g}^t(x, p) = (x', p')$ then $\hat{g}^t(x', -p') = (x, -p)$. By changing the sign of the momentum variables after running the leapfrog integrator, the proposal is symmetric and the Metropolis-Hastings acceptance probability is given by

$$\alpha(x', p' | x, p) = \min \left\{ 1, e^{H(x, p) - H(x', p')} \right\}. \quad (2.45)$$

As the numerical error due to the leapfrog integrator tends to be relatively small, it is possible to make large proposals using HMC whilst maintaining a high acceptance rate. An illustration for the bivariate Gaussian is shown in Fig. 2.5. The random walk behaviour apparent in the other MCMC samplers considered is greatly reduced. An appropriate specification of ϵ and L is required for HMC to perform well. For Fig. 2.5, the step size was tuned to give an acceptance rate close to 90%. The no-U-turn sampler offers a way to determine L adaptively [HG14]. Further improvements can be achieved by incorporating the local curvature of the log target density into the proposal [GC11, LMW18].

2.2.3 Parallel tempering

Multimodal distributions can pose severe problems for Markov chain simulation. Markov chains may spend a long time at each mode and may even become trapped in a mode over the finite simulation times. This occurs when regions of high-probability are separated by regions of very low probability, and the movement between the high-probability regions is a rare event. If this

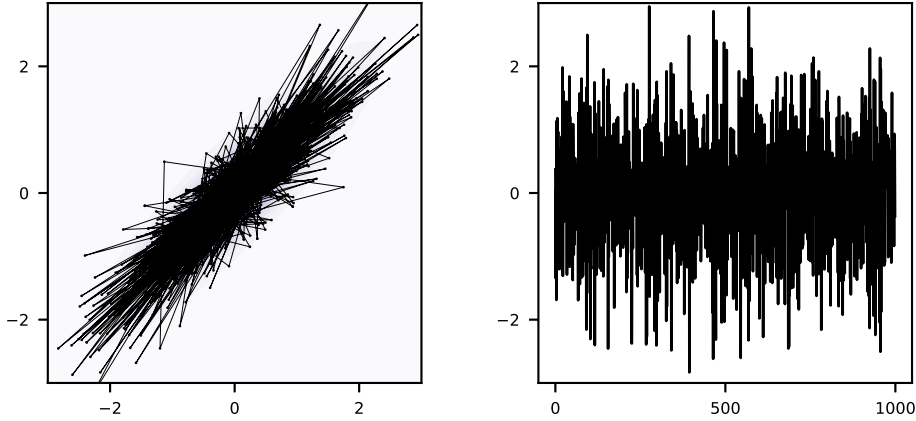


Figure 2.5: An illustration of the MCMC estimator in (2.21) for HMC for a bivariate Gaussian (purple). A Markov chain of length 1,000 has been simulated and plotted at the drawn locations (left). Trace plots are shown for the first component (right).

is the case, the resulting Markov chain is likely to have very high autocorrelation.

These problems can be remedied by introducing a sequence of tempered distributions. In Bayesian statistics, it is common to define the sequence of densities

$$\pi_i(x) \propto \pi_0(x)^{1-\beta_i} f(x)^{\beta_i}, \quad i = 1, \dots, M, \quad (2.46)$$

where $0 = \beta_0 < \beta_1 < \dots < \beta_M = 1$ are a sequence of tempering parameters such that $\pi_M = \pi$. The tempered distributions become more like the target as the tempering parameter increases.

One possibility is simulated tempering [MP92], in which MCMC is performed over an augmented space with the introduction of a temperature index. Simulated tempering can be very difficult to tune as the amount of time spent at each temperature depends on the unknown normalising constants of the distributions in (2.46). An alternative is parallel tempering [SW86, Gey91, HN96], in which MCMC is performed over a much larger state space where the joint distribution is given by

$$\pi(x_1, \dots, x_M) = \prod_{i=1}^M \pi_i(x_i). \quad (2.47)$$

Each x_i can be updated independently with Metropolis-Hastings transition kernels, however, additional steps are performed in which two randomly selected neighbouring pairs are swapped with probability

$$\alpha(x'_{i+1}, x'_i | x_{i+1}, x_i) = \min \left[1, \frac{\pi_i(x_{i+1})}{\pi_i(x_i)} \frac{\pi_{i+1}(x_i)}{\pi_{i+1}(x_{i+1})} \right], \quad (2.48)$$

to ensure that the parallel chains communicate. The Markov chain for π_M is retained and the remaining samples are discarded. An illustration is provided in Fig. 2.6 for a mixture of bivariate Gaussians, where parallel tempering is performed for $M = 5$ different temperatures and with HMC transition kernels. It is shown that for the target density, the HMC transition kernels do not frequently propose new moves between the two separated areas of higher probability density. These difficulties are alleviated by the introduction of swap proposals.

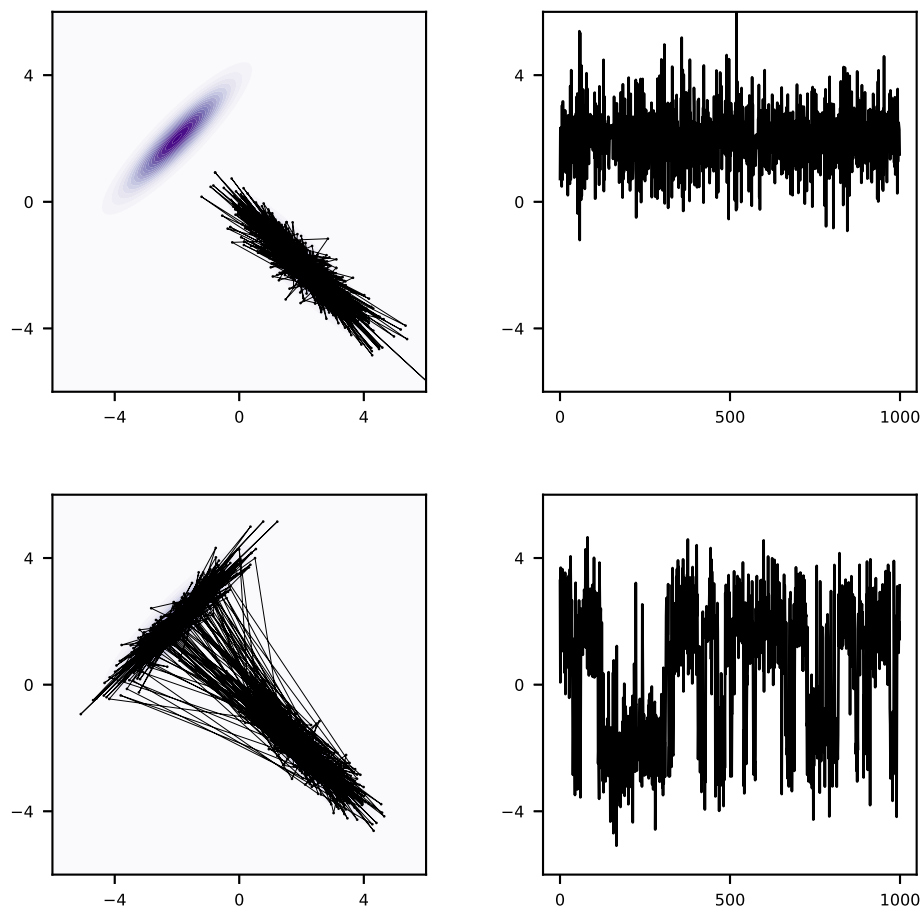


Figure 2.6: An illustration of the MCMC estimator in (2.21) for parallel tempering with HMC transition kernels for a mixture of bivariate Gaussians (purple): no swap transition kernels (top) and with swap transition kernels (bottom). For each, a Markov chain of length 1,000 has been simulated and plotted at the drawn locations (left). Trace plots are shown for the first component (right).

2.3 Sequential Monte Carlo

If the target probability distribution is complex or multimodal it can be difficult to generate a stationary Markov chain in a reasonable amount of time. For example, the parallel tempering

algorithm may be prohibitively expensive when the Markov chain is defined on a very large state space and needs to reach its stationary distribution. The unbiased estimator in (2.9) may also be of interest for the estimation of normalising constants, for example, for model comparison. It can be advantageous to combine importance sampling techniques with MCMC transition kernels. This can be achieved with sequential Monte Carlo (SMC) methods [Jar97a, Jar97b, DGA00, Nea01, Cho02, DMDJ06].

2.3.1 Sequential importance sampling

SMC is based on sequential importance sampling. The importance sampling techniques in Section 2.1.2 can be applied to the random variable $x = (x_1, \dots, x_M)$ by using a proposal distribution that is constructed sequentially:

$$\rho(x_1, \dots, x_M) = \rho_1(x_1)\rho_2(x_2|x_1) \dots \rho_M(x_M|x_1, \dots, x_{M-1}). \quad (2.49)$$

When the target distribution can be expressed as $\pi(x_1, \dots, x_M) = \pi_M(x_1, \dots, x_M)$, the importance weight in (2.6) is given by $w(x_1, \dots, x_M) = w_M(x_1, \dots, x_M)$, which can be evaluated recursively by

$$w_j(x_1, \dots, x_j) = w_{j-1}(x_1, \dots, x_{j-1}) \frac{\pi_j(x_1, \dots, x_j)}{\rho_j(x_j|x_1, \dots, x_{j-1})\pi_{j-1}(x_1, \dots, x_{j-1})}, \quad (2.50)$$

for $j = 1, \dots, M$. It is then possible to improve upon standard importance sampling techniques by using MCMC transition kernels in the proposal distribution in (2.49).

Annealed importance sampling is a particular instance of sequential importance sampling, and is based on the sequence of auxiliary densities

$$\pi_j(x) = \frac{f_j(x)}{Z_j}, \quad Z_j = \int_{\mathcal{X}} f_j(x) dx < \infty, \quad (2.51)$$

for $j = 1, \dots, M$, where it is assumed that each f_j is known pointwise and each normalising constant Z_j is unknown. It is assumed that it is possible to sample from the initial distribution π_0 and the target distribution is given by $\pi_M = \pi$. A sequence of transition kernels K_j is defined with each leaving the corresponding auxiliary density π_j invariant:

$$\pi_j(y) = \int K_j(x, y) \pi_j(x) dx. \quad (2.52)$$

A sequence of reverse transition kernels L_j is defined to satisfy

$$K_j(x, y)\pi_j(x) = L_{j-1}(y, x)\pi_j(y). \quad (2.53)$$

In other words, $K_j(x, y)$ is the conditional density of $Y|X$ and $L_{j-1}(y, x)$ is the conditional density of $X|Y$ when the random variables X and Y are marginally distributed by π_j . The sampler targets the joint density

$$\pi(x_1, \dots, x_M) = \pi_M(x_M) \prod_{j=1}^{M-1} L_j(x_{j+1}, x_j), \quad (2.54)$$

which admits the target density π_M as a marginal by construction. The following choice of proposal is used

$$\rho(x_1, \dots, x_M) = \pi_1(x_1) \prod_{j=1}^{M-1} K_{j+1}(x_j, x_{j+1}), \quad (2.55)$$

which may be simulated from by sampling from the initial distribution and by successively applying the transition kernels. The importance weights can be estimated recursively by using (2.50):

$$\begin{aligned} w_j(x_1, \dots, x_j) &= w_{j-1}(x_1, \dots, x_{j-1}) \frac{\pi_j(x_j) L_{j-1}(x_j, x_{j-1})}{\pi_{j-1}(x_{j-1}) K_j(x_{j-1}, x_j)}, \\ &= w_{j-1}(x_1, \dots, x_{j-1}) \frac{\pi_j(x_{j-1})}{\pi_{j-1}(x_{j-1})}, \end{aligned} \quad (2.56)$$

for $j = 1, \dots, M$. Therefore, the weights can be updated before the transition kernels are applied.

The scheme is valid for any Metropolis-Hastings transition kernels. The key advantage over standard MCMC techniques is that the transition kernels do not need to be repeatedly applied until the Markov chains mix well; the transition kernels perform a very different role here. For example, a valid choice of transition kernel is the ‘do nothing’ operator $K_j(x, y) = \delta(y - x)$, which yields a standard importance sampling estimator. On the other hand, this means a good choice of transition kernel is necessary to gain a significant improvement over the more standard importance sampling techniques. Unlike for standard MCMC methods, the transition kernels for sequential importance sampling can be determined adaptively after applying transition kernels for an auxiliary distribution. This means that there are many possibilities to gradually adapt the importance sampling proposal to the target distribution. Another key advantage is that the

transition kernels can be applied in parallel.

2.3.2 Sequential importance resampling

Another view of sequential importance sampling is that each density π_j is approximated by a mixture of Dirac masses $\hat{\pi}_j^N$, known as a particle approximation, given by

$$\hat{\pi}_j^N(x) = \sum_{n=1}^N W_j^{(n)} \delta_{X_j^{(n)}}(x), \quad W_j^{(n)} = \frac{w_j^{(n)}}{\sum_{k=1}^N w_j^{(k)}}. \quad (2.57)$$

Here, $\{X_j^{(n)}\}_{n=1}^N$ are the particle locations and $\{W_j^{(n)}\}_{n=1}^N$ and $\{w_j^{(n)}\}_{n=1}^N$ are the respective normalised and unnormalised particle weights. Each particle approximation is chosen to satisfy a SLLN:

$$\int_{\mathcal{X}} \phi(x) \pi_j(x) dx = \lim_{N \rightarrow \infty} \sum_{n=1}^N W_j^{(n)} \phi(X_j^{(n)}). \quad (2.58)$$

Sequential importance resampling is an extension of sequential importance sampling, and is based on the observation that the particles can be resampled and reweighted at any point during sequential importance sampling in such a way that a SLLN continues to hold. For example, multinomial resampling can be performed so that each particle is replaced with a random copy of another with the respective probabilities given by the normalised importance weights. After resampling, the particles are equally weighted and the particle approximation is given by

$$\bar{\pi}_j^N(x) = \frac{1}{N} \sum_{n=1}^N \delta_{\bar{X}_j^{(n)}}(x), \quad (2.59)$$

and satisfies

$$\int_{\mathcal{X}} \phi(x) \pi_j(x) dx = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \phi(\bar{X}_j^{(n)}), \quad (2.60)$$

where $\{\bar{X}_j^{(n)}\}_{n=1}^N$ are the updated particle locations. The unnormalised importance weights after resampling $\{\bar{w}_j^{(n)}\}_{n=1}^N$ are set to their average before resampling so that

$$\frac{Z_j}{Z_0} \approx \frac{1}{N} \sum_{i=1}^N w_j^{(i)} = \frac{1}{N} \sum_{i=1}^N \bar{w}_j^{(i)}, \quad \text{where } Z_j = \int_{\mathcal{X}} f_j(x) dx, \quad (2.61)$$

are unbiased and consistent estimators. The CLT result in (2.12) no longer applies when the particles are resampled. A CLT for sequential importance sampling with multinomial resampling performed at each step is given in [Cho04, DMDJ06].

Whilst the resampling step leaves the expectation of estimators unchanged, it causes a change to the variance. Information is inevitably lost by resampling the particles, however, resampling can be a particularly useful strategy to avoid applying expensive transition kernels to particles that have almost no weight. It is best to resample the particles before applying any expensive transition kernels to avoid wasted computation [DMDJ06]. In this way, the MCMC transition kernels perturb the resampled particles to restore lost diversity in the particle approximation. The benefit of resampling is demonstrated in Fig. 2.7, where AIS has been performed with and without resampling the particles for a mixture of bivariate Gaussians. Many of the particles have small importance weights for the case without resampling. For this example, 10 auxiliary distributions were used with a single HMC transition kernel per particle per auxiliary distribution.

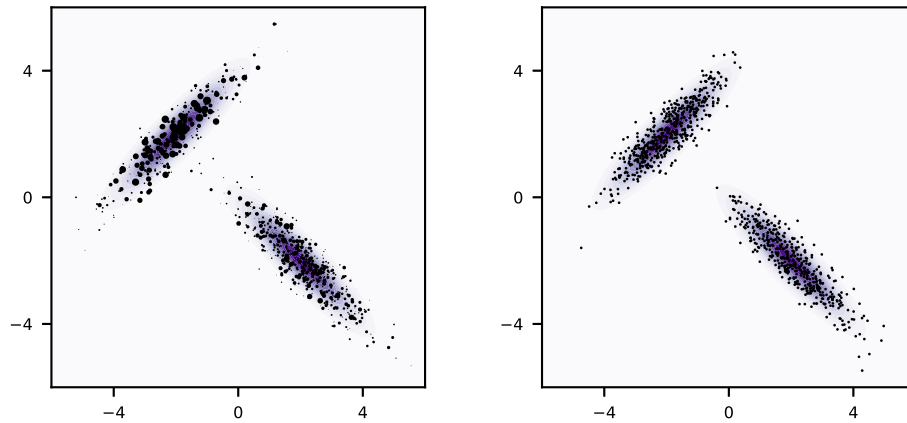


Figure 2.7: An illustration of the SMC estimator in (2.60) for a mixture of bivariate Gaussians (purple). SMC has been performed for 1,000 particles with no resampling (left) and resampling at every step (right). Particles are plotted at their drawn locations (black). The marker sizes are proportional to the respective importance weights.

2.4 Metropolis-Hastings methods for doubly-intractable distributions

In many cases the posterior density is of the form

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}, \quad \pi(y|\theta) = \frac{f(y;\theta)}{z(\theta)}, \quad z(\theta) = \int f(y;\theta)d\theta, \quad (2.62)$$

where $\pi(y)$ and $z(\theta)$ are both intractable. This scenario arises when the unknown terms do not have a closed-form solution or are prohibitively expensive to evaluate. The posterior densities

are said to be ‘doubly intractable’, and pose significant computational challenges. Whilst the densities only need be known up to a normalising constant for the Metropolis-Hastings algorithm to be used, the unknown normalising constant cannot depend on the variables being updated as the acceptance probability in (2.26) becomes

$$\alpha(\theta'|\theta) = \min \left\{ 1, \frac{f(y; \theta')\pi(\theta')}{f(y; \theta)\pi(\theta)} \frac{z(\theta)}{z(\theta')} \frac{q(\theta|\theta')}{q(\theta'|\theta)} \right\}, \quad (2.63)$$

in which the ratio $z(\theta)/z(\theta')$ does not cancel. One way of proceeding is to replace $z(\theta)$ with an estimate, for example, by fitting a model to importance sampling estimates. However, this is at the expense of asymptotic exactness.

2.4.1 Auxiliary variable methods

A solution was first proposed in [MPRB06], in which a Metropolis-Hastings algorithm is proposed with the pair

$$\pi(x, \theta|y) = \pi(x|y, \theta) \frac{f(y; \theta)\pi(\theta)}{z(\theta)\pi(y)}, \quad q(x', \theta'|x, \theta) = q(\theta'|\theta) \frac{f(x; \theta')}{z(\theta')}, \quad (2.64)$$

which yields the (tractable) acceptance probability

$$\alpha(\theta'|\theta) = \min \left\{ 1, \frac{\pi(x'|y, \theta')}{\pi(x|y, \theta)} \frac{f(y; \theta')\pi(\theta')}{f(y; \theta)\pi(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)} \frac{f(x; \theta)}{f(x'; \theta')} \right\}. \quad (2.65)$$

There are three significant drawbacks. First, it is necessary to produce exact samples from the intractable distribution $\pi(y|\theta)$. Second, the efficiency of the algorithm depends on being able to find a good auxiliary distribution $\pi(x|y, \theta)$. The choice of $\pi(x|y, \theta) = f(x; \theta)/z(\theta)$ would reintroduce the intractable term back into the acceptance probability. Third, proposing x with what is effectively an independence sampler can lead to high rejection rates.

An improved solution was proposed in [MGM06], in which the target distribution is augmented by

$$\pi(x, \theta, \theta'|y) = \pi(\theta) \frac{f(y'|\theta)}{z(\theta)} q(\theta'|\theta) \frac{f(x|\theta')}{z(\theta')}. \quad (2.66)$$

The augmented distribution is simulated from via the following two updates. First, the full conditional $x, \theta'|y, \theta$ can be drawn from exactly and therefore x and θ' can be updated via a block-Gibbs step. Second, a new value of θ can be proposed by a swap proposal with θ' that is

accepted with probability

$$\alpha(\theta'|\theta) = \min \left\{ 1, \frac{f(y; \theta') \pi(\theta')}{f(y; \theta) \pi(\theta)} \frac{q(\theta|\theta')}{q(\theta'|\theta)} \frac{f(x; \theta)}{f(x; \theta')} \right\}. \quad (2.67)$$

The scheme addresses two of the three issues, however, it is still necessary to be able to simulate from the intractable distribution.

2.4.2 Pseudo-marginal Markov chain Monte Carlo

It is not always possible to draw from the intractable distribution. Fortunately, one can still obtain an asymptotically exact scheme. First, it suffices to say that one can plug an unbiased positive estimate of $1/z(\theta')$, denoted $1/\hat{z}(\theta', u')$, for a newly proposed θ' into the Metropolis acceptance probability. Provided that the estimate $1/\hat{z}(\theta, u)$, for the previously accepted θ , is retained and used each time, then one obtains a valid MCMC scheme. This is known as pseudo-marginal MCMC [AR09]. It is a valid MCMC scheme because it is the Metropolis-Hastings algorithm performed on the augmented density

$$\pi(u, \theta|y) \propto \frac{f(y; \theta)}{\hat{z}(\theta, u)} q(u|\theta), \quad \text{with} \quad \pi(\theta|y) = \int \pi(u, \theta|y) du, \quad (2.68)$$

with the proposal $q(\theta', u'|\theta) \propto q(u'|\theta') q(\theta'|\theta)$. With this choice of proposal, the unknown $q(u|\theta)$ terms in the acceptance probability cancel. The fact that part of the proposal is an independence sampler means that there will be high rejection rates if there is high variability in the unbiased estimates of $1/z(\theta')$ – this can make the scheme prohibitive to use.

A consistent estimator of $1/z(\theta)$ can be obtained by taking the reciprocal of an importance sampling estimator, as done in [LGA⁺15, WM17]. The reciprocal of such an estimator is biased, so cannot be used in a pseudo-marginal MCMC scheme. It is, however, possible to construct a series $\{S_n\}_{n=0}^\infty$ satisfying $\lim_{n \rightarrow \infty} S_n = 1/z(\theta)$, so that the randomly truncated series

$$S = S_0 + \sum_{n=1}^N \frac{S_n - S_{n-1}}{P(n \leq N)}, \quad (2.69)$$

is an unbiased estimate of $1/z(\theta)$, which can be used in a pseudo-marginal MCMC scheme. This is the case because

$$\mathbb{E}[S] = \mathbb{E}[S_0] + \sum_{n=1}^{\infty} \frac{\mathbb{E}[S_n] - \mathbb{E}[S_{n-1}]}{P(n \leq N)} \mathbb{E}[\mathbb{I}_{[n \leq N]}(n)] = \lim_{n \rightarrow \infty} \mathbb{E}[S_n] = 1/z(\theta). \quad (2.70)$$

A distribution of N that seldom gives large values of N is less expensive to work with, but gives an estimator with high variance due to the dominator in the series.

The unbiased estimates of S can be negative due to variability in the underlying importance sampling estimates. It is possible to proceed by performing MCMC on the following distribution

$$\hat{\pi}(u, \theta|y) \propto \sigma(u, \theta) \frac{f(y; \theta)}{\hat{z}(\theta, u)} q(u|\theta), \quad (2.71)$$

where $\sigma(u, \theta)$ is the sign of $1/\hat{z}(\theta, u)$, since

$$\int f(\theta) \pi(\theta|y) d\theta = \frac{\iint \sigma(u, \theta) f(\theta) \hat{\pi}(u, \theta|y) du d\theta}{\iint \sigma(u, \theta) \hat{\pi}(u, \theta|y) du d\theta} = \lim_{N \rightarrow \infty} \frac{\sum_{i=1}^N \sigma(u^{(i)}, \theta^{(i)}) f(\theta^{(i)})}{\sum_{k=1}^N \sigma(u^{(k)}, \theta^{(k)})}, \quad (2.72)$$

for a suitably defined Markov chain $\{u^{(i)}, \theta^{(i)}\}_{i=1}^N$. It is noted that high variance in the S_n terms is problematic for two reasons. First, a large number of negative signs reduces the effective sample size of the estimator. Second, variance in the unbiased estimates can lead to high rejection rates in the pseudo-marginal MCMC scheme.

2.5 Discussion

It is possible to draw correlated samples from complex probability distributions by constructing a Markov chain from a sequence of transition kernels. A good sequence of transition kernels will allow the Markov chain to reach the stationary distribution rapidly, and will have low enough autocorrelation for reliable posterior summaries to be made. This is made possible because of the Metropolis-Hastings algorithm. The combination of Metropolis-Hastings transition kernels with importance sampling techniques can yield efficient Monte Carlo estimators, which is especially useful for the estimation of normalising constants. For challenging inference problems, the choice of method ultimately depends on the problem at hand, the properties of the target distribution and the computational considerations involved.

Chapter 3

Bayesian estimation of a multiscale spatially varying parameter

Many physical and socio-economic systems are characterised by a spatially varying parameter. When the spatial field is observed indirectly via a complex function, the correlation structure of the unknown must be specified a priori as the observation data contains limited information about the smoothness of the unknown. If the prior modelling is overly restrictive, the posterior may be concentrated around an undesirably smooth field. If the prior modelling is too vague, the posterior distribution can be severely multimodal and many of the solutions to the inverse problem are too rough to be plausible. In this chapter, a new multiscale approach is proposed in view of these challenges. A family of multiscale models are adaptively proposed and compared in a computationally efficient manner using sequential Monte Carlo methods.

The work in this chapter was performed in collaboration with Nicholas Zabaras and Mark Girolami, and parts have been published in [EZG16].

3.1 Context and literature review

The identification of a spatially varying parameter is an important task for modelling physical and socio-economic systems. The task is particularly challenging when the unknown parameter is observed indirectly via a complex function, for example, for an inverse problem. The problem is often ill-posed in that it may not have a unique solution, or its solution may be highly sensitive to noise in the measurement data. A prototypical example of an under-determined inverse

problem is the task of permeability estimation from indirect flow data. The approach developed here is primarily motivated by this example, however, the approach is general and applicable to other inverse problems.

Classical approaches to inverse problems formulate the task as an optimisation problem; the solution to the inverse problem is a point estimate without any quantification of uncertainty [Tik63, KS06]. For the reasons discussed in Section 1.3, the Bayesian approach to inverse problems is preferred. For permeability estimation, there are an infinite number of permeability fields that explain a partially observed flow and many of these fields are too rough to be plausible. Additional information must be provided via the prior modelling so that the posterior distribution also represents the practitioners beliefs about the unknown parameter. The Bayesian approach has been used for inverse problems across a variety of domains, including tomography [Wei97, KKSV00, DS16], hydrology [LHB⁺02a, WZ11, KBJ14, BJMS15], heat conduction [WZ04], solid mechanics [Kou09] and many other physical, biological and social systems [ABH03, Tar05, KS06, CS07, MZ09, CVGO11, RHL14, DS17].

Spatially varying parameters belong to an infinite-dimensional space. For the purposes of inference, the unknown parameter must be represented by a finite number of variables. It is common practice to discretise the unknown parameter using a grid that coincides with the finite element mesh used to numerically approximate the forward model. Standard models from spatial statistics can be used to represent the unknown parameter. Popular choices of model include Gaussian processes [OCR97, BCORR96, EHL06] and Markov random fields [BGHM95, LHB⁺02a]. These models assume a basic correlation structure, whereas permeability fields are essentially non-stationary with non-trivial correlation structures [WT02, XT04, FL07, LTT10]. Interpolation schemes with sparse grids have been used to represent the unknown parameter, however, these methods tend to perform poorly on fields with local variations or discontinuities [WZ11]. Other authors have considered the use of level sets for permeability estimation [MEMDG10, ILS15]. To reduce the dimensionality of the problem, the unknown can also be modelled by a truncated spectral expansion, for example, using Gaussian kernels [MZ09, Kou09], the Karhunen-Loève expansion [DEH08, DEM09, MN09, MEMDG10], Fourier basis functions or wavelets [BJMS15, DS17]. Wavelets are hierarchically structured basis functions that are localised in both frequency and spatial domains and are promising candidates to represent permeability fields with local variations [Mal08, SS09, DHS12].

The Bayesian approach to inverse problems provides a posterior density that must be explored numerically. The Markov chain Monte Carlo (MCMC) methods described in Section 2.1.3 can be used, however, the Markov chain is expensive to construct due to the cost of evaluating the likelihood function [OCR97]. A number of reduced-order models have been proposed [Jin08, MZ09] to reduce the computational cost. A two-stage MCMC algorithm is introduced in [EDGG⁺05, EHL06], in which where the acceptance rate of the proposals on the fine-scale model is improved by using a coarse-scale model. Sequential Monte Carlo (SMC) has been used to avoid the construction of a stationary Markov chain and to address concerns with multimodality [Cho02, DMDJ06, Kou09, WZ11, KBJ14, BJMS15]. SMC methods have recently been investigated with multilevel Monte Carlo [CGST11, BJL⁺17]. The nested sampling algorithm has also been used for permeability estimation [Ski06, EHW14].

It is common practice to fix either the length scale, the number of length scales or the number of terms in a representation of the spatially varying parameter. An overly parsimonious model will fail to capture the most basic features, but an overly complex model will seldom match the data. This is a model selection problem that can be approached using the model comparison techniques in Section 1.3. One possibility is to average over a family of candidate models using the reversible jump Metropolis-Hastings algorithm [Gre95], for example, as done for permeability estimation using level sets in [MEMDG10]. In practice, simulations over the entire model space are difficult to implement efficiently. Marginal likelihoods can be estimated with the nested sampling algorithm, however, each marginal likelihood must be estimated independently and the constrained-sampling step can be difficult to implement efficiently [Ski06, EHW14]. There is a potential to use SMC techniques for adaptive model selection [ZJA16].

Permeability fields are multiscale in nature. Some authors have advocated the use of multiscale methods for statistical inversion. According to [FL07], there are three cases when multiscale methods are particularly beneficial in the context of statistical inversion. First, when the parameter of interest is multiscale and better modelled this way. For example, multiscale Markov random field and sparse grid models have been developed to represent permeability fields in [LHB⁺02a] and [WZ11], respectively. Second, when there are algorithmic advantages for doing so. In some cases it is relatively easy to fit a coarse model but much more challenging to work with a finer model. This may be due to the raised computational cost or additional complexity associated with a finer model. For example, wavelets have been used to stabilise

inverse problems in [Rie97]. A multi-resolution approach is taken in [Kou09] where less accurate forward solvers are used to accelerate the inference procedure. Third, when the data is available at different scales and a multiscale approach provides a structured way of integrating the available information into the solution of the inverse problem. Another class of multiscale methods concerns the design of more efficient forward solvers, for example, the multiscale finite element method is designed to capture the coarse-scale behaviour of the forward model without resolving the fine-scale details [EH07, EH09].

In this chapter, Bayesian model selection techniques are investigated in the context of multiscale inverse problems using SMC techniques. The unknown spatially varying parameter is represented via a multi-resolution analysis comprising of hierarchically structured wavelet basis functions. The SMC framework is used to efficiently explore a sequence of posterior distributions, using information gathered from previous inferences. Each model is refined adaptively, and is compared to the previously proposed model using Bayes' factors. In this way, multiscale statistical inversion is addressed without presuming the number of scales. The proposed methodology is demonstrated with a computational study on permeability estimation using limited and noisy pressure data.

3.2 Mathematical modelling

It is assumed that the phenomenon of interest is described by a set of partial differential equations that depends on a spatially varying parameter as in (1.1). More specifically, it is assumed that the spatial domain $\mathcal{X} \subseteq \mathbb{R}^d$ is a bounded open set with a smooth boundary $\partial\mathcal{X}$, and that the normed spaces $\mathcal{K}$ and $\mathcal{Y} \subseteq \mathbb{R}$ correspond to the state space for the spatially varying parameter u and the observations y_i , for $i = 1, \dots, n$, respectively. It is assumed that (1.1) has a unique solution that can be expressed via the forward map $G : \mathcal{X} \times \mathcal{K} \rightarrow \mathcal{Y}$, and that the observations are made at locations $x_i \in \mathcal{X}$ and are each given by

$$y_i = G(x_i; u) + \epsilon_i. \quad (3.1)$$

Here, $\epsilon_i \in \mathcal{Y}$ represents the composite error, which is comprised of observation and model error at the location x_i . The primary interest is to recover the unknown parameter u from the observations $y = \{y_i\}_{i=1}^n$. Several processes fall under this setting in the physical sciences and engineering; whilst physical interpretations may differ, the underlying principles remain the

same.

3.2.1 Forward model

This chapter is motivated by a problem of central importance to hydrology: the task of estimating the permeability field from indirect flow data. An understanding of subsurface flow is required to design effective procedures to remove a soil contaminant and for efficient oil production. Subsurface flow is characterised by a heterogeneous permeability field, which is a function that describes how easily fluid flows through the aquifer at a given location. Whilst permeability can be measured directly by taking a core sample to a lab, it is impractical and prone to large measurement errors as the conditions of the lab differ greatly to those of the ground. Instead, permeability can be estimated from limited flow data, for example, using tracer break-through times or pressure head measurements.

The behaviour of groundwater flow through a heterogeneous medium is understood by an empirical relation known as Darcy's law. It may be combined with a relation describing mass conservation to obtain the following elliptic PDE:

$$v(x) = -\kappa(x)\nabla p(x), \quad x \in \mathcal{X}, \quad (3.2)$$

$$\nabla \cdot v(x) = f(x), \quad x \in \mathcal{X}, \quad (3.3)$$

$$v(x) \cdot \hat{n}(x) = 0, \quad x \in \partial\mathcal{X}. \quad (3.4)$$

Here, $p : \mathcal{X} \rightarrow \mathbb{R}$ is the pressure head, $v : \mathcal{X} \rightarrow \mathbb{R}^2$ is the fluid velocity, $\kappa : \mathcal{X} \rightarrow \mathbb{R}_{>0}$ is the permeability field and $\hat{n}$ denotes the unit normal to the boundary $\partial\mathcal{X}$. No flux boundary conditions have been used to confine the flow to $\mathcal{X}$, although in some cases other boundary conditions may be more appropriate and can be used instead. It is assumed that $\kappa \in L^\infty(\mathcal{X})$ satisfies the mathematically necessary and physically meaningful positivity constraint:

$$\text{ess inf}_{x \in \mathcal{X}} \kappa(x) \geq \kappa_{\min} > 0. \quad (3.5)$$

Lastly, $f : \mathcal{X} \rightarrow \mathbb{R}$ characterises the source/sink term and is assumed to be a continuous linear functional. A typical choice of source/sink function is

$$f(x) = \sum_{i=1}^S c_i \delta_{x_i}(x), \quad (3.6)$$

where the $c_i \in \mathbb{R}$ are the source/sink strengths, the $x_i \in \mathcal{X}$ are the source/sink locations and $\delta_{x_0}(x) = \delta(x - x_0)$ is the Dirac delta function. Strictly speaking, the mathematical formulation of the problem requires the Dirac delta functions to be mollified. This can be carried out in a way that makes little difference to the numerical approximation of the solution.

3.2.2 Existence, uniqueness and boundedness

It is important to check that the forward map G is well-defined so that it can be used as part of the likelihood function. This is covered by standard PDE theory [Eva98] and the key arguments are outlined here [DS17]. Two remarks are in order. First, solutions to the PDE in (3.2)-(3.4) are only defined up to an additive constant; if p is a solution, then so is $p + c$. An additional constraint can be placed on a selected point x^* , so that $p(x^*) = p^*$. For mathematical convenience, it suffices to consider the mean-zero solution:

$$\bar{p} = \frac{1}{|\mathcal{X}|} \int_{\mathcal{X}} p(x) dx = 0, \quad (3.7)$$

so that the Poincaré inequality holds. This means that for any $p \in L^2(\mathcal{X})$ with $\bar{p} = 0$, there exists a constant $C_p > 0$ such that

$$\|p\| \leq C_p \|\nabla p\|. \quad (3.8)$$

Second, a solution can only exist if the compatibility condition is satisfied

$$\int_{\mathcal{X}} f dx = - \int_{\mathcal{X}} \nabla \cdot v dx = - \int_{\partial\mathcal{X}} v \cdot \hat{n} ds = 0. \quad (3.9)$$

This is a mathematical consequence of the divergence theorem, but also consistent with the physical intuition behind no flux boundary conditions. In other words, any inflows from the source terms must be matched by outflows from the sink terms.

To show existence and uniqueness of a solution, one can apply the Lax-Milgram theorem to the weak problem, for example, Section 6.2 in [Eva98]. The Sobolev space $H_0^1(\mathcal{X})$ is defined as the subspace of $L^2(\mathcal{X})$ whose functions have first-order weak derivatives with a finite L^2 norm and satisfy the Poincaré inequality in (3.8). This space admits the Sobolev norm

$$\|\cdot\|_{H_0^1} = (\|\cdot\|^2 + \|\nabla \cdot\|^2)^{1/2}. \quad (3.10)$$

Let $H^{-1}(\mathcal{X})$ be the dual space of $H_0^1(\mathcal{X})$ – this is the space of continuous linear functionals on $\mathcal{X}$. The weak problem is then as follows: find $p \in H_0^1(\mathcal{X})$ such that

$$\langle \kappa \nabla p, \nabla p' \rangle = \langle f, p' \rangle \text{ for all } p' \in H_0^1(\mathcal{X}). \quad (3.11)$$

The weak problem is obtained by integrating the left-hand side of (3.3) against a test function $p' \in H_0^1(\mathcal{X})$ and applying the first Green's identity. Likewise, the right-hand side of (3.11) is obtained by integrating the right-hand side of (3.3) against the same test function.

The Lax-Milgram theorem then guarantees the existence of a unique solution for any $f \in H^{-1}(\mathcal{X})$ provided that the bilinear form satisfies two properties:

- (boundedness) there exists $C > 0$ such that $|\langle \kappa \nabla p, \nabla p' \rangle| \leq C \|p\|_{H_0^1(\mathcal{X})} \|p'\|_{H_0^1(\mathcal{X})}$ for all $p, p' \in H_0^1(\mathcal{X})$; and
- (coercivity) there exists $c > 0$ such that $|\langle \kappa \nabla p, \nabla p \rangle| \geq c \|p\|_{H_0^1(\mathcal{X})}^2$ for all $p \in H_0^1(\mathcal{X})$.

These conditions require that $\kappa \in L^\infty(\mathcal{X})$ with the positivity condition in (3.5). Then there exist $\kappa_{\min}$ and $\kappa_{\max}$ such that

$$0 < \kappa_{\min} \leq \kappa(x) \leq \kappa_{\max} < \infty, \quad (3.12)$$

almost everywhere. Then the Cauchy-Schwartz inequality¹ implies boundedness:

$$|\langle \kappa \nabla p, \nabla p' \rangle| \leq \kappa_{\max} |\langle \nabla p, \nabla p' \rangle| \leq \kappa_{\max} \|\nabla p\| \|\nabla p'\| \leq \kappa_{\max} \|p\|_{H_0^1(\mathcal{X})} \|p'\|_{H_0^1(\mathcal{X})}, \quad (3.13)$$

and the Poincaré inequality implies coercivity:

$$|\langle \kappa \nabla p, \nabla p \rangle| \geq \kappa_{\min} |\langle \nabla p, \nabla p \rangle| \geq \frac{\kappa_{\min}}{2C_p^2} (\|p\|^2 + C_p^2 \|\nabla p\|^2) \geq \frac{\kappa_{\min}}{2C_p^2} \min \{1, C_p^2\} \|p\|_{H_0^1(\mathcal{X})}^2. \quad (3.14)$$

Moreover, the unique solution is bounded by the norm of the right-hand side: there exists $\tilde{C} > 0$ such that

$$\|p\|_{H_0^1(\mathcal{X})} \leq \frac{\tilde{C}}{\kappa_{\min}} \|f\|_{H^{-1}(\mathcal{X})}. \quad (3.15)$$

Collectively, these results mean that a forward map of the solution $p(x)$ is well-defined and can be used as part of the likelihood function [DS17].

¹The Cauchy-Schwartz inequality is: $|\langle u, v \rangle| \leq \|u\| \|v\|$.

3.2.3 Spatially varying parameter

The unknown permeability field is modelled by $\kappa(x) = e^{u(x)}$, where $u : \mathcal{X} \rightarrow \mathbb{R}$ is represented by the finite series

$$u(x) = c_0 + \sum_{j=1}^N c_j \psi_j(x). \quad (3.16)$$

Here, $\{\psi_j\}_{j=1}^N$ are a set of bounded basis functions and $\{c_j\}_{j=0}^N$ are random variables with finite variance so that u is a stochastic process whose realisations are almost surely in $L^\infty(\mathcal{X})$. There are many possible choices for the basis functions, for example, radial basis functions, Fourier basis functions, wavelets to name a few.

3.2.4 Numerical approximation

The likelihood model that will be used will assume access to an approximate solution of the forward model, which is known as a forward solver. For the particular problem described here, the most popular method is the finite element method. In particular, the mixed finite element is used to jointly estimate the pressure and velocity fields. Although not required here, one is typically interested in estimating the velocity field for simulation purposes and numerically differentiating the pressure field leads to a loss of accuracy.

The mixed finite element method requires approximation spaces for pressure and velocity, and then solving the weak formulation in those spaces. It is well known that the discrete mixed problem can be an unstable approximation unless the approximation spaces are appropriately chosen. One suitable choice is the Raviart-Thomas space for the velocity solution and discontinuous elements for the pressure space [RT77, BHMM05]. To solve the approximate weak formulation, the mass matrix and right hand side are constructed via numerical quadrature and the unknown coefficients of the basis functions forming approximation spaces are obtained by solving the resulting linear system of equations. Solving a large system of linear equations, as required for fine meshes and high-order approximation spaces, makes each call to the forward solver an expensive operation.

3.3 Multiscale modelling

The reasons for adopting a multiscale approach were introduced in Section 3.1: permeability fields are multiscale processes that are naturally modelled this way; and there can be computational advantages for doing so. According to [FL07], there are three categories of multiscale methods to consider: multiscale decomposition methods that use hierarchically structured basis functions, explicit methods that model inter-scale dependencies and implicit methods that link independently defined models via numerical computations. The approach taken here is to specify a family of models using hierarchically structured basis functions, and then to implicitly link the models – from coarse to fine scales – using the SMC framework described in Section 2.3. Then by performing the inferences sequentially, each model can be evaluated and the next candidate model can be chosen adaptively as part of the fitting procedure [Mac03a, ZJA16].

The rest of this section describes a multiscale representation of the unknown spatially varying parameter. Tent map basis functions, wavelets and radial basis functions are popular choices of multiscale basis functions [FL07, WZ11]. Here, wavelet basis functions are used, which have been extensively studied for image analysis and signal processing [Mal89, Mal08]. There are two reasons for their widespread use. First, wavelet basis functions form a multi-resolution analysis and possess a natural hierarchical structure. Each wavelet basis function has a number of children basis functions that represent a finer level of detail. Second, wavelets are localised in both space and frequency. Wavelets are therefore good candidates to represent permeability fields that are long memory processes with local variations. Sharp variations can be more difficult to represent with global basis functions that are only localised in frequency, for example, Fourier basis functions, due to Gibbs phenomenon [GS97].

3.3.1 Multi-resolution analysis

This section describes a multi-resolution analysis and wavelet basis functions. The one-dimensional case is considered first, and is followed by the two-dimensional case.

Wavelets can be used to analyse functions that belong to $L^2(\mathbb{R})$, which is the space of functions with finite energy. A wavelet is a function $\psi \in L^2(\mathbb{R})$ that has zero mean $\int_{-\infty}^{\infty} \psi(x)dx = 0$ and has a unit square norm $\|\psi\|^2 = 1$. A wavelet function ψ can be used to generate an entire family

of wavelets by translation and scaling operations:

$$\psi_{j,n}(x) = 2^{-j/2} \psi(2^{-j}x - n), \quad (j, n) \in \mathbb{Z}^2. \quad (3.17)$$

If the family of wavelets is sufficiently rich, it is possible to create a basis of the real line:

$$L^2(\mathbb{R}) = \text{span}\{\psi_{j,n}\}_{(j,n) \in \mathbb{Z}^2}. \quad (3.18)$$

In (3.17), the scaling caused by a positive value of j is an expansion and the scaling caused by a negative value of j is a contraction.

In practice, wavelet coefficients are only considered up to a chosen value of j . In this case, to create a basis of the real line, it is necessary to consider additional basis functions. The additional basis functions are derived from a scaling function ϕ . Again, a family of scaling functions can be defined by translation and scaling operations:

$$\phi_{j,n}(x) = 2^{-j/2} \phi(2^{-j}x - n), \quad (j, n) \in \mathbb{Z}^2. \quad (3.19)$$

The approximation spaces V_j and detail spaces W_j are subspaces of $L^2(\mathbb{R})$ and are defined for the scale j . They are constructed from the scaling function and wavelet function, respectively. The approximation and detail spaces are defined by

$$V_j = \text{span}\{\phi_{j,n}\}_{n \in \mathbb{Z}}, \quad W_j = \text{span}\{\psi_{j,n}\}_{n \in \mathbb{Z}}. \quad (3.20)$$

The approximation spaces satisfy the nesting property

$$\{0\} \subset \cdots \subset V_1 \subset V_0 \subset V_{-1} \subset \cdots \subset L^2(\mathbb{R}), \quad (3.21)$$

and a finer approximation space is obtained by augmenting a coarser approximation space with the set of wavelets for that scale:

$$V_{j-1} = V_j \oplus W_j. \quad (3.22)$$

It is assumed that the coarsest scale occurs for $j = k$ and then (3.21)-(3.22) are applied to represent any $u \in L^2(\mathbb{R})$ by

$$u(x) = \sum_{n=-\infty}^{\infty} \langle u, \phi_{k,n} \rangle \phi_{k,n}(x) + \sum_{n=-\infty}^{\infty} \sum_{j=-\infty}^k \langle u, \psi_{n,j} \rangle \psi_{n,j}(x). \quad (3.23)$$

Following [Mal08], an orthonormal basis spanning $L^2(\mathbb{R}^2)$ is provided by two-dimensional approximation spaces and detail spaces, where $\otimes$ denotes the tensor product and $\oplus$ denotes the orthogonal sum. The spaces are defined by

$$V_j^2 = V_j \otimes V_j, \quad \text{and} \quad V_{j-1}^2 = V_j^2 \oplus W_j^2, \quad (3.24)$$

respectively. Combining these definitions and using (3.22) along with the distributivity of $\oplus$ over $\otimes$, it follows that

$$W_j^2 = (V_j \otimes W_j) \oplus (W_j \otimes V_j) \oplus (W_j \otimes W_j). \quad (3.25)$$

Therefore two-dimensional wavelets and scaling functions are obtained as follows:

$$\psi_0(x, y) = \phi(x)\psi(y), \quad (3.26)$$

$$\psi_1(x, y) = \psi(x)\phi(y), \quad (3.27)$$

$$\psi_2(x, y) = \psi(x)\psi(y), \quad (3.28)$$

$$\phi(x, y) = \phi(x)\phi(y). \quad (3.29)$$

Two-dimensional families of wavelets and scaling basis functions are defined in a similar way to the one-dimensional case by a series of translation and scaling operations

$$\phi_{j,n,m}(x, y) = 2^{-j}\phi(2^{-j}x - n, 2^{-j}y - m), \quad (3.30)$$

$$\psi_{d,j,n,m}(x, y) = 2^{-j}\psi_d(2^{-j}x - n, 2^{-j}y - m), \quad (3.31)$$

where $d \in \{0, 1, 2\}$ and $(j, n, m) \in \mathbb{Z}^3$. Here, translations occur in two directions, as indexed by n and m , and the index d is introduced to specify the wavelet direction. Any $u \in L^2(\mathbb{R}^2)$ can be represented by

$$\begin{aligned} u(x, y) = & \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \langle u, \phi_{k,n,m} \rangle \phi_{k,n,m}(x, y) \\ & + \sum_{j=-\infty}^k \sum_{d=0}^2 \sum_{n=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \langle u, \psi_{j,d,n,m} \rangle \psi_{j,d,n,m}(x, y). \end{aligned} \quad (3.32)$$

In [Dau92], it is shown that a necessary and sufficient condition to ensure a stable reconstruction is that wavelet coefficients must define a Riesz basis. This imposes finite energy constraints on the wavelet coefficients in that their infinite sum of squares must be finite.

3.3.2 Multiscale modelling of a spatially varying parameter

A multi-resolution parametrisation is provided for the spatially varying parameter. The spatially varying parameter is sampled over the unit square as opposed to the real plane. For simplicity it is assumed that wavelets are defined on the 2-dimensional unit torus $\mathbb{T}^2$. As the unit square is finite, there is a lower limit on the coarsest scale possible. The coarsest scale is specified to occur for $s = 0$, where $s = -j$, and has a single constant-valued scaling basis function. A spatially varying parameter $u \in L^2(\mathbb{T}^2)$ is represented by

$$u(x, y) = c_0 + \sum_{s=1}^S \sum_{d=0}^2 \sum_{n=0}^{2^{s-1}} \sum_{m=0}^{2^{s-1}} c_{s,d,n,m} \psi_{s,d,n,m}(x, y). \quad (3.33)$$

where S is the finest scale, c_0 is constant-valued and $c_{s,d,n,m}$ are the wavelet coefficients. The wavelet representation is illustrated in Fig. 3.1 for the corresponding one-dimensional case. It can be seen that each one-dimensional wavelet has two children wavelets that are more localised in space than its parent.

The wavelet representation being considered is periodic, which is uncharacteristic of the unknown spatially varying parameter. To remove the unwanted effects of periodicity, the reconstruction is carried out over an extended domain of twice the size. The spatially varying parameter is reconstructed using the (inverse) fast wavelet transform [Mal89, Mal08] in $O(n \ln n)$ operations, where n is the number of grid points. Not only is it a computationally efficient reconstruction, but many families of wavelets are not known in closed form and must be implemented in this way.

3.4 Bayesian inverse problem

The solution to the inverse problem is given by the posterior density $\pi(\theta|y)$, which is proportional to the product of the prior density $\pi(\theta)$ and the likelihood function $\pi(y|\theta)$:

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)}, \quad (3.34)$$

where $\pi(y)$ is the marginal likelihood. The specification of the prior distribution and likelihood function is discussed here.

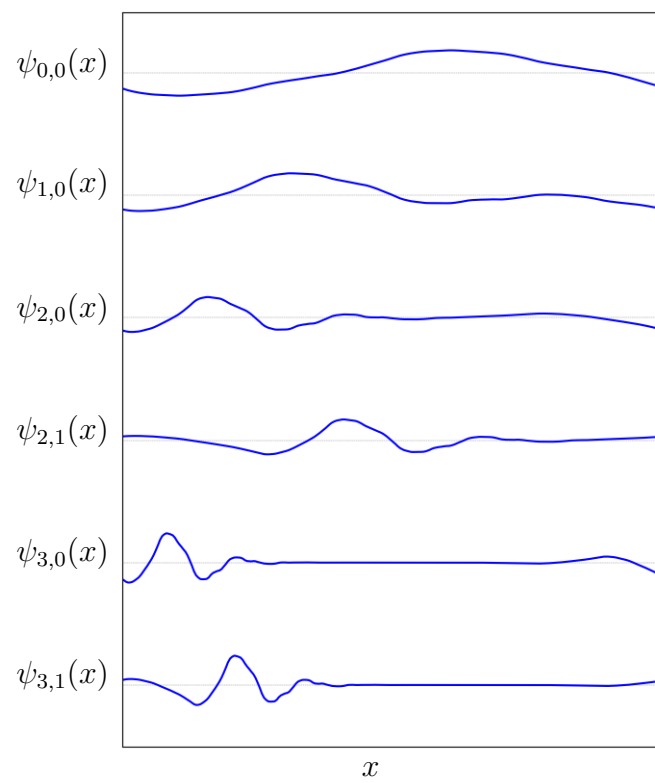


Figure 3.1: Illustration of one-dimensional Daubechies wavelets with 4 vanishing moments (the db8 wavelet). The wavelets have been rescaled to share a common amplitude. The hierarchical structure can be seen: $\psi_{2,0}$ and $\psi_{2,1}$ are children of $\psi_{1,0}$; and $\psi_{3,0}$ and $\psi_{3,1}$ are children of $\psi_{2,0}$.

3.4.1 Prior modelling

Following (3.33), the unknown variables are the wavelet coefficients $\{c_0\} \cap \{c_{s,d,n,m}\}$. To obtain a reasonable representation of the unknown field, the basis functions are first adjusted so that they share a common amplitude. A rate of decay a_s is then introduced for each scale s . The unknown wavelet coefficients are related to the model parameters $\{\theta_0\} \cap \{\theta_{s,d,n,m}\}$ by

$$c_0 = a_0\theta_0, \quad c_{s,d,n,m} = 2^{-s}a_s\theta_{s,d,n,m}. \quad (3.35)$$

When the unknown parameters are modelled as independent and identically distributed (iid) random variables, the wavelet coefficients for the coarser scales have more energy (on average) than the wavelet coefficients for the finer scales. The coarser scales therefore represent the more salient features. The following choice is used for the rate of decay:

$$a_s = \frac{a}{(s+1)^\alpha}, \quad \alpha > 1, \quad a > 0. \quad (3.36)$$

Here a is the amplitude that controls the marginal variance of the process and α is the rate of decay that controls the relative importance of the finer scales. The rate of decay allows the process to have a degree of smoothness. Without rescaling the wavelets, when the wavelet coefficients are iid random variables their infinite sum of squares would diverge. The wavelets at higher frequencies would overwhelm the wavelets at lower frequencies and the process would resemble white noise. Conversely, if the decay is too aggressive, then the process would be overly smooth. Although the infinite-dimensional case reaffirms the appropriateness of the wavelet decay, only finite dimensional representations are used in this chapter.

It is necessary to specify a suitable prior distribution for the unknown parameters. Exponential priors such as the Gaussian prior have been popular amongst geophysicists. Here, uniform priors are used, as done in [BJMS15]. Uniform priors have been adopted in the computational mathematics literature in the context of uncertainty quantification in [SS12, DS17]. The prior used is

$$\theta_0, \theta_{s,d,n,m} \sim U(-1, 1), \quad (3.37)$$

although the properties of the reconstructed field depend on the choices of a and α . As discussed in the opening sections, the parameters a and α must be carefully specified on a case-by-case basis as the measurement data typically contains little information about the smoothness of the

unknown.

Lastly, it remains to specify the family of wavelet; there are many to choose from, each with differing properties. Despite being the simplest and most well known, the Haar wavelet is not suitable for approximating smooth functions as it contains discontinuities [FL07]. On the other hand, the Daubechies wavelets are characterised by a maximal number of vanishing moments for a given support [Dau92]. The Daubechies wavelet with 4 vanishing moments, known as the db8 wavelet, is used here.

The prior modelling for the spatially varying parameter described here is applicable for a general class of inverse problems. It is acknowledged that domain-specific knowledge can be used to improve the performance of the proposed methodology. Such knowledge is not assumed here to avoid restricting the proposed methodology to specific examples only.

3.4.2 Likelihood specification

It is assumed that the mathematical model described in (1.1) can be approximated by the forward map $G(x; \theta)$ so that, after taking into account any uncertainty,

$$y_i = G(x_i; u) + \epsilon_i, \quad (3.38)$$

for $i = 1, \dots, N$. Here, each ϵ_i is a composite error representing model error and observation noise, and N is the number of observations which are made at locations $x_i \in \mathcal{X}$. As noted in [KO01], the model error and observation noise are not identifiable, meaning that several different values can be consistent with the data. For this reason, they are modelled together as a composite error. Each error is modelled as an iid Gaussian random variable:

$$\epsilon_i \sim \text{Normal}(0, \sigma^2). \quad (3.39)$$

As the variance of the composite error is difficult to specify a priori, a prior is used for the precision

$$\sigma^{-2} \sim \text{Gamma}(a, b). \quad (3.40)$$

This allows the model to be robust to varying levels of noise, which is particularly important as the magnitude of model error is dependent on the representation of the spatially-varying

parameter as well as the accuracy of the forward solver used. The additional parameter can be marginalised over, analytically, to obtain the following likelihood function:

$$\begin{aligned}
\pi(y|\theta) &= \int_0^\infty \pi(y|\theta, \sigma^2) \pi(\sigma^2) d\sigma^2, \\
&= \int_0^\infty \left(\frac{\tau}{2\pi} \right)^{K/2} \exp \left(-\frac{\tau}{2} \sum_{i=1}^K (y_i - G(x_i; u))^2 \right) \times \frac{b^a}{\Gamma(a)} \tau^{a-1} \exp(-b\tau) d\tau, \\
&= \left(\frac{1}{2\pi} \right)^{K/2} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + K/2)}{\left(b + \frac{1}{2} \sum_{i=1}^K (y_i - G(x_i; u))^2 \right)^{a+K/2}} \\
&\quad \times \underbrace{\int_0^\infty \text{Gamma} \left(\tau \middle| a + \frac{K}{2}, b + \frac{1}{2} \sum_{i=1}^K (y_i - G(x_i; u))^2 \right) d\tau}_{=1}, \\
&\propto \left(b + \frac{1}{2} \sum_{i=1}^K (y_i - G(x_i; u))^2 \right)^{-(a+K/2)}.
\end{aligned} \tag{3.41}$$

The hyperparameters a and b are given the uninformative choices of $a = b = 0.001$ [Gel06]. This is because in the limit $a, b \rightarrow 0$, the prior

$$\pi(\sigma^{-2}) \propto (\sigma^{-2})^{a-1} \exp(-b\sigma^{-2}), \tag{3.42}$$

is the Jeffreys prior $\pi(\sigma^{-2}) \propto \sigma^2$, which is the only prior that is invariant under a change of scale [GCS⁺14]. With this choice, one has a prior that is conjugate with the Gaussian likelihood that is also an approximation of a scale invariant prior.

3.5 Sequential Monte Carlo samplers

The posterior distribution in (3.34) is analytically intractable and quantities, such as credible intervals of the spatially-varying parameter, must be estimated numerically. It is known that MCMC methods do not always perform well for statistical inversion, especially when the target density is multimodal and the likelihood function is expensive to evaluate. For this reason, SMC methods have been suggested to explore the posterior distribution [Kou09, WZ11, KBJ14, BJMS15]. A new multiscale approach is proposed to sample from the more complex posterior distributions.

3.5.1 Sequential Monte Carlo

First, a SMC sampler is described to explore the posterior distribution in (3.34). The SMC sampler is based on the sequential importance resampling techniques outlined in Section 2.3.2. Here, it is assumed that the posterior distribution is defined for the wavelet representation in (3.33) with wavelets defined up to scale S . A particle approximation of $\pi(\theta|y)$, denoted $\{\theta^{(i)}, W^{(i)}\}_{i=1}^N$, is obtained such that a strong law of large numbers holds:

$$\int \phi(\theta) \pi(\theta|y) d\theta = \lim_{N \rightarrow \infty} \sum_{i=1}^N W^{(i)} \phi(\theta^{(i)}), \quad (3.43)$$

for any integrable test function ϕ .

A sequence of bridging densities $\pi_0(\theta), \dots, \pi_M(\theta)$ is introduced to smooth the transition from the prior density $\pi(\theta)$, with $\pi_0(\theta) = \pi(\theta)$, to the posterior density $\pi(\theta|y)$, with $\pi_M(\theta) = \pi(\theta|y)$. The following bridging densities are used

$$\pi_j(\theta) = \frac{1}{Z_j} \pi(\theta|y)^{\gamma_j} p(\theta), \quad (3.44)$$

for $j = 0, 1, \dots, M$, where $0 = \gamma_0 < \gamma_1 < \dots < \gamma_M = 1$ is a sequence of tempering parameters. Each normalising constant is given by

$$Z_j = \int \pi(\theta|y)^{\gamma_j} p(\theta) d\theta. \quad (3.45)$$

The proceeding implementation uses uniformly spaced tempering parameters, as done in [WZ11]. It is also possible to determine the sequence of tempering parameters adaptively to reduce the number of bridging distributions required [Kou09, KBJ14, BJMS15]. The chosen sequence of distributions is natural for a Bayesian analysis as the prior distribution is easy to sample from and covers the support of the posterior distribution. The marginal likelihood $\pi(y)$ can be estimated using (2.61), which can be used for model comparison.

The SMC sampler obtains a particle approximation for each bridging density π_j . Each particle approximation is denoted by $\{\theta_j^{(i)}, W_j^{(i)}\}_{i=1}^N$, respectively. Here, $W_j^{(i)}$ and $w_j^{(i)}$ are the normalised and unnormalised importance weights, respectively. A particle approximation for $j = 0$ is obtained by drawing the particle locations from the prior distribution, which is easy to sample from, and by setting $W_0^{(i)} = 1/N$ and $w_0^{(j)} = 1$. Then for $j = 1, \dots, M$, the particle

approximation of π_{j-1} is updated to obtain a particle approximation of π_j . Each update comprises of a reweighting, resampling and rejuvenation step as follows:

- Reweighting. The unnormalised importance weights are updated by

$$w_j^{(i)} = w_{j-1}^{(i)} \frac{\pi_j(\theta_{j-1}^{(i)})}{\pi_{j-1}(\theta_{j-1}^{(i)})}, \quad (3.46)$$

for $i = 1, \dots, N$. Each normalised importance weight is also updated by

$$W_j^{(i)} = w_j^{(i)} / \sum_{k=1}^N w_j^{(k)}. \quad (3.47)$$

- Resampling. The particle degeneracy is measured by the variability of the importance weights, for which it is common practice to calculate the so-called effective sample size, given by

$$\text{ESS}_j = \left(\sum_{i=1}^N (W_j^{(i)})^2 \right)^{-1}, \quad (3.48)$$

which takes a value between 1 and N . The particles are resampled via multinomial resampling whenever ESS_j exceeds the cut-off defined by $\text{ESS}_{\min} = \xi N$ for some $\xi \leq 1$ [DMDJ06]. The new unnormalised importance weights are set to the average of the previous ones. The resampling step prevents a large amount of computational resource being spent on particles with little weight.

- Rejuvenation. The rejuvenation step is performed with a Metropolis-Hastings transition kernel K_j that leaves the particles π_j invariant. If the particle locations are marginally distributed according to η_j , then after rejuvenation the particles are marginally distributed according to

$$\eta_j(\theta_j) = \int \eta_{j-1}(\theta_j) K_j(\theta_{j-1}, \theta_j) d\theta_{j-1}. \quad (3.49)$$

Here, an isotropic random walk with step size σ_j is used to propose a new locations for each particle:

$$\tilde{\theta}_j^{(i)} \sim N(\theta_j^{(i)}, \sigma_j^2 I). \quad (3.50)$$

The newly proposed locations are accepted with probability $\min\{1, \pi_j(\tilde{\theta}_j^{(i)})/\pi_j(\theta_j^{(i)})\}$. This choice of proposal allows for local updates such that the covariance of the proposal is consistent with the prior modelling. The transition kernel K_j and therefore the step

size parameter σ_j can be determined after applying the previous transition kernels K_{j-1} . The step size parameter for the next iteration σ_{j+1} can therefore be determined from the current acceptance rate AR_j , by

$$\sigma_{j+1} = \begin{cases} 2\sigma_j & \text{if } \text{AR}_j < 0.30, \\ \frac{1}{2}\sigma_j & \text{if } \text{AR}_j < 0.15, \\ \sigma_j & \text{otherwise.} \end{cases} \quad (3.51)$$

This scheme is similar to the approach taken in [BJMS15], which aims to achieve an acceptance rate in the neighbourhood of 0.2.

Lastly, the normalizing constant $Z_M = \pi(y)$ can be estimated by averaging over the importance weights:

$$Z_M \approx \frac{1}{N} \sum_{i=1}^N w_M^{(i)}. \quad (3.52)$$

Therefore, model comparison can be performed alongside the inference procedure.

The SMC algorithm is directly parallelisable. Communication between parallel processes is only required when calculating the effective sample size and when resampling the particles. This is a major benefit for statistical inversion as the expensive likelihood evaluations can be carried out in parallel.

3.5.2 Multiscale approaches

A new multiscale approach is proposed in this section. The motivating aspects are:

- It is easier to fit a coarse model than a fine model, and there is scope for multiscale methods to improve the inference procedure. First, coarsening has the effect of smoothing the unknown and therefore ameliorating the multimodality; there are fewer modes for the sampler to visit. Second, when using hierarchical basis functions, a coarse model has a lower-dimensionality and takes fewer steps to explore via a random walk proposal. This is because a larger step size can be used to achieve a desired acceptance rate. Importance sampling has been considered in [FL07], however, it is noted that importance sampling will not perform well unless the likelihood function is used by the proposal distribution.

One possibility is to use SMC, which have been used for other multiscale approaches in [Kou09, WZ11].

- The number of scales in the multiscale representation is unknown. Whilst hyperparameters akin to the amplitude and decay parameters in (3.36) are usually specified as part of the prior modelling, the hierarchical representation of the unknown spatially varying parameter can be specified with varying degrees of complexity. As marginal likelihoods can be estimated with SMC, it is investigated whether the model complexity can be determined from data when moving from coarse to fine models. SMC also provides an opportunity for adaptive refinement [WZ11].

The specifics of the new methodology are described in the subsections below, followed by a summary of the proposed inference procedure.

A multiscale approach

From (3.33), one can define a family of models $\mathcal{M}_1, \dots, \mathcal{M}_{S_{\max}}$ such that each model $\mathcal{M}_S$ only contains wavelet basis functions up to the finest scale $s = S$. If Λ_S denotes the set of multi-indices to be included in the model $\mathcal{M}_S$, so that the wavelet representation in (3.33) for $\mathcal{M}_S$ can be written as

$$u(x, y) = c_0 + \sum_{i \in \Lambda_S} c_i \psi_i(x, y), \quad (3.53)$$

then the models are chosen such that $\Lambda_0 \subseteq \Lambda_1 \subseteq \dots \subseteq \Lambda_{S_{\max}}$. Revisiting (3.34), the probability density functions are now conditioned on the choice of model $\mathcal{M}_S$:

$$\pi(\theta_S | y, \mathcal{M}_S) = \frac{\pi(y | \theta_S, \mathcal{M}_S) \pi(\theta_S | \mathcal{M}_S)}{\pi(y | \mathcal{M}_S)}, \quad (3.54)$$

where θ_S denotes the N_S model parameters for $\mathcal{M}_S$. The likelihood $\pi(y | \theta_S, \mathcal{M}_S)$ is given by (3.41), as before, but the spatially varying parameter is reconstructed using (3.53). The prior $\pi(\theta_S | \mathcal{M}_S)$ can be written as

$$\pi(\theta_S | \mathcal{M}_S) \propto \mathbb{I}_{[-1,1]^{N_S}}(\theta_S). \quad (3.55)$$

A model $\mathcal{M}_S$ can be refined by adding up to $3 \times 2^{S+1} \times 2^{S+1}$ wavelets at the scale $S+1$ to obtain $\mathcal{M}_{S+1}$, as can be seen in Fig. 3.2. Let $\theta_{S'}$ denote the $N_{S'} = N_{S+1} - N_S$ additional parameters

needed to refine $\mathcal{M}_S$, so that $\theta_{S+1} = \{\theta_S, \theta_{S'}\}$ are the N_{S+1} model parameters for $\mathcal{M}_{S+1}$. The prior $\pi(\theta_{S+1}|\mathcal{M}_{S+1})$ can be written in terms of $\pi(\theta_S|\mathcal{M}_S)$ by

$$\pi(\theta_{S+1}|\mathcal{M}_{S+1}) \propto \pi(\theta_S|\mathcal{M}_S)\pi(\theta'_S|\mathcal{M}_{S'}), \quad (3.56)$$

for $S = 1, \dots, S_{\max} - 1$. Here, $\pi(\theta'_S|\mathcal{M}_{S'}) = \mathbb{I}_{[-1,1]^{N_{S'}}}(\theta')$ is the additional term required for the additional parameters.

Bridging between resolutions

A particle approximation for each $\pi(\theta_S|y, \mathcal{M}_S)$ can be obtained using the SMC sampler described in the previous section. This approach can be inefficient for a larger number of scales S due to the complexity of the posterior. Instead, a multiscale strategy is adopted.

First, the SMC sampler described in the previous section is used to sample from the posterior for $S = 0$. The following sequence of bridging densities is used:

$$\pi_j(\theta_0) = \frac{1}{Z_{0,j}} \pi(\theta_0|y, \mathcal{M}_0)^{\gamma_j} \pi(\theta_0|\mathcal{M}_0)^{1-\gamma_j}. \quad (3.57)$$

where $0 = \gamma_{0,0} < \gamma_{0,1} < \dots < \gamma_{0,M_0} = 1$. Then for $S = 0, \dots, S_{\max} - 1$, the particle approximation of $\pi(\theta_S|y, \mathcal{M}_S)$ is updated to obtain a particle approximation of $\pi(\theta_{S+1}|y, \mathcal{M}_{S+1})$. The following two updates are performed. First, the missing part is drawn (augmentation). Second, a sequence of $M_S + 1$ bridging densities is used to update the particle approximation (evolution).

- Augmentation. First, the missing parameters $\theta_{S'}$ are drawn from a (suitable) proposal distribution ρ_S . The particles are reweighted to obtain a particle approximation of

$$\pi_0(\theta_{S+1}) \propto \pi(\theta_S|y, \mathcal{M}_S)\pi(\theta'_S|\mathcal{M}_{S'}), \quad (3.58)$$

denoted $\{\theta_{S+1,0}^{(i)}, W_{S+1,0}^{(i)}\}_{i=0}^N$ with unnormalised importance weights $\{w_{S+1,0}^{(i)}\}_{i=0}^N$. With reference to (2.56), the new importance weights are given by

$$w_{S+1,0}^{(i)} = w_{S,M_S}^{(i)} \frac{\pi(\theta'_S|\mathcal{M}_{S'})}{\rho_S(\theta_S^{(i)})}. \quad (3.59)$$

By drawing the missing parameters from their prior distribution, the importance weights are left unchanged by this step. A resampling step is performed immediately before

this step to ensure that the particles are equally weighted when drawing the missing parameters.

- Evolution. The following sequence of bridging densities are used to move from $\pi_0(\theta_{S+1})$ to $\pi(\theta_{S+1}|y, \mathcal{M}_{S+1})^{\gamma_{S,j}}$:

$$\pi_j(\theta_{S+1}) = \frac{1}{Z_{S,j}} \pi(\theta_{S+1}|y, \mathcal{M}_{S+1})^{\gamma_{S,j}} \pi_0(\theta_{S+1})^{1-\gamma_{S,j}}, \quad (3.60)$$

where $0 = \gamma_{S+1,0} < \gamma_{S+1,1} < \dots < \gamma_{S+1,M_{S+1}} = 1$. The reweighting, resampling and rejuvenation steps are performed as in Section 3.5.1 to obtain the particle approximations $\{\theta_{S+1,j}^{(i)}, W_{S+1,j}^{(i)}\}_{i=0}^N$ and unnormalised importance weights $\{w_{S+1,j}^{(i)}\}_{i=0}^N$, for $j = 0, \dots, M_{S+1}$. The particle approximation with $j = M_{S+1}$ is the required particle approximation of $\pi(\theta_{S+1}|y, \mathcal{M}_{S+1})$.

Model comparison

A multi-resolution representation of the unknown spatially varying parameter has been proposed, however, the finest scale S needs to be determined. One possibility is to marginalise over a set of candidate models by treating the model as a parameter to be inferred. As already noted in [WZ11], model averaging can be achieved with Metropolis-Hastings methods using the reversible jump algorithm [Gre95], however, it can be very difficult to propose two significantly different models whilst maintaining a reasonable acceptance ratio. Instead, the approach taken here is to compare a family of competing models by their Bayes factors using the methodology described in Section 1.3.4.

For a family of competing models $\mathcal{M}_1, \dots, \mathcal{M}_{S_{\max}}$, Bayes theorem gives

$$p(\mathcal{M}_S|y) \propto p(y|\mathcal{M}_S)p(\mathcal{M}_S). \quad (3.61)$$

With the choice of a uniform prior $p(\mathcal{M}_S) = 1/S_{\max}$, the family of models can be compared by their Bayes factors

$$\text{BF}_{S_2, S_1} = \frac{\pi(y|\mathcal{M}_{S_2})}{\pi(y|\mathcal{M}_{S_1})} = \frac{Z_{S_2, M_{S_2}}}{Z_{S_1, M_{S_1}}}. \quad (3.62)$$

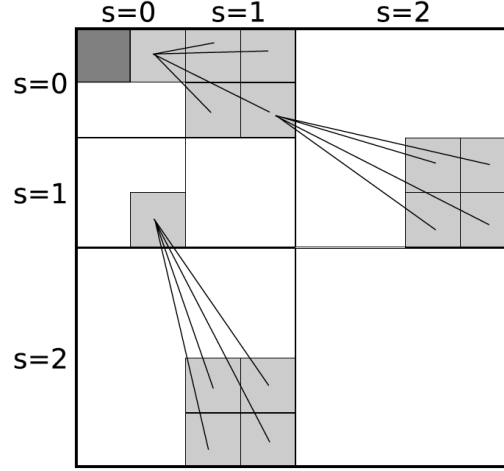


Figure 3.2: The wavelet coefficients of a two-dimensional periodic function can be represented by three quadtrees, for horizontal, vertical and diagonal directions, respectively. The first coefficient at scale $s = 0$ (shaded dark grey) is for the constant scaling basis function. The remaining three coefficients at scale $s = 0$ are the root nodes for each of the quad-trees. Selected coefficients have been shaded light grey to illustrate the quad trees; each wavelet has four children. At each scale, the top right block corresponds to horizontal wavelets ψ_0 , the bottom left block corresponds to vertical wavelets ψ_1 and the bottom right block corresponds to the diagonal wavelets ψ_2 .

The Bayes factors can be estimated using the output of the multiscale SMC sampler:

$$\text{BF}_{S_2, S_1} \approx \frac{\sum_{i=1}^N w_{S_2, M_{S_2}}^{(i)}}{\sum_{i=1}^N w_{S_1, M_{S_1}}^{(i)}}. \quad (3.63)$$

A Bayes factor of $\text{BF}_{S, S+1} > 1$ indicates that a refinement of $\mathcal{M}_S$ has provided an improved model $\mathcal{M}_{S+1}$. Conversely, a Bayes factor of $\text{BF}_{S, S+1} < 1$ indicates that the previous model is preferred.

Adaptive refinement and termination criterion

A wavelet representation can be refined by adding more wavelets at the next scale, following the nesting property in (3.21). Each wavelet at scale $s + 1$ is a local correction to the representation at the coarser scale s . If a wavelet does not make a contribution at scale s , it is unlikely that its children wavelets will make a contribution at scale $s + 1$. The intra-scale dependencies are represented by a quad-tree structure for each direction of wavelet in Fig. 3.2. It is possible to suggest a more parsimonious refinement instead of adding all possible wavelets to the model.

The refinement strategy is to identify ‘significant’ wavelets at the finest scale number $s = S$, and to add their children to the model. As the Bayesian framework provides a posterior distribution

over each wavelet coefficient, it is necessary to use summary statistics to determine which wavelets will have their children added to the model. The posterior mean is used here, as done in [WZ11], although other summary statistics are possible. A wavelet coefficient $c_{S,d,n,m}$ for $\mathcal{M}_S$ is ‘significant’ if the absolute of the posterior mean is in excess of a threshold $\epsilon \geq 0$:

$$\left| E[c_{S,d,n,m}|y, \mathcal{M}_S] \right| \geq \epsilon. \quad (3.64)$$

The threshold ϵ is a tuning parameter to be determined. The implementation here uses $\epsilon = 0.05$. When ϵ is small, all wavelet coefficients are added to the model until refinement terminates, or when $S_{\max}$ is reached. When ϵ is large, refinement will not take place and the procedure will terminate early.

By using the posterior mean for the refinement criterion, it is ensured that wavelets are only added to the model when the corresponding posterior marginal is not concentrated around zero and differs from its prior marginal. When the posterior marginal is similar to its prior marginal, the likelihood is insensitive to that wavelet coefficient and this is likely to be the case for the corresponding children wavelet coefficients. When the wavelet coefficient is concentrated around zero, no correction is required from the corresponding wavelet and it is likely that no correction is required from its children wavelets. Another possibility is to use the posterior variance for the refinement criterion, however, this would result in unwanted refinement for the wavelet coefficients concentrated around zero.

Wavelets are added to the model until a maximum scale $S_{\max}$ is reached, or until a termination criterion is satisfied. A suitable strategy is to move from one model $\mathcal{M}_S$ to the next $\mathcal{M}_{S+1}$ until the log-Bayes factor $\ln \text{BF}_{S+1,S}$ is negative or beneath a small positive threshold. A decrease in the log-Bayes factor suggests that the additional model complexity does not, on average, provide an improved explanation of the data. In this case, the previous model $\mathcal{M}_S$ should be preferred. Whilst early termination of the algorithm is theoretically possible due to ‘scale separation’, it would be uncharacteristic of a permeability field to have a missing scale. This belief is reflected in the prior modelling where the variation per scale is assumed to decrease as the scale number increases.

3.5.3 Computational considerations

For the standard SMC algorithm in Section 3.5.1, the maximum number of forward solver evaluations required is equal to the number of bridging densities $M_S + 1$. When using the multiscale SMC sampler by moving from one model to the next, the bridging densities contains two likelihood terms and the maximum number of forward solver evaluations is doubled. In practice, the number of forward solver evaluations is fewer, as whenever a proposal is made outside of the finite domain, the sample is rejected without calling the forward solver. As the SMC algorithm is directly parallelisable, the forward solver calls can be performed in parallel.

A multiscale scheme is proposed in [WZ11], however, informative priors are specified from the posterior of a previously fitted model. Whilst this means that only a single forward solver call is required when bridging between two densities, the measurement data appears in both the prior and likelihood. It may also be possible to accelerate the inference procedure by sometimes solving the forward model on a coarser mesh, for example, as done in [Kou09]. In doing so, only the most expensive computations are required $M_S + 1$ times during the final run of SMC, as is the case for standard SMC. This idea is not investigated here, but may further reduce the cost described above.

3.5.4 Summary of proposed approach

The multiscale approach presented in the previous subsection is summarised by the following algorithm:

1. Initialize the scale number $S \leftarrow 0$, set of multi-indices $\Lambda_0 \leftarrow \{0, d, 0, 0\}_{d=0,1,2}$ for $\mathcal{M}_0$ and the refinement set $R \leftarrow \emptyset$.
2. Use the SMC sampler described in Section 3.5.1 to obtain a particle approximation of $\pi(\theta_0|y, \mathcal{M}_0)$.
3. For each wavelet at scale S that satisfies the refinement criterion in (3.64), add the multi-indices of its children to the refinement set R .
4. **repeat**:
 - (a) Define the next model $\mathcal{M}_{S+1}$ by adding indices to the next set of multi-indices $\Lambda_{S+1} \leftarrow \Lambda_S \cup R$, then empty the refinement set $R \leftarrow \emptyset$.
 - (b) Update the particle approximation of $\pi(\theta_S|y, \mathcal{M}_S)$ to obtain a particle approximation of $\pi(\theta_{S+1}|y, \mathcal{M}_{S+1})$, using the augmentation and evolution steps described in Section 3.5.2.
 - (c) For each wavelet at scale S that satisfies the refinement criterion in (3.64), add the multi-indices of its children to the refinement set R .
 - (d) Estimate the Bayes factor $\text{BF}_{S+1,S}$ using (3.63).
 - (e) Set $S \leftarrow S + 1$.

until $\text{BF}_{S,S-1} \leq 1$, $R = \emptyset$ or $S > S_{\max}$.

In summary, the idea is to adaptively specify a family of candidate models $\mathcal{M}_1, \dots, \mathcal{M}_{S_{\max}}$ where each model is a refinement of the previously proposed model. A particle approximation is obtained for each model, and each particle approximation is used to provide information about the next, which improves the inference procedure for the finer models that have more complex posteriors. Bayes factors are used to assess whether a newly proposed model provides a better explanation of the data than the previous, and the procedure terminates once the additional

complexity no longer provides an improved fit. This is an improvement on the termination criterion provided in [WZ11].

3.6 Numerical results

This section provides numerical results to demonstrate the proposed methodology. Summary statistics are presented for the standard SMC sampler and for the multiscale approach, which are estimated using (3.43).

3.6.1 Implementation details

The forward solver that approximates the solution to (3.2)-(3.4) uses the mixed finite element method. Here, zero-order Raviart-Thomas elements are used for the velocity solution [RT77] and zero-order discontinuous elements are used for the pressure solution [BHMM05]. These correspond to piecewise-linear and piecewise-constant function spaces, respectively. The basis functions are defined on a 32×32 quadrilateral mesh. The forward solver is implemented using the Dealii C++ library [BHK07]. The spatially varying parameter is represented on a 32×32 grid, which coincides with the discretisation used by the forward solver.

The software used in the experiments is implemented in C++ using a Linux environment. Parallel computation was used wherever possible using MPI routines from OpenMP libraries. The workload was distributed over 64 processes, which ran in parallel. As SMC is embarrassingly parallel, this meant that the runtime for 1,280 particles in parallel is approximately the runtime for 20 particles in serial. The computational bottleneck arises from the forward solver calls. In the implementation, all likelihood computations are cached for subsequent use.

Synthetic measurement data is generated using the forward solver run on a 32×32 quadrilateral mesh. Synthetic data is used so that inferences can be compared with the truth, whereas this is not possible using real datasets. The data is either generated for a 10×10 or 5×5 pressure sensor network that is equally spaced across the solution domain with additive noise. The noise added to each measurement is Gaussian with either 5% or 1% relative noise depending on the example. This means that the standard deviation of the noise is equal to the given percentage multiplied by the mean of the pressure observations.

The key parameter values used in this section are summarised in Table 3.1. For SMC, $M = 1000$ auxiliary distributions with a resampling threshold of $\text{ESS}_{\min} = 0.95N$ was used to avoid resampling the particles on consecutive iterations as determined by preliminary runs. The resampling threshold is conservative to avoid applying expensive transition kernels to particles with small weights. Lastly, $N = 1280$ particles were chosen so that calculations could be performed within 48 hours on the available computer hardware.

Parameter	Value
Likelihood parameter α_e	0.001
Likelihood parameter β_e	0.001
Amplitude a	1.5
Decay rate α	1.25
Number of particles N	1280
Number of auxiliary distributions M	1000
MCMC Steps per auxiliary distribution	1
Mesh (quadrilateral elements)	32×32
Resampling threshold	$0.95N$
Refinement threshold ϵ	0.05

Table 3.1: Table of key parameter values for numerical examples.

3.6.2 Preliminary studies

To gain an empirical sense of the effect of fixing the scale number S , a number of realisations are drawn from the prior distribution. These are given in Fig. 3.3, and reaffirm the appropriateness of the prior modelling. A preliminary study was performed by replacing the forward solver with the identity map. For this study only, synthetic measurement data was obtained by taking the original 32×32 permeability field with 1% relative noise. In all cases, a good reconstruction was obtained with a sufficient number of wavelets.

3.6.3 Presentation of results

Example 1

First, a straightforward but well-studied problem is considered where the data is generated from a linear log-permeability field [LHB⁺02b, MZ09, WZ11]. The purpose of this example is to investigate the proposed methodology for an example with relatively few features; a good fit

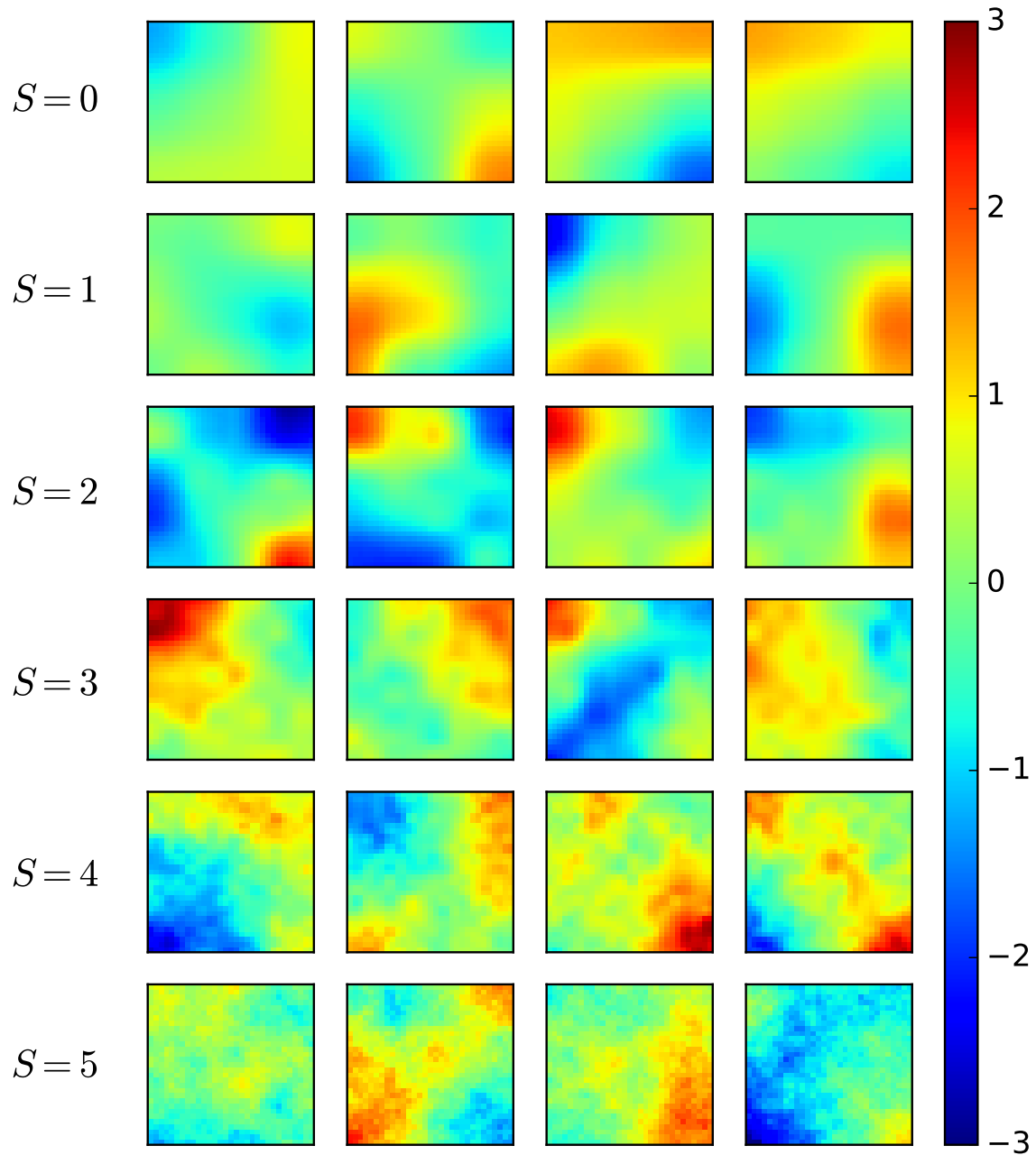


Figure 3.3: Each row shows 4 log-permeability fields drawn from the prior distribution for $S = 0, \dots, 5$ (top to bottom). There are 4^{S+1} basis functions for each model. It can be seen that the complexity of the prior increases with S , as each model includes wavelets up to scale S .

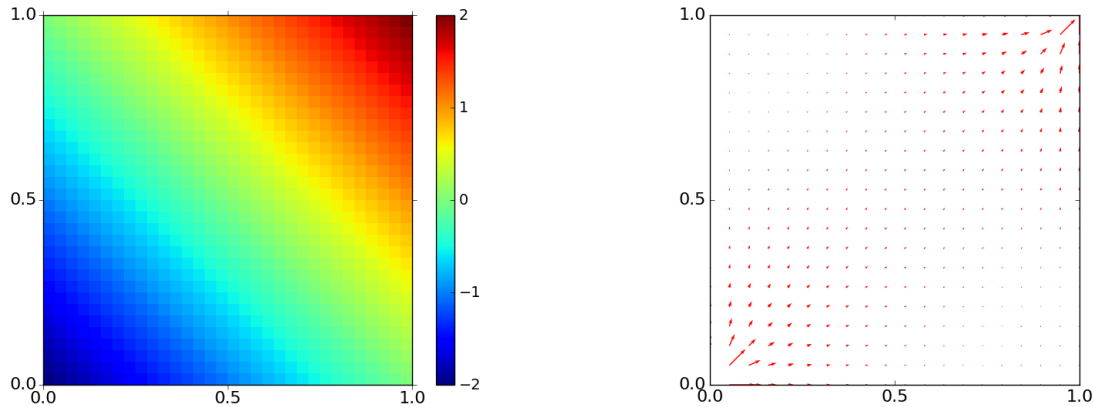


Figure 3.4: Left: True log-permeability for example 1, which is a linear function. Right: Plot of the velocity vector field over the true log-permeability.

is expected with a parsimonious model. The effect of using additional data with less noise is studied. The effect of adding additional scales to the model is also studied.

The log-permeability field used to generate the data is

$$u(x, y) = 2(x + y - 1), \quad (3.65)$$

and is illustrated in Fig. 3.4. The fluid velocity profile is also shown in Fig. 3.4. It can be seen that the fluid velocity is greater along the diagonal between the wells, and in the vicinity of the wells. It is expected to be more difficult to recover the permeability values in regions of limited flow, for example, in the top-left and bottom-right corners.

First, the fixed scale model is investigated for $S = 0$ as defined by (3.33) using the standard SMC sampler. Two synthetic datasets are used: 10×10 measurements with 1% relative noise; and 5×5 measurements with 5% relative noise. The former represents high-quality measurement data, whereas the latter represents low-quality data. It is expected that a reasonable representation of a linear log-permeability field can be obtained with relatively few basis functions. In the preliminary study, using the identity map, most salient features were recovered for the model with $S = 0$.

The estimated posterior means and quantiles for both cases (high and low noise datasets) are presented in Figs. 3.5-3.7. As expected, the posterior mean provides a good estimate of the log-permeability with $S = 0$. In both cases, the estimated field is closer to truth in locations with greater flow. The uncertainty indicated by the quantiles is higher in regions of limited

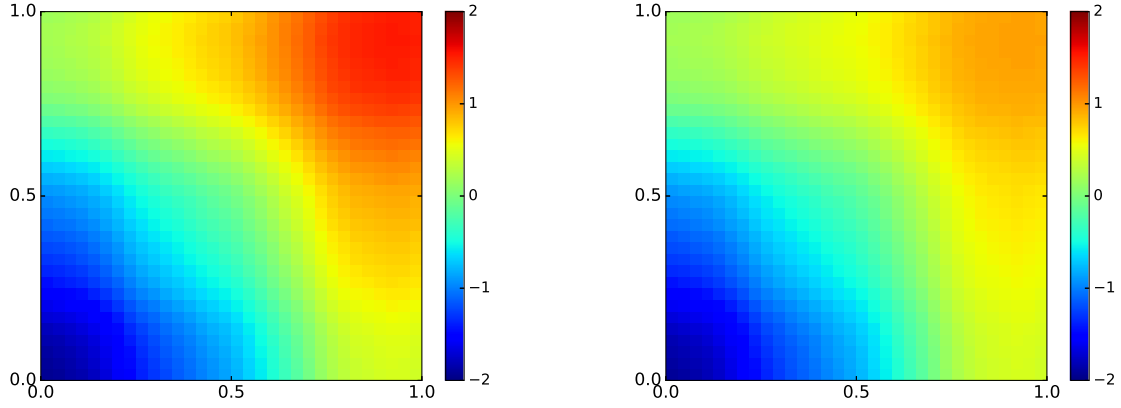


Figure 3.5: Inferred posterior means for example 1 for the low-noise 10×10 dataset (left) and for the high-noise 5×5 dataset (right).

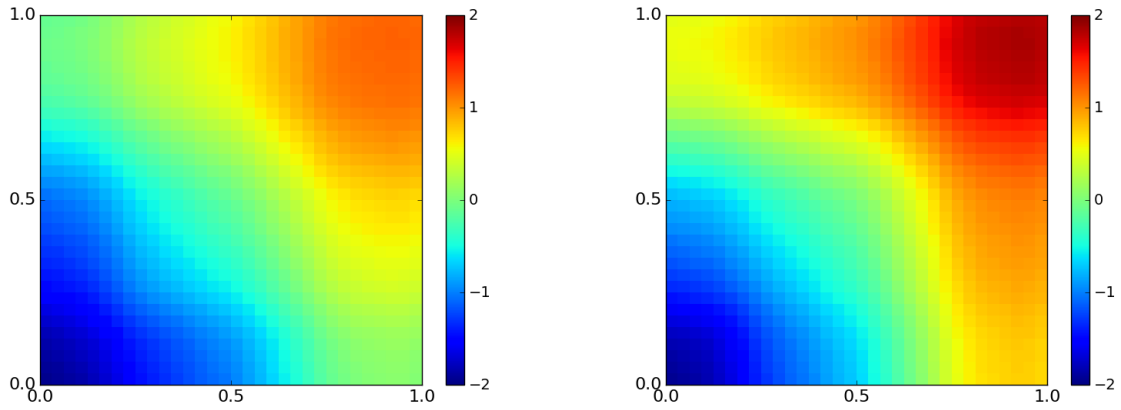


Figure 3.6: Inferred posterior quantiles for example 1 for the low-noise 10×10 dataset: 5% quantile (left) and 95% quantile (right).

flow. In line with expectations, the posterior means obtained for the 5×5 dataset are less accurate than those obtained for the 10×10 dataset, and also show increased uncertainty. The acceptance rates for the SMC sampler are in Fig. 3.8. It can be seen that the adaptive strategy maintains an acceptance rate around 20%, as desired.

Lastly, the multiscale algorithm is used for the 10×10 measurement data to study the effect of adding more basis functions to the model. The termination criterion is satisfied after $S = 0$ because the estimated Bayes factor decreased. This suggests a preference towards the simpler model. For demonstration purposes, the procedure was continued until $S = 5$ and is presented in Fig. 3.9 to demonstrate the unwanted effects of adding excessive basis functions to the model. The log-Bayes factors are computed for each S and are presented in Fig. 3.10. It can be seen

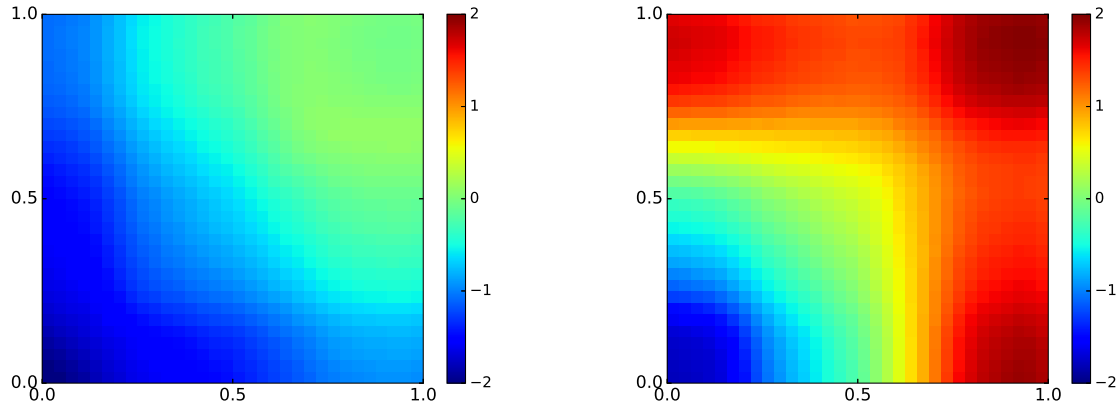


Figure 3.7: Inferred posterior quantiles for example 1 for the high-noise 5×5 dataset: 5% quantile (left) and 95% quantile (right).

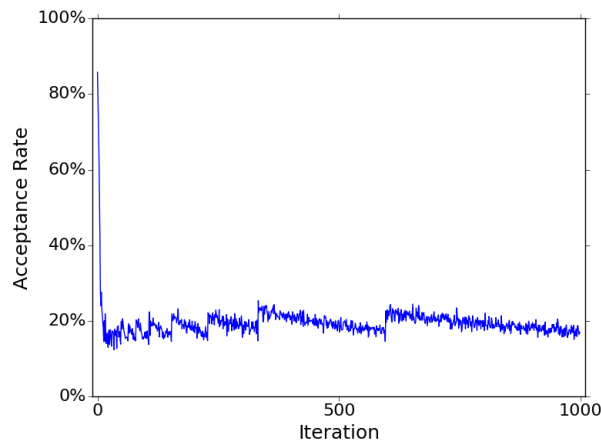


Figure 3.8: Average acceptance rates against iteration number for the random walk Metropolis transition kernels. This plot was obtained for example 1 with $S = 0$ using the 10×10 dataset.

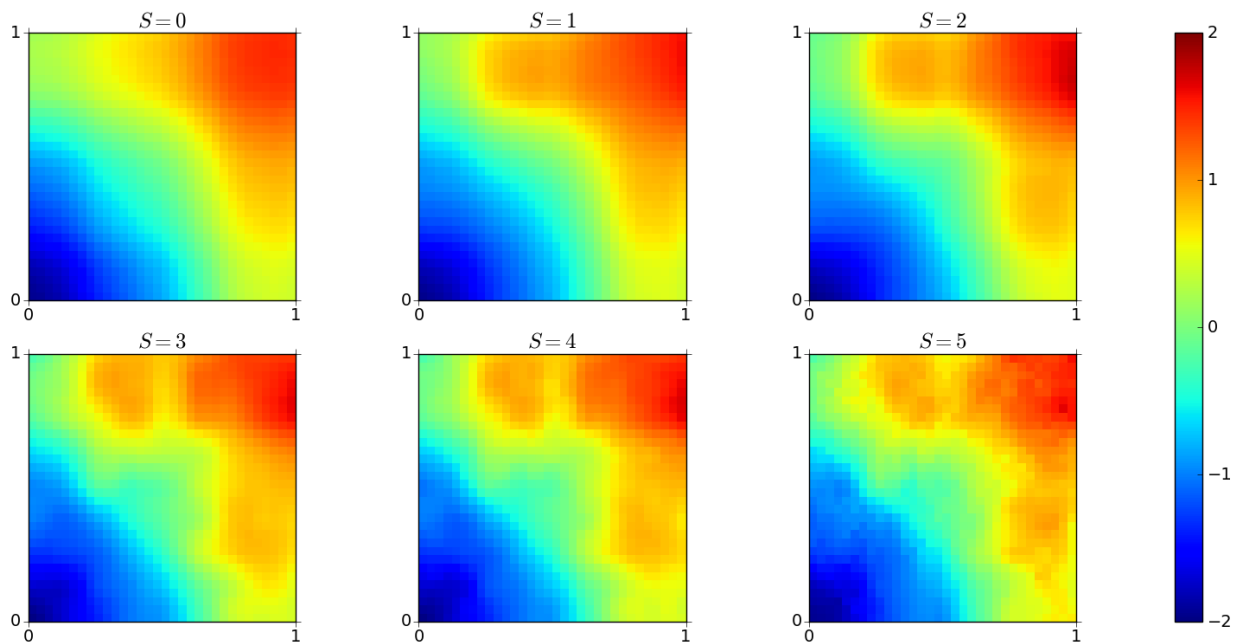


Figure 3.9: Inferred posterior means of the log-permeability for example 1 using the multiscale approach for $S = 0, \dots, 5$ (from left to right and then top to bottom). The algorithm terminates after $S = 0$ with just 4 basis functions, however, it is demonstrated that by adding additional basis functions the posterior mean is further from the truth. The number of basis functions in each model is 4, 16, 60, 208, 664 and 1848 for $S = 0, \dots, 5$, respectively.

that the marginal likelihood is maximised for $S = 0$ and continues to decrease by adding more basis functions. For further exploration of the posterior, Fig. 3.11 shows one-dimensional plots along the line $y = x$ of the posterior mean and quantiles against the true log-permeability. It can be seen that the estimate is closest to the truth for $S = 0$ and that the variance of the posterior log-permeability increases by adding further basis functions to the model.

Example 2

For the second example, the log-permeability used to generate the synthetic data is a realisation of a centred Gaussian process. The Gaussian process has a squared exponential covariance function with unit marginal variance and a length scale of $l = 0.3$. The log-permeability and velocity field used to generate the data are illustrated in Fig. 3.12. For this example, pressure is measured over a network of 10×10 sensors with 1% relative noise. There is limited flow through the top-left corner of the domain and most flow is directed around the right and bottom regions.

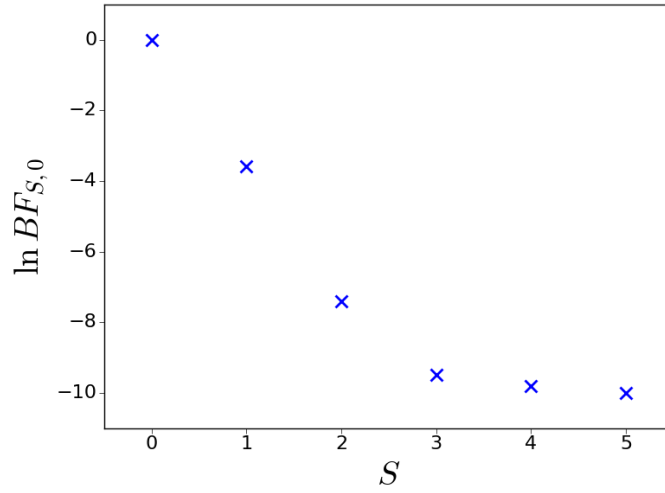


Figure 3.10: Plot of estimated log-Bayes factors for example 1 for the multiscale approach for $S = 0, \dots, 5$. The procedure terminates after $S = 0$, for which the estimated marginal likelihood is maximised. The scale determination has been run for subsequent scales to demonstrate that adding an excessive number of basis functions causes the marginal likelihood to decrease.

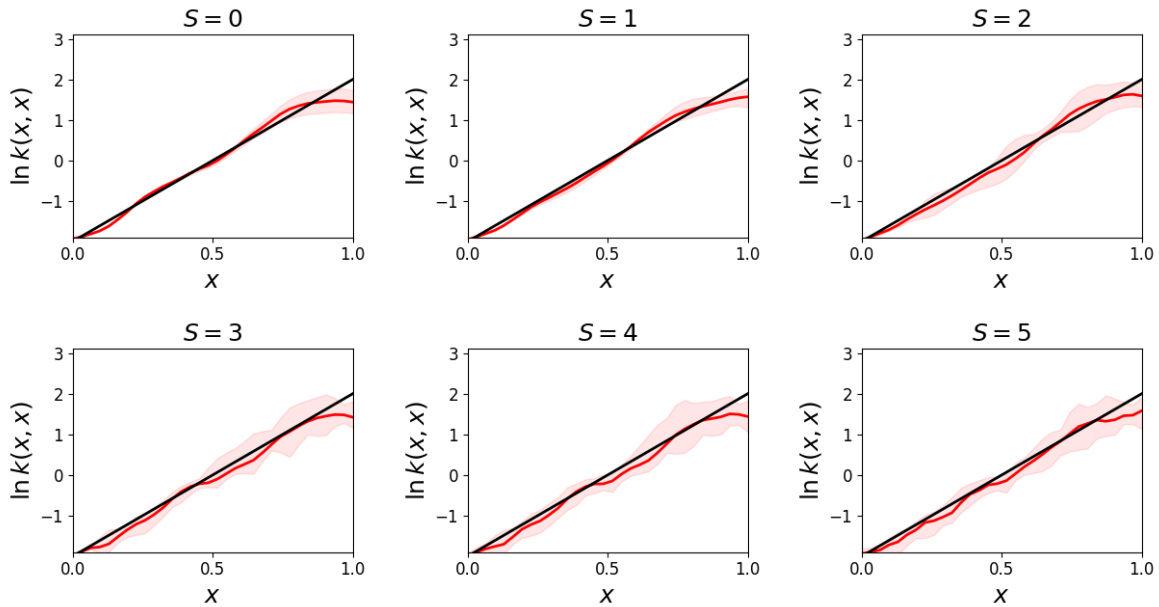


Figure 3.11: The inferred posterior mean (red line) against the true log-permeability (black line) along the line $y = x$ for example 1 using the multiscale approach for $S = 0, \dots, 5$ (from left to right and then top to bottom). The shaded region shows the 95% credible interval.

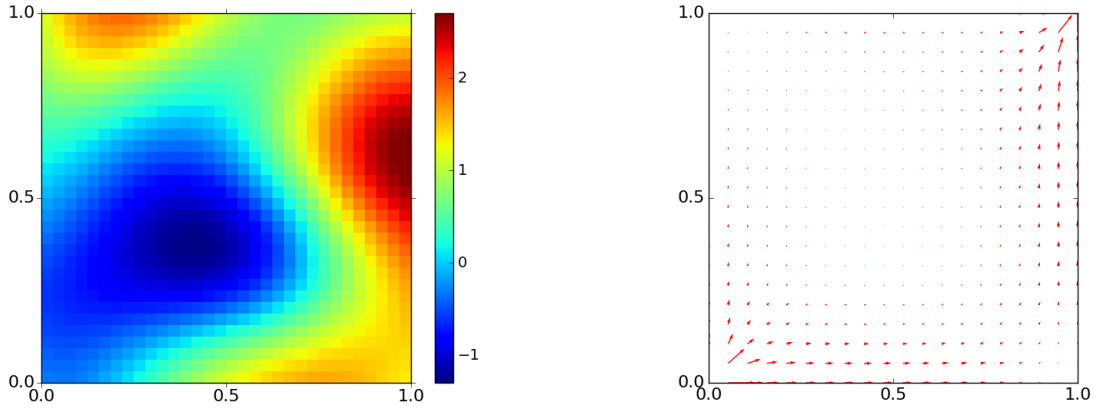


Figure 3.12: Left: True log-permeability for example 2, which is a realisation of a Gaussian process with an exponential squared covariance function. Right: Plot of the velocity vector field over the true log-permeability.

A number of fixed scale models are investigated by fixing $S = 0, \dots, 3$. The posterior mean estimates of log-permeability are illustrated in Fig. 3.13. It can be seen that the model with $S = 0$ provides a poor estimate of log-permeability. The additional basis functions in models with $S = 1$ and $S = 2$ provide improved estimates. For models with $S = 2, 3$, the model shows signs of increased complexity and the sampler can fail to find a significant mode.

The multiscale approach is used to provide a comparison. The results are illustrated in Fig. 3.14. The termination criterion is satisfied after $S = 2$. It can be seen that the multiscale SMC sampler provides superior results when compared to the standard SMC sampler. First, the multiscale approach is able to find a significant mode. Second, the multiscale approach provides an estimate that captures the most salient features. It is noted that the top-left corner is incorrectly captured, which appears to coincide with the lack of flow in the top-left corner of the domain. The Bayes factors are plotted in Fig. 3.15. There is a notable improvement in moving from $S = 0$ to $S = 1$ and terminating after $S = 2$ appears optimal by inspecting the posterior means. The posterior mean and quantiles are plotted against the true log-permeability along the line $y = x$ in Fig. 3.16. It can be seen that with additional basis functions, the posterior mean is closer to the truth. Upon moving to $S = 3$, the posterior variance increases without providing a better fit, which is consistent with the decrease in the marginal likelihood.

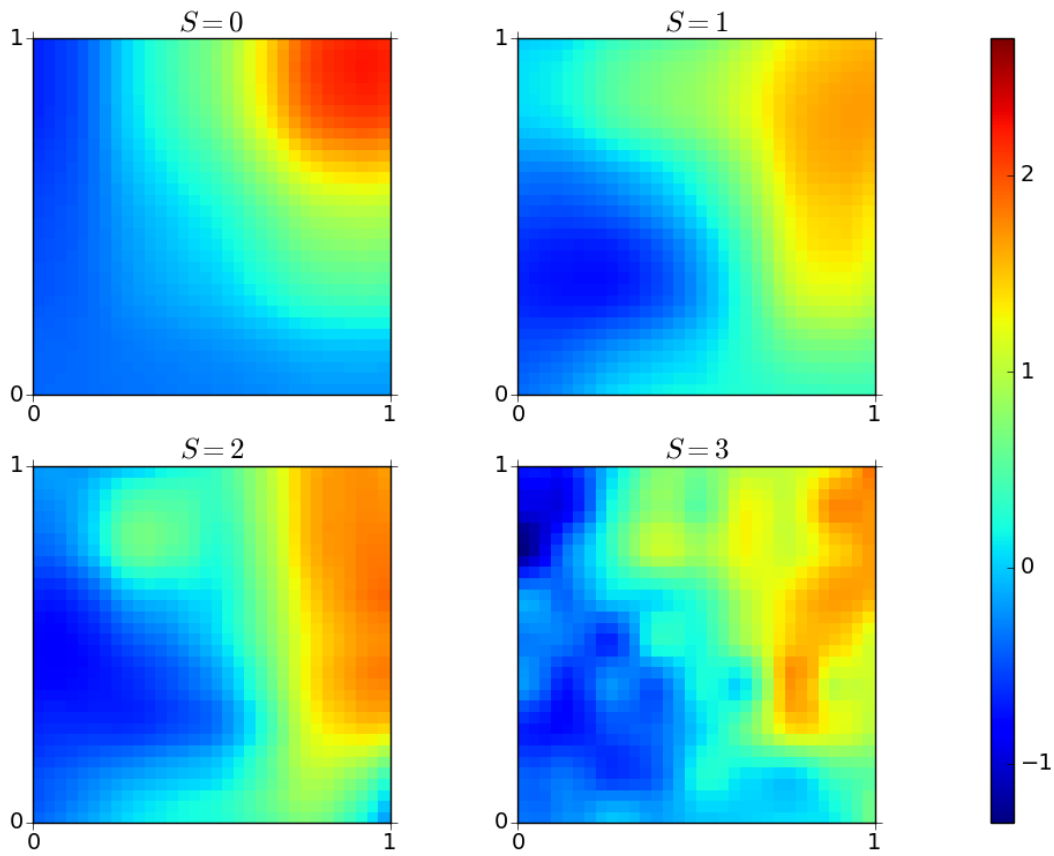


Figure 3.13: Inferred posterior means for example 2 for fixed $S = 0, \dots, 3$ (left to right and then top to bottom). The number of basis functions in each model is 0, 16, 64 and 256 for $S = 0, \dots, 3$, respectively.

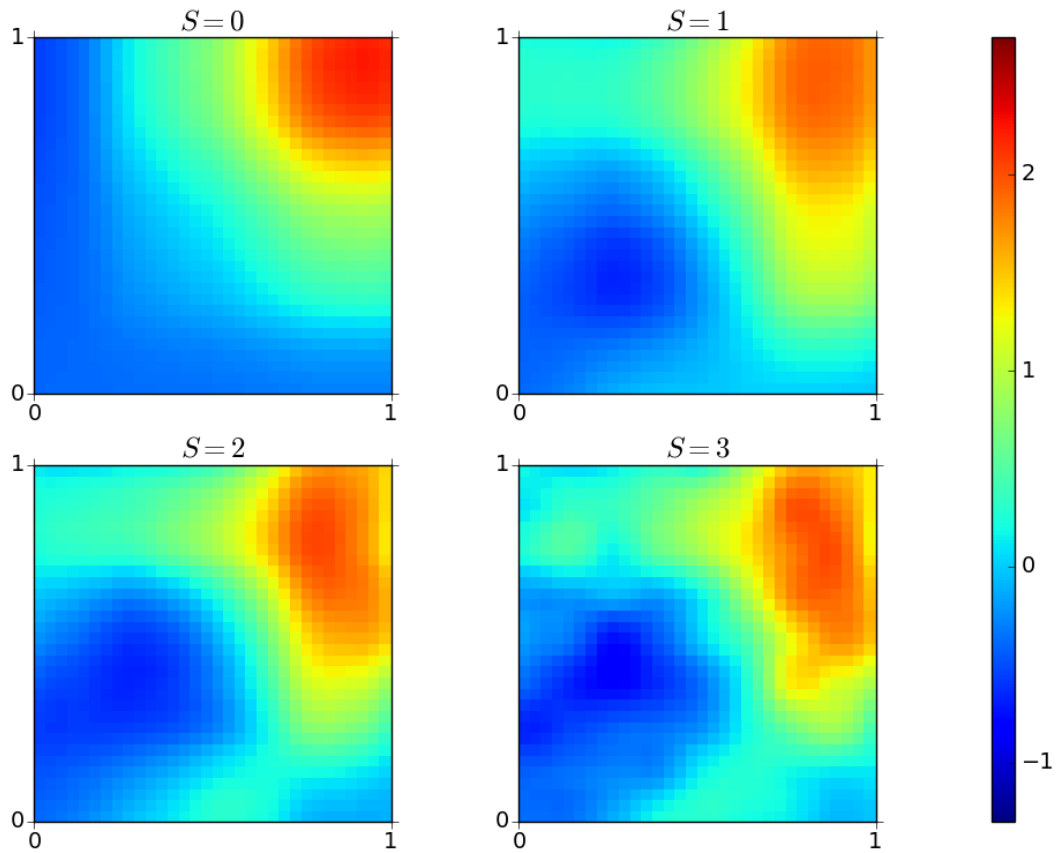


Figure 3.14: Inferred posterior means of the log-permeability for example 2 using the multiscale approach for $S = 0, \dots, 3$ (from left to right and then top to bottom). The number of basis functions in each model is 4, 16, 60 and 228 for $S = 0, \dots, 3$, respectively.

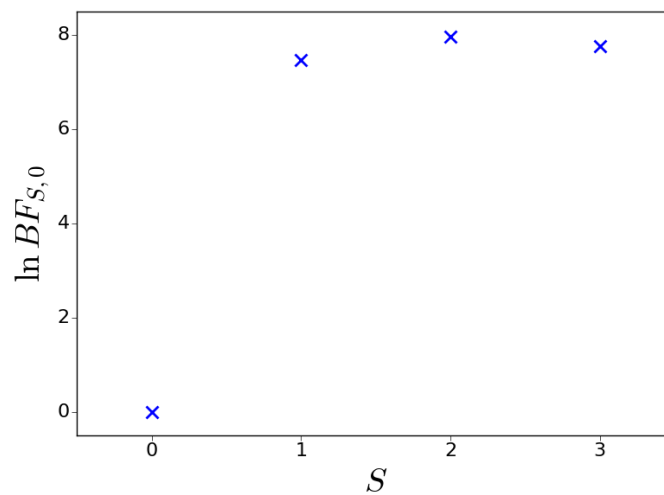


Figure 3.15: Plot of estimated log-Bayes factors for example 2 for the multiscale approach for $S = 0, \dots, 3$. The procedure terminates after $S = 2$, for which the estimated marginal likelihood is maximised.

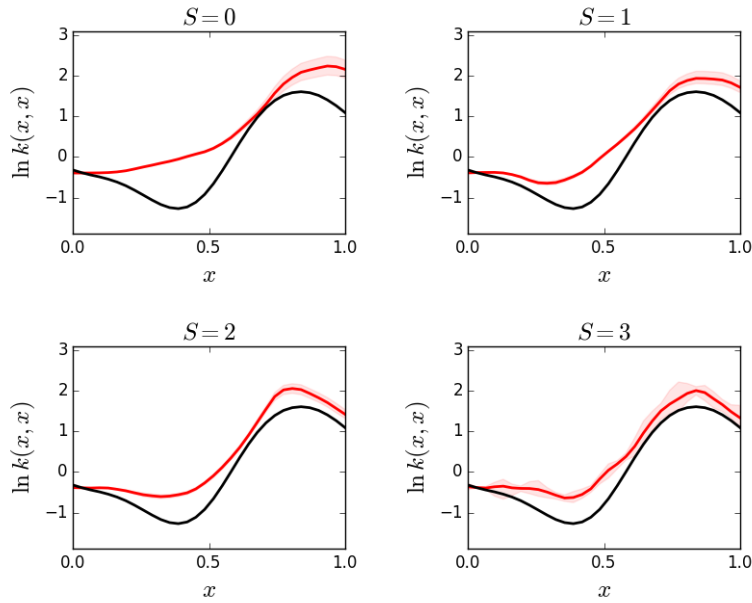


Figure 3.16: The inferred posterior mean (red line) against the true log-permeability (black line) along the line $y = x$ for example 2 using the multiscale approach for $S = 0, \dots, 3$ (from left to right and then top to bottom). The shaded region shows the 95% credible interval.

Example 3

For this example, the log-permeability field used to generate the data is also a Gaussian process. An exponential covariance function is used with the marginal variance and length scale as before. This covariance function generates a random field with multiscale characteristics, as illustrated in Fig. 3.17. The flow covers most of the domain, but with little flow in the top-left and bottom-right corners; it is expected that the log-permeability estimates are less accurate for these areas. Two synthetic datasets are used for this example: a 10×10 sensor network with 1% relative noise and a 5×5 sensor network with 5% relative noise, respectively. The multiscale approach is compared for the two different datasets.

For the 10×10 dataset, the procedure provides models for scales $S = 0, \dots, 5$ with 4, 16, 64, 224, 848 and 1740 basis functions, respectively, indicating that several areas for refinement were identified. The termination criterion was satisfied for $S = 4$, as the marginal likelihood decreased for $S = 5$. The posterior means and quantiles, Bayes factors and approximate posterior realisations are plotted in Figs. 3.18-3.21. The posterior means show evidence of refinement until $S = 3$. Whilst some of the finer characteristics are not visible after averaging, they can be seen in the (approximate) realisations from the posterior distributions in Fig. 3.20. Several

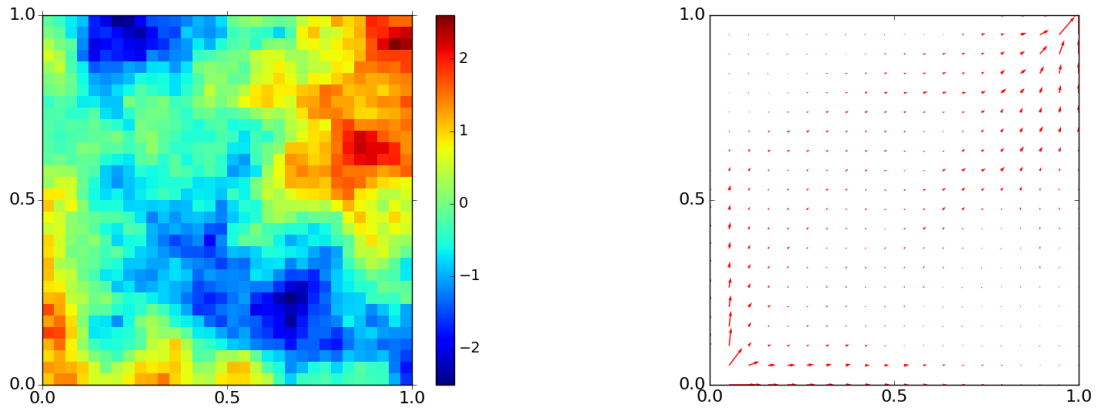


Figure 3.17: Left: True log-permeability for example 3, which is a realisation of a Gaussian process with an exponential covariance function. Right: Plot of the velocity vector field over the true log-permeability.

artefacts appear because the fluid can easily go around small regions of lower permeability. This is also likely to be the reason for the small changes in the Bayes factors after $S = 2$. As before, the posterior mean and quantiles are plotted against the true log-permeability along the line $y = x$ in Fig. 3.22. With refinements, the posterior mean of the log-permeability along the line $y = x$ becomes closer to the truth. As with all previous examples, estimates of permeability are less accurate in the regions with limited flow.

For the 5×5 pressure sensor network, the procedure provides models for scales $S = 0, \dots, 2$ with 4, 12 and 44 basis functions, respectively, as the termination criterion is satisfied after $S = 1$. As expected, the limited data resulted in earlier termination with fewer refinements. The posterior means and quantiles, Bayes factors and approximate posterior realisations are plotted in Figs. 3.23-3.26. The 10×10 dataset resulted in estimates closer to the truth with reduced uncertainty.

Lastly, the empirical distributions of the wavelet coefficients are examined at different scales. For the 10×10 dataset, the empirical distributions of four selected wavelet coefficients are presented in Fig. 3.28. The distributions for the selected coarse features are uni-modal with a strong peak. For the selected wavelet coefficient that was not refined, the empirical distribution closely resembles the prior. This suggests that the prior is a good choice of initial distribution for the SMC sampler, especially for the higher scales. Similar findings were made in Fig. 3.29 for the 5×5 dataset.

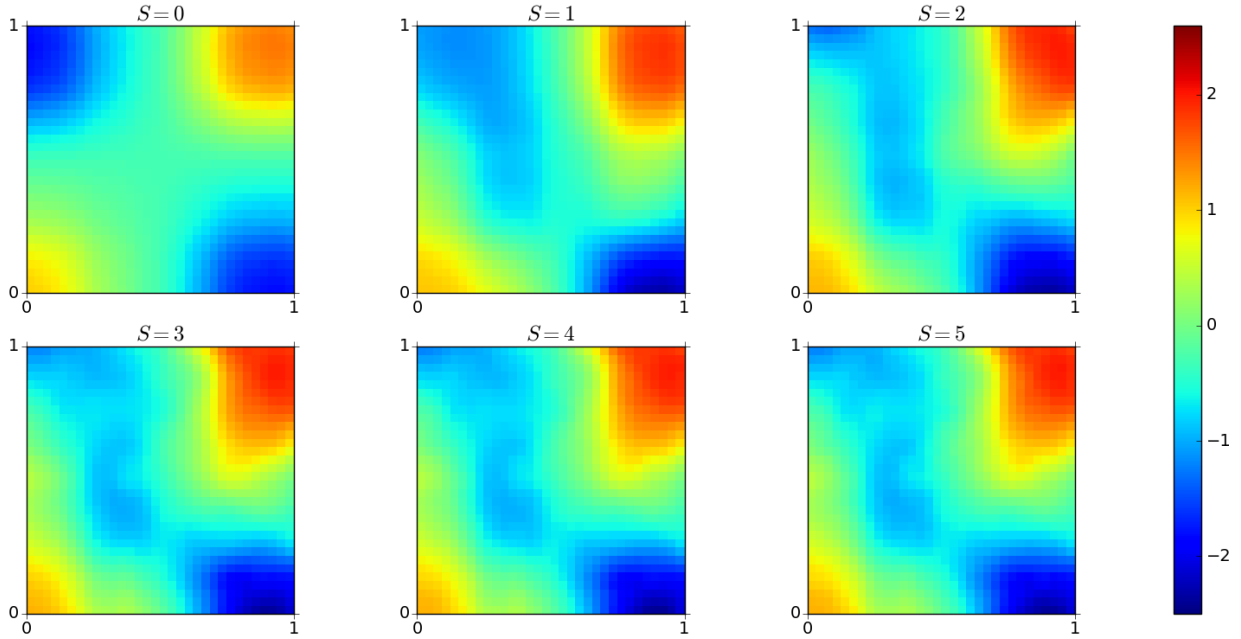


Figure 3.18: Inferred posterior means of the log-permeability for example 3 using the multiscale approach for $S = 0, \dots, 5$ for the low-noise 10×10 dataset (from left to right and then top to bottom). The number of basis functions in each model is 4, 16, 64, 224, 848 and 1740 for $S = 0, \dots, 5$, respectively.

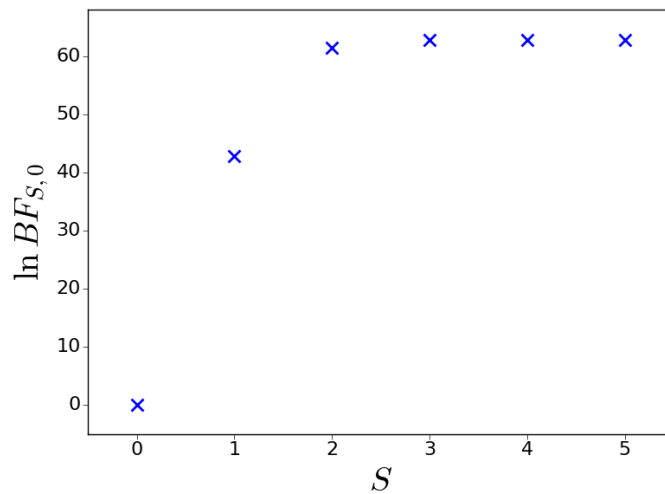


Figure 3.19: Plot of estimated log-Bayes factors for example 3 for the multiscale approach for $S = 0, \dots, 5$ for the low-noise 10×10 dataset. The procedure terminates after $S = 4$, for which the estimated marginal likelihood is maximised.

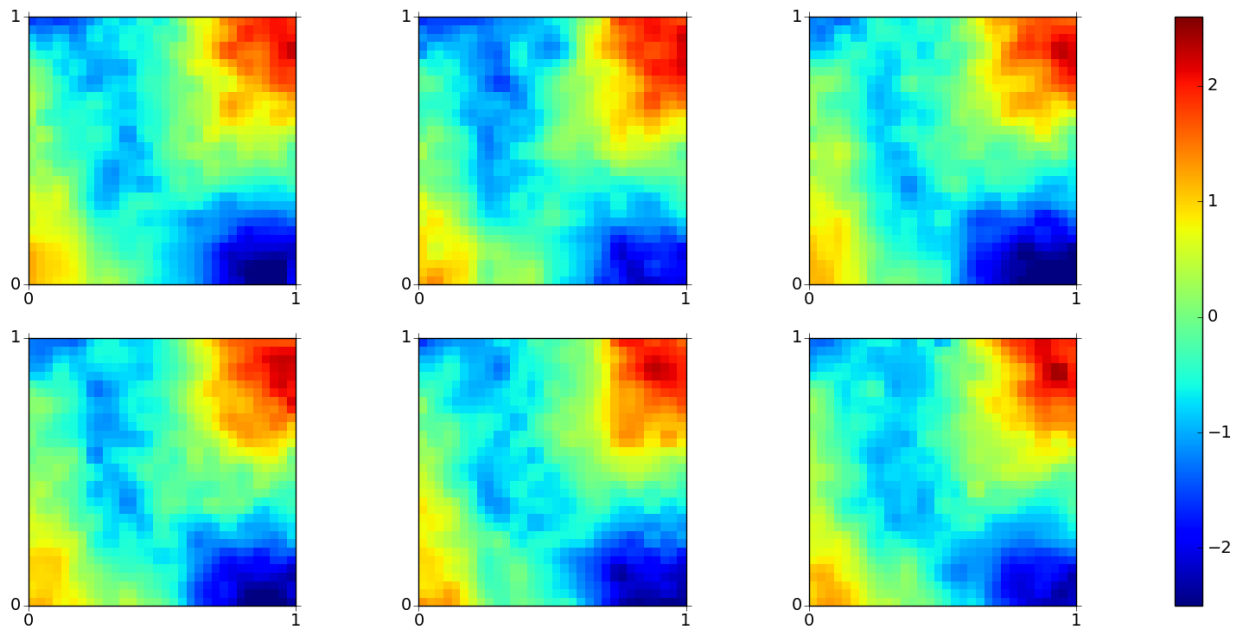


Figure 3.20: Approximate draws from the posterior density for $S = 4$ for example 3 for the low-noise 10×10 dataset.

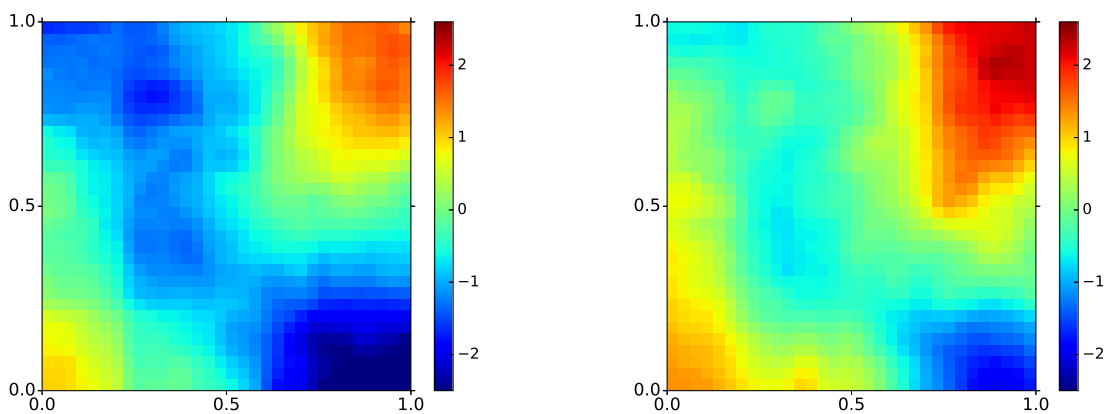


Figure 3.21: Inferred posterior quantiles for example 3 for the low-noise 10×10 dataset: 5% quantile (left) and 95% quantile (right).

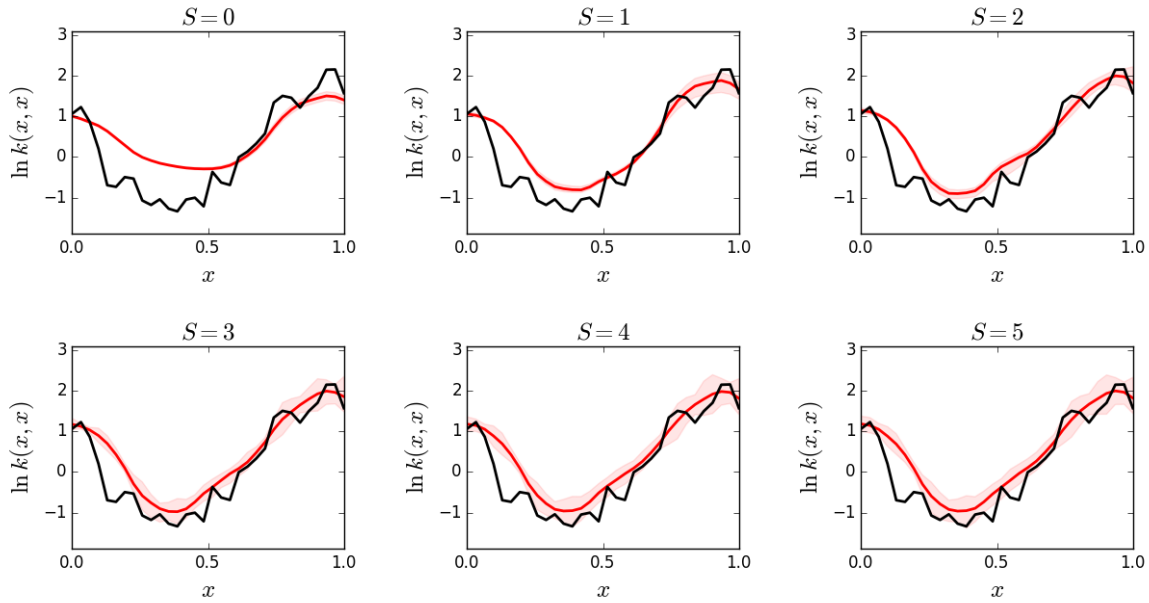


Figure 3.22: The inferred posterior mean (red line) against the true log-permeability (black line) along the line $y = x$ for example 3 using the multiscale approach for $S = 0, \dots, 5$ (from left to right and then top to bottom). The shaded region shows the 95% credible interval.

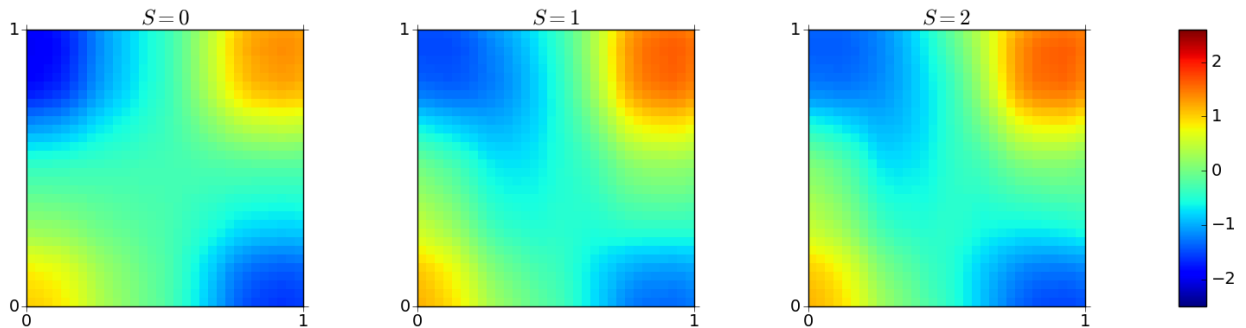


Figure 3.23: Approximate draws from the posterior density for $S = 1$ for example 2 for the high-noise 5×5 dataset.

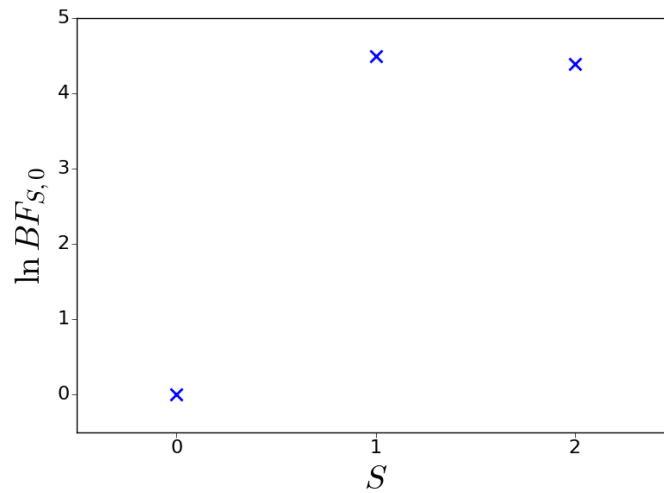


Figure 3.24: Plot of estimated log-Bayes factors for example 3 for the multiscale approach for $S = 0, \dots, 2$ for the high-noise 5×5 dataset. The procedure terminates after $S = 2$, for which the estimated marginal likelihood is maximised.

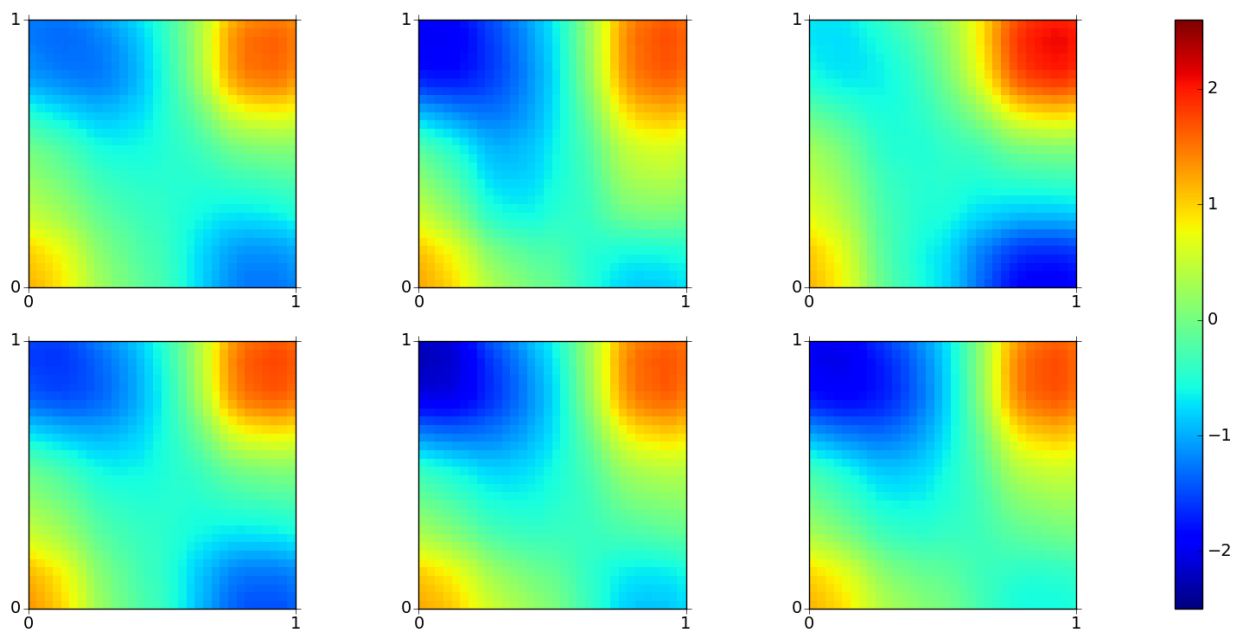


Figure 3.25: Approximate draws from the posterior density for $S = 1$ for example 3 for the low-noise 10×10 dataset.

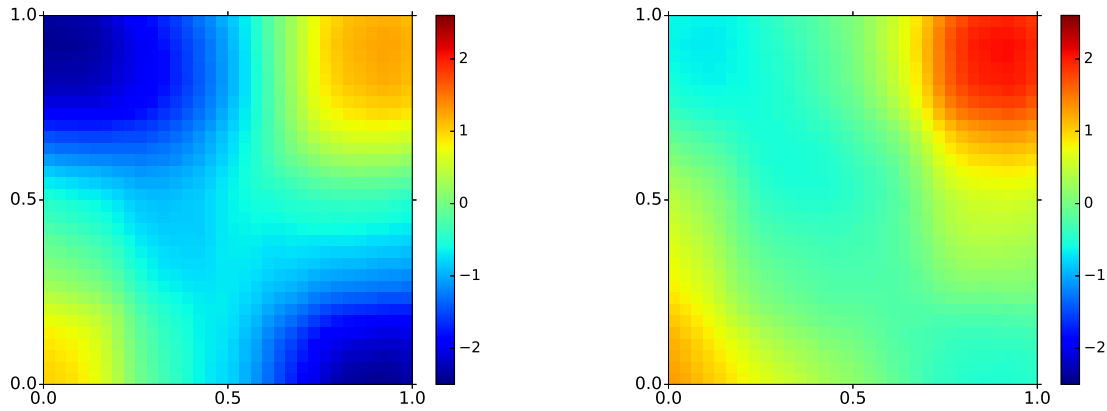


Figure 3.26: Inferred posterior quantiles for $S = 1$ example 3 for the high-noise 5×5 dataset: 5% quantile (left) and 95% quantile (right).

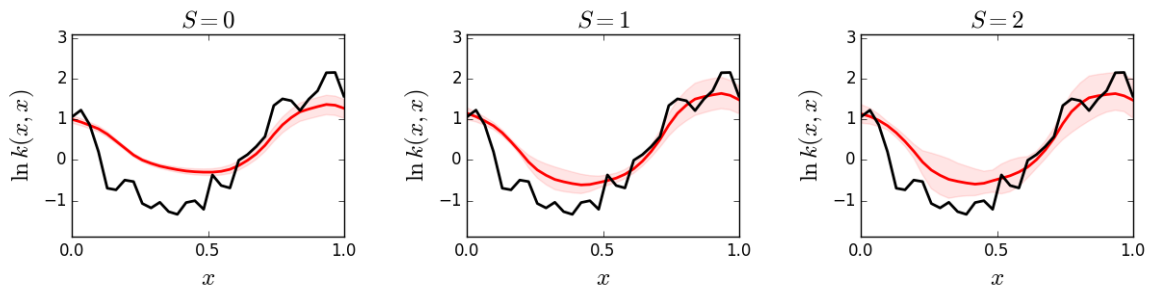


Figure 3.27: The inferred posterior mean (red line) against the true log-permeability (black line) along the line $y = x$ for example 3 using the multiscale approach for $S = 0, \dots, 2$ (from left to right and then top to bottom). The shaded region shows the 95% credible interval.

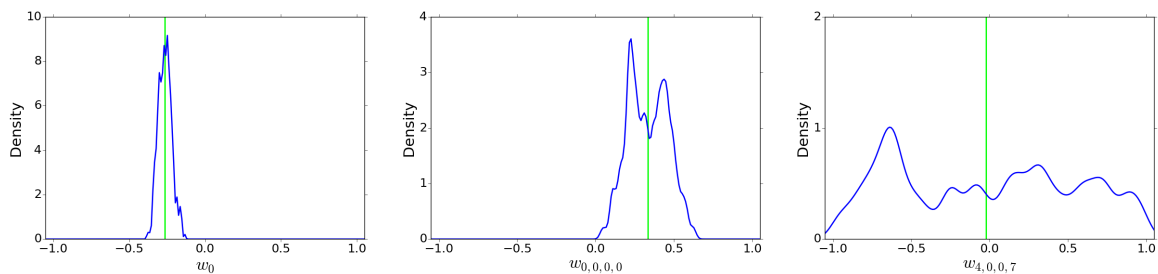


Figure 3.28: Estimated posterior marginal density plots for selected wavelet coefficients for the low-noise 10×10 dataset, using kernel density estimation for $S = 4$. The green line indicates the posterior mean of the coefficients. The coefficient $w_{4,0,0,7}$ closely resembles the prior and did not satisfy the refinement criterion.

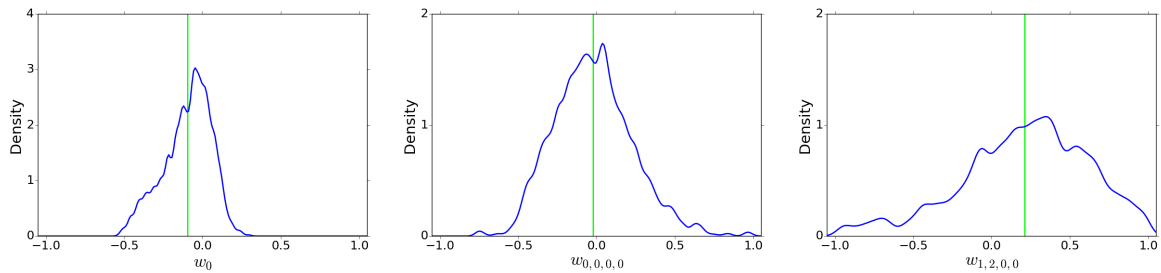


Figure 3.29: Estimated posterior marginal density plots for selected wavelet coefficients for the high-noise 5×5 dataset, using kernel density estimation for $S = 1$. The green line indicates the posterior mean of the coefficients. In contrast to the inference for the low-noise 10×10 dataset, the coefficients $w_{0,0,0,0}$ was concentrated around zero and was not selected for further refinement.

3.7 Discussion and conclusions

A new multiscale approach has been proposed and investigated for the Bayesian estimation of a spatially varying parameter that is observed via a complex function. The unknown parameter is represented by wavelet basis functions, up to a given scale, whose coefficients are modelled as random variables that are conditioned on the measurement data. Wavelet basis functions form a multi-resolution analysis, therefore allowing a sequence of Bayesian models to be defined from coarse to fine scales, in order of increasing model complexity.

It was found that wavelet basis functions provide an effective representation of an unknown spatially varying parameter with multiscale characteristics. The multi-resolution structure of the unknown can be exploited by a multiscale SMC sampler that efficiently moves between models from coarse to fine scales. The multiscale approach was shown to be well-suited to exploring high-dimensional and multimodal posterior distributions. The hierarchical nature of wavelet basis functions was also used to suggest a more parsimonious refinement of the previous model by excluding wavelets that are unlikely to improve the model fit.

As the proposed models are of increasing complexity, the chosen model effectively determines the characteristic length scale of the unknown. The model fit can be assessed via the marginal likelihoods, which can be estimated as a by-product SMC sampler. It is demonstrated that a more complex model does not always provide a better explanation of the data. The use of the marginal likelihood has been suggested as a termination criterion for adding additional scales to the model, which is an improvement of the work in [WZ11].

The proposed method was demonstrated for the task of permeability estimation in corner-to-corner groundwater flow, however, it is emphasised that the approach presented is applicable to other inverse problems that have an unknown spatially varying parameter. The SMC algorithm was implemented in parallel, so the expensive calls to the forward solver were distributed over the available computer hardware. It is highlighted that there is potential to reduce the cost of the multiscale SMC sampler by reducing the accuracy of calculations at the coarser scales.

Many challenges and open questions remain to be addressed. For example, dealing with data that is multiscale, multi-resolution or multi-physics is an essential component of many inverse problems in domains such as climate modelling, geological flows, and many more. The Bayesian approach to inverse problems remains a computationally intensive task and variational approximations may be useful in this context [TBKZ16, AZ19]. For example, variational approximations may be used to accelerate the computation at coarser scales to assist asymptotically exact computations at the finer scales. In the work presented here, the model error was assumed to be independent and identically distributed throughout, which is a reasonable assumption when the forward solver uses a fine grid. Being able to relax this assumption and adapt the accuracy of the forward model to the current estimate of permeability maybe an important step to controlling the computational cost in complex inversion tasks. Surrogate forward models may also be used to reduce the cost of likelihood calculations, especially at the coarser scales [MZ09].

Chapter 4

Bayesian parameter estimation for high-dimensional Gaussian models

The Gaussian distribution provides a flexible class of tractable models that are extensively used for the spatial and spatio-temporal modelling of physical and socio-economic systems. The probability density function of a Gaussian random variable contains a matrix determinant, which poses tractability issues when working in higher dimensions. In this chapter, a new approach is proposed to replace the determinant term with auxiliary variables so that simulation-based Bayesian inference is tractable for a larger class of problems.

The work in this chapter was performed in collaboration with Heiko Strathmann, Iain Murray and Mark Girolami, and parts have been published in [ESGM17].

4.1 Context and literature review

Gaussian models are one of the most fundamental tools for statistical analysis and arise frequently in the spatial and spatio-temporal modelling of physical and socio-economic systems [KS06, FL07, GDGF10]. Their wide applicability is due to at least two reasons. First, the average of a large number of independent and identically distributed random variables approximately has a Gaussian distribution. Second, the conditional and marginal distributions for a collection of jointly Gaussian random variables are also Gaussian, and their probability densities are known in closed form. As there are many possible ways to specify the second-order properties of a Gaussian random variable, the Gaussian distribution provides a rich and flexible class of

tractable models.

When a number of Gaussian latent variables are observed indirectly via a linear transformation plus additive Gaussian noise, the observations also have a Gaussian distribution. In a Bayesian setting, it is possible to obtain closed form expressions for key quantities such as the marginal likelihood and posterior predictions. Popular examples with this set-up are non-parametric regression using Gaussian processes (GPs) [RW06] and Gaussian Markov random field (GMRF) models [RH05, LRL11]. GPs are often specified by full covariance matrices, which means that each variable is correlated with the others. In contrast, GMRFs model conditional independence via zero entries in the precision matrix, which is typically sparse. When there are n modelled variables, an evaluation of the probability density function generally requires $O(n^3)$ operations and $O(n^2)$ storage. The costs arise from the solution of a linear system and the evaluation of a matrix determinant. There is a large literature on approximations to the model itself, for example, via sub-sampling, low-rank approximations or by assuming a Kronecker or Toeplitz structure [ZLL05, QCR05, QCRW07, CJ08, KR10, Saa12, HFL13, LPS14, FWN⁺15]. The scalability concerns are also recognised in data assimilation, for example, the Kalman filter is an exact algorithm to determine the filtering distribution for linear state-space models with additive Gaussian noise [LSZ15]. The required updates eventually become prohibitively expensive when there are a large number of modelled variables. The ensemble Kalman filter instead provides an inversion-free approach to approximate large covariance matrices [Eve03, LSZ15]. Recently, an inversion-free approach to infer point-estimates of covariance parameters of large-scale Gaussian distributions was proposed by [ACS17].

When the observation process is non-linear or when the latent Gaussian model is corrupted by non-Gaussian noise, the observations do not have a Gaussian distribution. For example, Poisson model is used for data in [DMRT13], or a probit or logit model is used for binary or categorical data, respectively, in [RW06]. Non-linear observation processes also arise in inverse problems [DS17]. Several MCMC methods have been proposed to simulate from the posterior distribution in this setting, for example, the Metropolis-adjusted Langevin algorithm (MALA), Hamiltonian Monte Carlo (HMC) and the preconditioned Crank-Nicolson Langevin algorithm have been used [Nea97, GC11, DMRT13, CRSW13]. More recently, auxiliary variable techniques have been considered for latent Gaussian models in [TP18]. Each of these methods suffer the aforementioned limitation in that an evaluation of the probability density function for a new set

of model parameters generally requires $O(n^3)$ operations and $O(n^2)$ storage. An alternative to marginalising over the model parameters is to maximise an approximate marginal likelihood with respect to the model parameters, however, type II maximum likelihood is an approximate method and can lead to over-fitting [RW06].

For a high-dimensional Gaussian with a sparse precision matrix, it may not be possible to obtain the required Cholesky factorisation due to the memory requirements imposed by the so-called fill-in effect [Aun12, ASE14]. Noting the sparsity, it can be possible to use iterative methods to solve the linear system and to estimate the determinant instead. It has also been recognised that one can simulate from these models via iterative methods by applying an inverse-matrix-square-root to a vector [Aun12, AEP13]. Iterative methods exploit the fact that matrix-vector products can be computed and do not require the storage of any additional matrices. A number of numerical approximations are available for estimating the log-determinant term [BFG96, HMS15, SAI17]. Rational approximations reduce the computation of matrix functions, such as the determinant, inverse or inverse-square-root, to a series of shifted linear systems. Rational approximations have error guarantees that are straight-forward to control and can even be set to floating point precision [Ken04, HL13]. Iterative methods such as the conjugate gradient method and multi-shift variants [Jeg96, Sim08] can be used. The application of iterative methods for type II maximum likelihood estimation is considered in [Aun12, ASE14] and continues to be an active area of research [DEN⁺17].

Bayesian inference of the model parameters, however, remains an open problem. The applicability of Markov chain Monte Carlo (MCMC) is limited by the need to compute determinants to calculate the Metropolis-Hastings acceptance probability [MA10]. In [LGA⁺15], it is recognised that the log-determinant estimator proposed in [Aun12, ASE14] is unbiased, and can be combined with the pseudo-marginal MCMC algorithm [AR09]. A complication is that the required exponentiation transforms an unbiased estimator of the log-determinant into a biased estimator determinant. In [LGA⁺15], an unbiased estimator of the determinant is obtained via the Russian roulette methodology. The performance of the proposed scheme is dependent on the variance of the unbiased estimator; there are a number of tuning parameters that must be carefully specified. The cost of controlling the variance of the unbiased estimator makes the scheme infeasible in certain cases.

In this chapter, an alternative MCMC approach is proposed and investigated to perform asymptotically exact Bayesian inference for large-scale Gaussian models. Auxiliary variables are introduced so that the joint distribution to be simulated from contains no determinant terms. It is instead necessary to update the auxiliary variables that replace the determinant terms, which can be achieved by applying an inverse-matrix-square-root to a vector. In the case that the inverse-matrix-square-root of the covariance is unknown, rational approximations are used to perform tuning-free updates. The scheme is considerably simpler than the approach taken in [LGA⁺15], and by avoiding the pseudo-marginal framework there is no need to tune the internal unbiased estimators. The approach scales well to high-dimensional Gaussian models, provided that the application of an inverse-matrix-square-root can be carried out reliably. In the case of a poorly-conditioned model the underlying iterative method may converge slowly, or not at all to within the desired tolerance.

4.2 Linear Gaussian models

By a linear Gaussian model, it is meant that the finite-dimensional vector of latent variables $x \in \mathbb{R}^m$ is Gaussian-distributed

$$x \sim N(\mu, \Sigma), \quad (4.1)$$

as specified by the probability density function

$$\pi(x) = \left(\frac{1}{(2\pi)^n |\Sigma|} \right)^{1/2} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right). \quad (4.2)$$

Here, $\mu \in \mathbb{R}^m$ is the mean vector and $\Sigma \in \mathbb{R}^{m \times m}$ is the symmetric positive definite covariance matrix. A linear transformation of a Gaussian is also Gaussian:

$$Ax + b \sim N(\mu + b, A\Sigma A^T), \quad (4.3)$$

where $A \in \mathbb{R}^{n \times m}$ and $b \in \mathbb{R}^n$. Therefore, when the Gaussian model is observed via a linear operator plus Gaussian noise, the observations are also Gaussian. The latent variables x may be a finite-dimensional realisation of a continuous model of spatial variation, or a realisation of a discrete-space model. Two possibilities are given below [GDGF10].

Gaussian Markov random fields A Gaussian Markov random field (GMRF) is a Gaussian random variable $(x_1, \dots, x_m)^T \in \mathbb{R}^m$ that satisfies a Markov property. More specifically, let j index the nodes of a graph, and let each variable x_j be the value the process takes at the respective node j . The Markov property is

$$P(x_i \in A | \{x_j : j \neq i\}) = P(x_i \in A | \{x_j : j \in \mathcal{N}_i\}), \quad (4.4)$$

where $\mathcal{N}_j$ is the set indices of the nodes that x_j is connected to. It is common to specify a GMRF model by a system of linear equations with a Gaussian-distributed right-hand-side. The linear system provides the square-root of the precision matrix.

Gaussian processes A Gaussian process (GP) is a mean-square continuous stochastic process $\{x(s) \in \mathbb{R} : s \in \mathcal{X}\}$, for a spatial domain $\mathcal{X}$, whose finite-dimensional distributions (FDDs) are multivariate Gaussian distributed random variables. This means that for any finite set of locations $s_1, \dots, s_m \in \mathcal{X}$, the random vector $(x(s_1), \dots, x(s_m))^T \in \mathbb{R}^m$ is a Gaussian random variable. The Kolmogorov existence theorem states that if the FDDs of a stochastic process are consistent under reordering and marginalisation, then the stochastic process is uniquely defined by its FDDs. A GP is specified the first and second order statistics of its FDDs.

4.2.1 Covariance functions

A (centred) GP can be uniquely specified by a covariance function that provides the covariance between any two locations. Certain GMRFs can be obtained by discretising the stochastic partial differential equation (SPDE) representation of a GP with a Matérn covariance function.

The requirement that the covariance matrix is always positive definite is equivalent to the requirement that the covariance function is positive definite [RW06, GDGF10, Pav16]. Two simplifying assumptions are stationarity and isotropy. This means that the covariance function only depends on the Euclidean distance between the two locations. The covariance function of a stationary and isotropic process can be written as

$$\Sigma(s_1, s_2) = \sigma^2 \varphi(r/l), \quad (4.5)$$

where $r = |s_1 - s_2|$ is the distance between the two observations locations s_1 and s_2 , respectively,

σ^2 is the marginal variance at any location, l is the length scale controlling the rate at which correlations decay, and $\varphi : [0, \infty) \rightarrow \mathbb{R}_{\geq 0}$ is a suitably defined function. A stationary GP is mean-square continuous if and only if its covariance function is continuous. A suitable function φ can be determined by Bochner's theorem, which states that a real-valued continuous (covariance) function is positive definite if and only if it is the inverse Fourier transform of a symmetric positive measure. It is therefore possible to construct valid covariance functions via the spectral density function, which is the Lebesgue density of a symmetric positive measure. Bochner's theorem has lead to two popular covariance families that possess a sparse structure: Matérn covariance functions and compactly supported polynomial covariance functions.

Matérn covariance functions

The Matérn covariance functions are obtained when

$$\varphi(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} (r)^\nu K_\nu(r), \quad (4.6)$$

where K_ν is the modified Bessel function of the second kind, and $\nu > 0$ is a smoothness parameter [GDGF10]. The Matérn covariance function is valid in all dimensions $d \geq 1$. When $\nu = 1/2$, one obtains the exponential covariance function $\varphi(r) = \exp(-r)$ and in the limit $\nu \rightarrow \infty$ one obtains the squared exponential covariance function $\varphi(r) = \exp(-r^2/2)$.

GP with Matérn covariance function are also stationary solutions of the fractional SPDE [GDGF10, LRL11] given by

$$(\kappa^2 - \Delta)^{\alpha/2} x(s) = \xi(s), \quad \alpha = \nu + d/2. \quad (4.7)$$

The underlying sparse structure is apparent for $\alpha = 2$, as the discrete Laplacian can be written as

$$\Delta x \approx \frac{1}{h^2} \sum_{j=1}^d \left\{ x(s - he_j) - 2x(s) + x(s + he_j) \right\}, \quad (4.8)$$

where e_j is the unit vector in the j th direction. Moreover, the d -dimensional discrete Laplacian can be written as d Kronecker sum [KR10], so can requires very little memory to work with. By discretising the SPDE in (4.7), one obtains a GMRF model that has a sparse precision matrix.

Compactly supported polynomial covariance functions

Compactly supported polynomial covariance functions have a computational appeal because the covariance matrices they create are often sparse. For example, the spherical covariance function

$$\varphi(r) = \left(1 - \frac{3}{2}r + \frac{1}{2}r^3\right) \mathbb{I}_{\{0 \leq r \leq 1\}}(r), \quad (4.9)$$

is a valid covariance function for an input dimension number $d \leq 3$ [FL07]. Many other compactly supported covariance functions are available. For example, covariance functions that are $2q$ -times continuously differentiable are proposed in [Wen95]. For example, for $q = 2$ the following function can be used

$$\varphi(r) = \left((1 - r)^4(1 + 4r)\right) \mathbb{I}_{\{0 \leq r \leq 1\}}(r), \quad (4.10)$$

which again is valid for $d \leq 3$. The challenge in identifying suitable covariance functions is determining positive definiteness; naively setting small covariances to zero does not guarantee this.

4.2.2 Bayesian inference for linear Gaussian models

Here, it is assumed that a Gaussian model is observed via a linear transformation plus additive Gaussian noise:

$$y = Ax + e, \quad (4.11)$$

where $y, e \in \mathbb{R}^n$, $x \in \mathbb{R}^m$ and $A \in \mathbb{R}^{n \times m}$ with

$$x|\theta \sim N(\mu_\theta, \Sigma_\theta), \quad (4.12)$$

$$e|\theta \sim N(\delta_\theta, \Gamma_\theta). \quad (4.13)$$

For example, A may be the identity matrix, in which case x represents the process at the locations where observations are made. This means that the dimensionality of the latent process is equal to the number of observations. A may also be a matrix that performs linear interpolation, so that the dimensionality of x does not depend on the number of observations. The dimensionality of the latent process can be a key modelling assumption for GMRF models because they are not always discretisation invariant. To model spatial or spatio-temporal data, the process can

be centred around a mean value μ_θ comprising of a linear expansion of explanatory variables, as recently done for crime data in [MHCC18]. In the machine learning literature, μ_θ has also been taken to be a low-rank approximation of a GP as part of a multiscale approach [VV08]. For both of these examples, the latent process describes spatially correlated noise with a relatively short length scale. It is usual to assume that the observation noise is centred and independent and identically distributed (iid) Gaussian.

For Bayesian inference, the required likelihood $p(y|\theta)$ is available in closed form. By letting $p(e|\theta) = p(y - Ax|\theta)$ and $p(x|\theta)$ denote the Gaussian densities of $e|\theta$ and $x|\theta$, respectively, the marginal likelihood $p(y|\theta)$ is obtained as follows:

$$p(y|\theta) = \int p(y - Ax|\theta)p(x|\theta)dx, \quad (4.14)$$

$$= \int N(y|Ax + \delta_\theta, \Gamma_\theta)N(x|\mu_\theta, \Sigma_\theta)dx, \quad (4.15)$$

$$= N(y|A\mu_\theta + \delta_\theta, S_\theta), \quad (4.16)$$

where the covariance matrix S_θ is given by

$$S_\theta = \Gamma_\theta + A\Sigma_\theta A^T. \quad (4.17)$$

To infer the model parameters θ conditioned on observation data y , a prior is specified for the model parameters $p(\theta)$ that is reweighted according to the likelihood $p(y|\theta)$ using Bayes' theorem:

$$p(\theta|y) \propto p(y|\theta)p(\theta). \quad (4.18)$$

A single evaluation of the posterior density requires a matrix inverse and a determinant of S_θ . Note that even if the likelihood is only needed up to normalisation constant, the determinant term cannot be ignored as it depends on θ . If the determinant term cannot be evaluated, then it is a doubly-intractable inference problem that can be approached using the ideas in Section 2.4.

Whilst non-Gaussian likelihoods can be more appropriate, such as when modelling categorical or count data, the likelihood $p(y|\theta)$ is no longer available in closed. Only the more tractable Gaussian case is studied here to allow for a fair comparison of methods to be performed. For similar reasons, μ_θ and δ_θ are assumed to be constant in all examples.

4.3 Computational aspects

Despite the tractability of linear Gaussian models, there are scalability issues due to the required matrix operations. This section reviews direct and iterative methods for the required matrix operations.

4.3.1 Direct methods

For dense matrices, direct methods require $O(n^2)$ storage and $O(n^3)$ operations, where n is the dimension number. For example, one usually computes the Cholesky factorisation $A = LL^T$, where L is a lower triangular Cholesky factor whose entries are

$$L_{jj} = \sqrt{A_{jj} - \sum_{k=1}^{j-1} L_{jk}^2}, \quad (4.19)$$

$$L_{ij} = \frac{1}{L_{ij}} \left(A_{ij} - \sum_{k=1}^{j-1} L_{ik} L_{jk} \right), \quad i > j. \quad (4.20)$$

This factorisation allows any linear system of the form $A^{-1}x = b$ to be solved via back-substitution in $O(n^2)$ operations. The log-determinant can also be obtained in $O(n)$ by summing over the logarithm of the diagonal:

$$\log |A| = 2 \sum_{j=1}^n \log(L_{jj}).$$

Currently standard computers struggle to hold dense Cholesky factors in memory from around $n=10,000$.

In special cases, the Cholesky factorisation of a sparse matrix is also sparse. For example, when a matrix is band diagonal with bandwidth b , the Cholesky factorisation also has bandwidth b and requires $O(nb)$ memory to store and $O(nb^2)$ operations to compute [GDGF10]. In many cases, the same level of sparsity is not preserved and the Cholesky factorisation can be impossible to store due to the storage requirements. This is known as the fill-in effect, and is illustrated in Fig. 4.1. It is possible to reduce the extent of the fill-in effect via a permutation matrix that reorders the indices to obtain a matrix with a lower bandwidth [Saa03, GDGF10]. For example, let P be a square permutation matrices such that $P^T P = I$. Each row and column of P only has a single non-zero entry which is equal to one. Then the linear system $Ax = b$ can be

rewritten as

$$\underbrace{(PAP^T)}_{\tilde{A}} \underbrace{(Px)}_{\tilde{x}} = \underbrace{(Pb)}_{\tilde{b}}, \quad (4.21)$$

where the bandwidth of $\tilde{A}$ is much less than the bandwidth of A . By solving $\tilde{A}\tilde{x} = \tilde{b}$, one can obtain the solution x by $x = P^T\tilde{x}$. As there are $n!$ different permutation matrices to choose from, finding a suitable permutation matrix can be a non-trivial task. Whilst the fill-in effect can be reduced, it can remain a problem to some extent. In this Chapter, a new approach is suggested for if the fill-in effect, after reordering, is so great that the Cholesky factorisation cannot be stored in memory. This was also the motivation for the work in [Aun12].

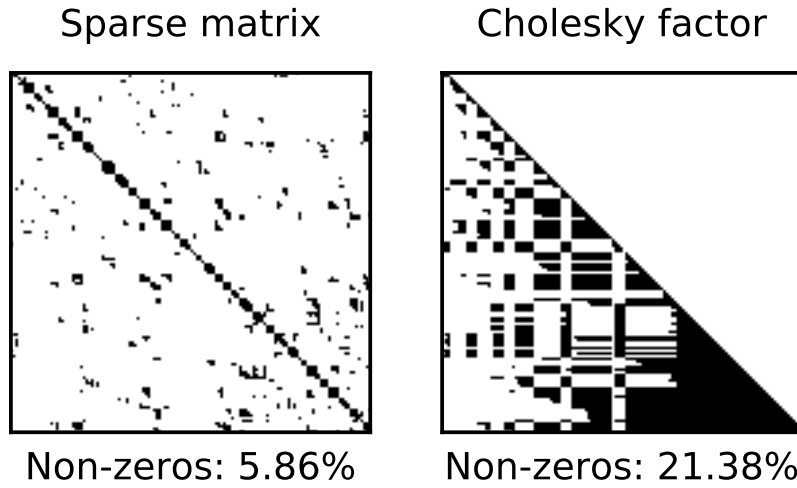


Figure 4.1: An illustration of the so-called ‘fill-in’ effect for a Cholesky decomposition. The sparsity pattern of a sparse matrix (left) and the lower Cholesky factor of the same matrix (right).

4.3.2 Iterative methods and rational approximations

It can be possible to solve sparse linear systems of the form $Ax = b$ via iterative methods, even when they cannot be approached by direct methods due to the number of operations or storage required [NW06]. For example, the conjugate gradient method provides a sequence of improving approximate solutions $x_1, \dots, x_m$ with the property that

$$x_k = \underset{x \in \mathcal{K}_k(A, b)}{\operatorname{argmin}} \|x - A^{-1}b\|_A^2, \quad (4.22)$$

where $\|x\|_A = \|A^{1/2}x\|$ and $\mathcal{K}_k(A, b) = \operatorname{span}\{b, Ab, \dots, A^{k-1}b\}$. An exact solution can be found when A is full rank for $k = n$. In practice, the system is never solved exactly due to numerical

error. The conjugate gradient method exploits the fact that A is a symmetric positive definite matrix to provide a new search direction that is orthogonal to the previous ones. The dominant cost at each iteration is the matrix-vector product that provides the new search direction. As a result, the algorithm requires at most $O(kn^2)$ operations, and much less when the cost of a matrix-vector product is significantly reduced due to the sparsity of A . It is also necessary that A is well-conditioned for the conjugate gradient method to converge to within an acceptable tolerance.

It is also necessary to consider iterative methods for the determinant term. The matrix function is defined by a diagonal matrix D by the diagonal matrix with $f(D)_{ii} = f(D_{ii})$. The definition of a matrix function extends to a symmetric matrix A , which is diagonalised by the orthogonal matrix U , by

$$f(A) = Uf(D)U^T. \quad (4.23)$$

Fortunately explicit diagonalisation of the matrix A is not necessary, as a polynomial or rational approximation can be used instead [Ken04]. A rational approximation of a matrix function applied to a vector b can be written as a series of shifted linear system solves

$$f(A)b \approx \sum_{i=1}^N \alpha_i (A + \sigma_i I)^{-1} b, \quad (4.24)$$

for suitably chosen weights $\{\alpha_i\}_{i=1}^N$ and shifts $\{\sigma_i\}_{i=1}^N$. Equation (4.24) can be evaluated via iterative methods, for example, with multiple runs of the conjugate gradient method. It turns out that a number of shifted linear systems can be solved at approximately the cost of solving a single linear system using a multi-shift variant of the conjugate gradient method [Jeg96, ASE14]. Alternatively, the systems can be solved in parallel.

Matrix determinant

When A is a symmetric positive definite matrix, it can be shown that

$$\ln |A| = \text{trace}(\ln A), \quad (4.25)$$

by using (4.23). The following Monte Carlo approximation can be used:

$$\text{trace}(\ln A) = \sum_{j=1}^N e_j^T (\ln A) e_j \approx \frac{1}{S} \sum_{j=1}^S z_j^T (\ln A) z_j, \quad (4.26)$$

where e_j is the unit vector in the j th direction and $z_j \in \{-1, 1\}^n$ is a vector of Rademacher random variables [Hut90, BFG96]. The Monte Carlo approximation of the log-determinant has recently been used to obtain point estimates of the hyperparameters for large-scale GPs in [DEN⁺17] and for GMRFs in [Aun12, ASE14]. The exponentiation of the approximation that would be required for the Metropolis-Hastings algorithm gives a biased estimate of the determinant. This led the authors of [LGA⁺15] to use the Russian roulette framework described in Section 2.4.2, however, the scheme is complex and sensitive to the variance in the Monte Carlo approximation of the log-determinant.

Matrix inverse square root

Using the MCMC methods described in Section 2.4.1, it is possible to specify a Metropolis-Hastings algorithm that avoids the determinant computations altogether, provided that it is possible to simulate from the likelihood model. This is possible by applying a rational approximation of the inverse-matrix-square-root applied to a vector. This is because if z solves

$$A^{-1/2}z = w, \quad w \sim N(0, I), \quad (4.27)$$

then $Az \sim N(0, A)$.

One way to obtain a rational approximation is to use quadrature method to approximate the Cauchy integral formula for a matrix function [HHT08, HL13]. It is assumed that the function is analytic on and inside the contour C that encloses the spectrum of A . The formula is given by

$$f(A) = \frac{1}{2\pi i} \int_C f(z)(zI - A)^{-1} dz. \quad (4.28)$$

The contour C must be carefully chosen so that a good approximation is obtained with relatively few quadrature nodes. For the function $f(z) = z^{-1/2}$, the change of variables $z = \omega^2$ yields

$$A^{-1/2} = -\frac{1}{\pi i} \int_{C'} (A - \omega^2 I)^{-1} d\omega. \quad (4.29)$$

Since A is positive definite and its eigenvalues are contained inside of an interval $[m, M] \subset \mathbb{R}_{>0}$, it is only necessary to integrate over the complex plane without passing through the origin. Following [HHT08], the quadrature points can be chosen efficiently with the following conformal map $\omega(t) = m^{1/2} \text{sn}(t|m/M)$ defined in terms of the Jacobi elliptic function sn . The rational approximation, with reference to (4.24), is then given by:

$$\alpha_i = \frac{2K' m^{1/2} \text{cn}(t_i) \text{dn}(t_i)}{\pi N}, \quad \sigma_i = \omega(t_i)^2, \quad (4.30)$$

where $t_j = i(j - 1/2)K'/N$ and $1 \leq j \leq N$, and cn and dn are further Jacobi elliptic functions. Remarkably, this choice of approximation coincides with the optimal¹ rational approximation discovered by Zolotarev. Crucially, the Frobenius norm or the error in estimating $A^{-1/2}$ decays rapidly as

$$O\left(\exp\left(\frac{-2\pi^2 N}{\ln(M/m) + 3}\right)\right),$$

and the number of quadrature points N can be chosen to reach the desired level of accuracy. The contour C is chosen to be the smallest region that contains the spectrum of A . From preliminary studies, it was found that $N \approx 20$ quadrature points were needed to reach floating point precision when $m=10^{-6}$ and $M=10^6$.

4.4 Methodology

The ideas introduced in the previous sections are used to provide a new Metropolis-Hastings algorithm. There are two cases to consider: when the covariance matrix of the latent process Σ_θ is sparse, and when the precision matrix of the latent process $\Lambda_\theta = \Sigma_\theta^{-1}$ is sparse.

It is possible to apply the auxiliary variable methods described in Section 2.4.1 to avoid the need to estimate the determinant term. This is due to a careful choice of proposal that requires simulating from the model. Instead, a number of auxiliary variables can be introduced to remove

¹Optimal in the sense of minimising worse-case error.

the determinant term altogether. This is less restrictive because then any proposal can be used. The following auxiliary variables are introduced:

$$z|\theta \sim \mathcal{N}(0, S_\theta^{-1}), \quad (4.31)$$

where S_θ is as defined in (4.17). The joint distribution over the augmented state space is

$$p(\theta, z|y) \propto p(z|\theta)p(y|\theta)p(\theta) \propto p(\theta)e^{-\frac{1}{2}(y-A\mu_\theta)^T S_\theta^{-1}(y-A\mu_\theta) - \frac{1}{2}z^T S_\theta z}, \quad (4.32)$$

and the determinant terms cancel. That is, Equation (4.32) is left with only the quadratic forms for the latent and auxiliary Gaussian variables. The original posterior in (4.18) is restored by marginalising over z ,

$$p(\theta|y) = \int p(\theta, z|y) dz.$$

In other words, one can sample from the augmented distribution and subsequently discard the auxiliary variables to be left with samples from the target distribution.

A block Gibbs scheme is proposed as it is difficult to update both the latent variables and model parameters together. The two steps are:

1. Update the auxiliary variables $z|y, \theta$ using a Metropolis-Hastings transition kernel that targets the full conditional $p(z|y, \theta) \propto e^{-\frac{1}{2}z^T S_\theta z}$.
2. Update the model parameters $\theta|y, z$ using a Metropolis-Hastings transition kernel that targets the full conditional $p(\theta|y, z) \propto p(\theta)e^{-\frac{1}{2}(y-A\mu_\theta)^T S_\theta^{-1}(y-A\mu_\theta) - \frac{1}{2}z^T S_\theta z}$.

These updates are considered in more detail in the proceeding subsections.

4.4.1 Model parameter updates

The model parameters are updated with standard Metropolis-Hastings (MH) transition kernels. As gradient evaluations with respect to θ are generally prohibitively expensive, a random walk proposal with an appropriately tuned covariance matrix is an obvious choice. The covariance of the posterior $p(\theta|y)$ is estimated empirically from a preliminary run, and is rescaled to achieve an acceptance rate in the range of 20–40% [Ros11].

4.4.2 Auxiliary variable updates

As gradient evaluations with respect to z can be obtained by solving linear systems, it is possible to update the auxiliary variables using MALA or HMC transition kernels. As described in Section 2.2.2, the effectiveness of HMC depends on the step size and the number of steps for the leapfrog integrator. It may be beneficial to apply a change of variables, for example, as done for the Gibbs sampler in Section 2.2.1, to improve the mixing of HMC if the auxiliary variables are strongly correlated [Nea11]. Optimal tuning can be costly and difficult to achieve in practice.

Instead, a Gibbs update can be performed by resampling the auxiliary variables from their full conditional $z|\theta \sim \mathcal{N}(0, S_\theta^{-1})$. The iterative methods and rational approximations described in Section 4.3.2 can be used to resample the auxiliary variables, which only require matrix-vector products. In particular, it is possible to sample from high-dimensional Gaussian distributions as long as it is possible to apply the covariance or precision matrix to a vector, and that the matrix is reasonably well-conditioned. Two strategies are specified depending on whether one has direct access either to a sparse covariance matrix Σ_θ or sparse precision matrix $\Lambda_\theta = \Sigma_\theta^{-1}$.

Sparse latent covariance

For the case where one has direct access to a sparse latent covariance matrix Σ_θ , it is possible to evaluate $S_\theta^{-1/2}w$ in its current form, where w is a vector of Gaussian white noise, because $S_\theta = \tau^{-1} + A\Sigma_\theta A^T$ is sparse. The following two steps are performed to update the auxiliary variables:

1. Draw $w \sim \mathcal{N}(0, I)$.
2. Set $z = S_\theta^{-1/2}w$.

Sparse latent precision

The case where one has access to a sparse latent precision matrix Λ_θ requires an alternative approach because $S_\theta = \tau^{-1} + A\Lambda_\theta^{-1}A^T$ is non-sparse. Instead, it is possible to draw a random variable $\tilde{y} \sim N(0, S_\theta)$, then solve the linear system $S_\theta z = \tilde{y}$ so that one obtains a random

variable z whose covariance is equal to the precision of the model $z \sim N(0, S_\theta^{-1})$. To avoid working with non-sparse matrices, it is necessary to use the matrix inversion lemma:

$$S_\theta^{-1} \tilde{y} = \tau \tilde{y} - \tau^2 A(\Lambda_\theta + \tau A^T A)^{-1} A^T \tilde{y},$$

which only requires an additional linear solve. In summary, to draw an auxiliary variable given a sparse latent precision, the following three steps are performed:

1. Draw $w' \sim \mathcal{N}(0, I)$ and $\tilde{\epsilon} \sim \mathcal{N}(0, \tau^{-1} I)$.
2. Set $\tilde{x} = \Lambda_\theta^{-\frac{1}{2}} w'$ and $\tilde{y} = A\tilde{x} + \tilde{\epsilon}$.
3. Set $z = \tau \tilde{y} - \tau^2 A(\Lambda_\theta + \tau A^T A)^{-1} A^T \tilde{y}$.

Note that in both cases, the proposed scheme requires fewer linear system solves than the comparable MCMC scheme described by (2.66)-(2.67).

4.5 Numerical results

In this section the proposed methodology is applied to three selected linear Gaussian models: random pattern precision matrices, GPs with sparse covariance and GMRFs with sparse precisions. These examples have been selected to show the capabilities and limitations of the proposed method.

4.5.1 Random pattern precision matrices

This example is for a model that has a well-conditioned covariance matrix with an underlying sparse structure. This way, certain matrix-vector products and linear system solves are inexpensive operations that can be performed to a desired level of accuracy. The following single-parameter model is assumed:

$$y|\theta \sim \mathcal{N}(0, S_\theta), \tag{4.33}$$

$$S_\theta = (\gamma^{-1} Q + \gamma I)^{-1}, \tag{4.34}$$

where Q is a randomly generated symmetric and positive definite matrix that has a condition number close to 1 and approximately $3n$ non-zero elements. Synthetic data is generated for $n = 10^3, 10^4, 10^5$ and 10^6 . The sparsity pattern for $n = 10^3$ is illustrated in Fig. 4.1. The value of $\ln \gamma = -3$ is used to generate the data for this example. A log-uniform prior is specified for γ .

For each example, a standard MCMC algorithm (standard MCMC) and the determinant-free approach (determinant-free MCMC) were used to simulate from the posterior of $\ln \gamma$. Markov chains of length 10,000 are run for each example, after tuning the random walk proposal to achieve an acceptance rate close to 20%. The chains are initialised near the values of $\ln \gamma$ used to generate the data. For the determinant-free MCMC, 20 terms are used in the rational approximation and the multi-shift conjugate gradient solver is used to solve the linear systems. The covariance matrices in this example are well-conditioned, so the conjugate gradient method converges rapidly without the help of a preconditioner.

A comparison is performed between standard MCMC and the determinant-free method. The MCMC output is summarised in Table. 4.1. It is seen that, in all cases, standard MCMC produces more independent samples for a fixed number of iterations, as measured by the effective sample size (ESS). This is to be expected as the coupling between the model parameters and auxiliary variables cause the Markov chain to mix slower. The standard MCMC method requires more time per iteration, because of the cost of evaluating the Cholesky decomposition. Determinant-free MCMC scales better with the matrix size and was found to produce more effective samples per unit time than standard MCMC for $n = 10^4$. For $n \geq 10^5$ and beyond, a standard laptop does not have enough memory to run the Cholesky based scheme. Of greatest importance is that determinant-free MCMC allows much larger problems to be approached when compared to standard MCMC.

4.5.2 Sparse covariances for spatial modelling of anti-social crime data

The proposed approach is applied to a model that has a sparse covariance matrix. For this example, a real dataset is used to infer the spatial distribution of anti-social criminal activity, in London (UK), using count data obtained from the UK government website [Ser16]. Log-crime-rates were calculated for each lower layer super output area (LSOA), defined as the number

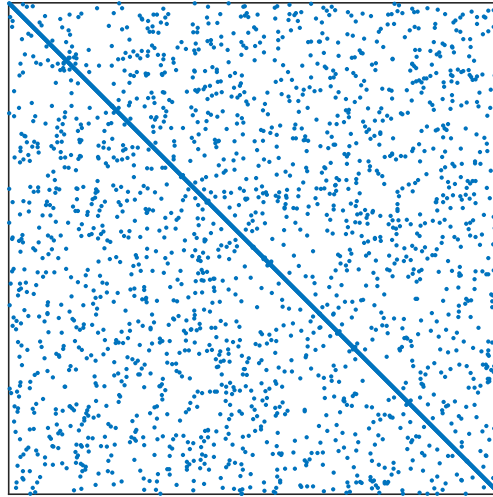


Figure 4.2: The sparsity pattern of a well-conditioned randomly generated $10^3 \times 10^3$ matrix. Each shaded region represents a non-zero matrix entry, of which there are 2996.

Matrix size	10^3	10^4	10^5	10^6
Standard MCMC				
ESS	1778	1532	NA	NA
Time (s)	2.56×10^2	4.22×10^4	NA	NA
ESS/Time (s^{-1})	6.9025	0.0363	NA	NA
Determinant-free MCMC				
ESS	793	679	613	1272
Time (s)	1.04×10^2	1.67×10^3	1.94×10^3	2.28×10^4
ESS/Time (s^{-1})	7.6189	0.4055	0.3160	0.0558

Table 4.1: Comparison of MCMC efficiency between standard MCMC and determinant-free MCMC for the random pattern precision matrices example. The Markov chains are run for a length of 10,000.

Parameter	σ^2	l	τ
Standard MCMC			
Mean	0.13	0.06	2.11
Standard deviation ($\times 10^{-2}$)	1.60	0.57	4.91
ESS	868	930	956
Time ($s \times 10^4$)	1.59	1.59	1.59
ESS/time ($s^{-1} \times 10^{-2}$)	5.48	5.87	6.03
Determinant-free MCMC			
Mean	0.16	0.06	2.10
Standard deviation ($\times 10^{-2}$)	2.33	0.54	4.61
ESS	213	242	340
Time ($\times 10^4$)	1.24	1.24	1.24
ESS/time ($s^{-1} \times 10^{-2}$)	1.72	1.96	2.75

Table 4.2: Comparison of efficiency of standard MCMC and determinant-free MCMC for the anti-social crime dataset. The Markov chains are run for a length of 10,000.

of crimes divided by the LSOA population. The location of each crime-rate is taken to be the LSOA centroid location. The preprocessed dataset is comprised of 4,826 log-crime-rates, one for each LSOA, which is illustrated in Fig. 4.2 (top).

Returning to the linear Gaussian model in (4.11), the choice of mean function μ_θ is given by a linear expansion of radial basis functions:

$$\mu(s) = \beta_0 + \sum_{h=1}^5 \beta_h \exp \left\{ -\frac{1}{2\sigma_h^2} |s - s_h|^2 \right\}, \quad (4.35)$$

which is evaluated at each observation location. Here, the constant β_0 represents the background log-crime-rate. The radial basis functions capture the general trend of higher crime towards more central areas. As the focus is to perform a comparison between standard MCMC algorithm and determinant-free MCMC, the coefficients β_h are held fixed to values obtained from a linear regression. If Gaussian priors were placed on each coefficient β_h , then (4.35) would approximate a squared-exponential GP with a slow correlation decay. The radial basis functions are centred at locations s_h as determined by the k-means clustering algorithm and the length scale parameter σ_h is set equal to the smallest pairwise distance between the s_h . Crime tends to be spatially correlated with localised ‘hot-spots’. A compactly supported Wendland kernel is used to model the localised behaviour that is not captured by the global mean function [RW06]. For this example, the covariance function specified by (4.10) is used. Weakly informative independent

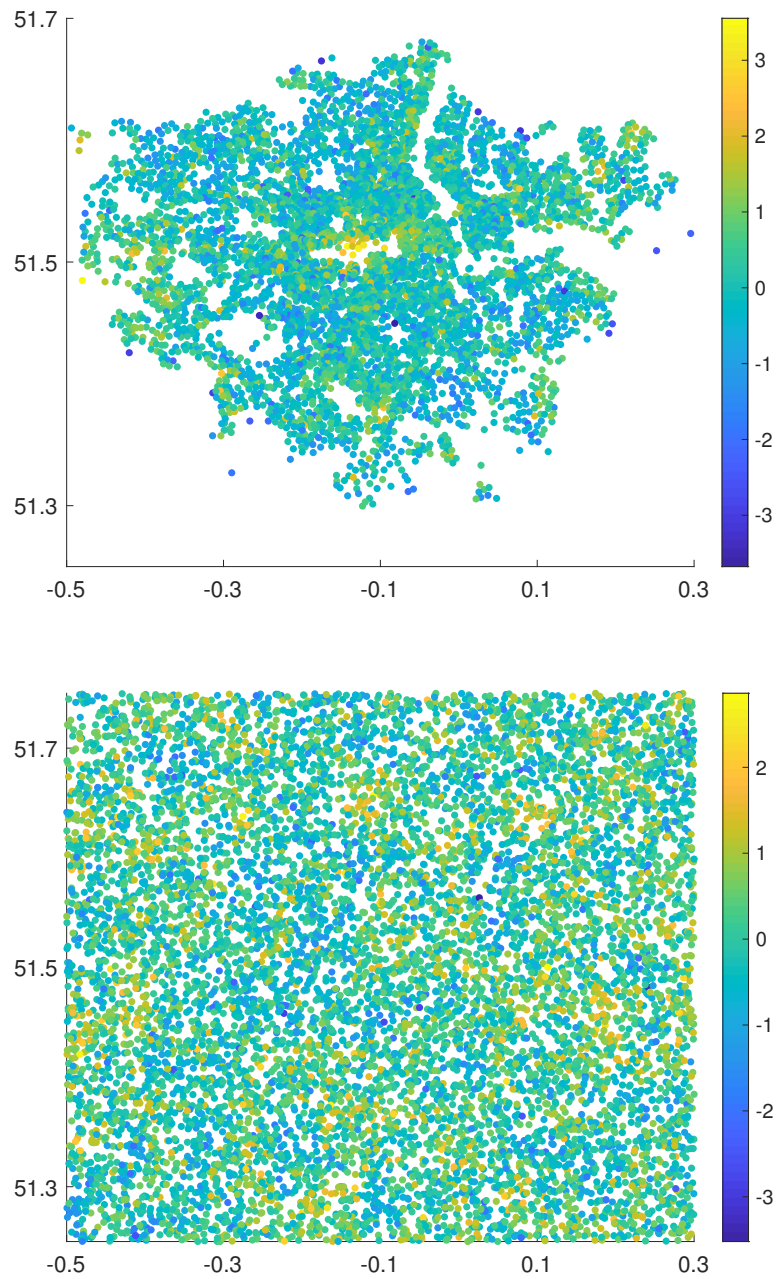


Figure 4.3: The crime dataset for example 1 comprising of 4,826 observations, after subtracting a linear model (top). The synthetic dataset for example 1 comprising of 10,000 observations (bottom). For the crime dataset, each point is shaded according to the adjusted log-crime-rate for anti-social crimes and is plotted at the centroid location of the respective LSOA.

Parameter	σ^2	l	τ
Standard MCMC			
Mean	0.13	0.06	2.11
Standard deviation ($\times 10^{-2}$)	0.91	0.25	3.01
ESS	243	184	220
Time ($s \times 10^4$)	3.21	3.21	3.21
ESS/Time ($s^{-1} \times 10^{-2}$)	0.76	0.57	0.68
Determinant-free MCMC			
Mean	0.13	0.06	2.11
Standard deviation ($\times 10^{-2}$)	0.94	0.23	3.31
ESS	82	172	84
Time ($\times 10^4$)	0.66	0.66	0.66
ESS/time ($s^{-1} \times 10^{-2}$)	1.25	2.48	1.27

Table 4.3: Comparison of efficiency of standard MCMC and determinant-free MCMC for the synthetic dataset for the spatial modelling of crime example. The Markov chains are run for a length of 2,000.

log-normal priors are placed on the unknown parameters $\theta = (\sigma^2, l, \tau)$.

The posterior distribution is simulated from using a standard MCMC algorithm and determinant-free MCMC. Markov chains are run for 10,000 iterations. The covariance matrix for the random walk proposal is set to the empirical covariance obtained from a preliminary run. An appropriate rescaling of the covariance ensures that the acceptance rate is around 20%-40%. The MCMC output is summarised in Fig. 4.2. As before, the standard MCMC approach is slower per iteration but produces more independent samples per iteration, as measured by the ESS. Due to the relatively small size of the dataset, the standard MCMC approach here produces slightly more independent samples per unit time. Whilst the determinant-free MCMC runs faster, it mixes slower due to the coupling between the auxiliary variables and model parameters.

It is expected that the determinant-free method will provide more effective samples per unit time for a similar system but with more observations gathered over a greater spatial extent. To test the idea, $n = 10^4$ data locations were drawn from the uniform distribution over $[-0.5, 0.3] \times [51.25, 51.75]$ and drawn from a similar model with the covariance parameters set to their previously inferred posterior means. The synthetic dataset is visualised in Fig. 4.3 (bottom). The above procedures were performed for a chains of length 2,000. The MCMC output is summarised in Fig. 4.3. It can be seen that for the larger dataset, the determinant-free method ran 4.9 times faster and produced more effective samples per unit time for each sampled

parameter.

4.5.3 Gaussian Markov random field models specified by a whitening matrix

Lastly, the proposed approach is applied to a case where the model is specified by a sparse precision matrix Λ_θ and is a GMRF that is noisily observed via a linear function. The covariance of the marginal likelihood in (4.16) is therefore given by

$$S_\theta = \tau^{-1}I + A\Lambda_\theta^{-1}A^T, \quad (4.36)$$

which is typically non-sparse. To express the marginal likelihood in terms of sparse matrices only, the matrix inversion lemma is used to obtain:

$$\begin{aligned} \ln p(y|\theta) = & \frac{1}{2} \left\{ \ln |\Lambda_\theta| + n \ln \tau - \ln |\Lambda_\theta + \tau A^T A| \right. \\ & \left. - \tau y^T y + \tau^2 y^T A (\Lambda_\theta + \tau A^T A)^{-1} A^T y \right\} \\ & + \text{const.} \end{aligned} \quad (4.37)$$

For standard MCMC, the log-determinants in (4.37) and linear system solve is evaluated with 2 sparse Cholesky decompositions. This is compared against the determinant-free MCMC scheme.

This example studies a poorly-conditioned system where the iterative methods are suspected to run into difficulty. Here, the precision matrix Λ_θ specifies a smooth latent process that is derived from discrete Laplacian

$$L = \frac{1}{m^2} \begin{bmatrix} 2\delta & 0 & & & \\ -1 & 2 & -1 & & \\ & -1 & 2 & -1 & \\ & & \ddots & \ddots & \ddots \\ & & & -1 & 2 & -1 \\ & & & & 0 & 2\delta \end{bmatrix},$$

by $\Lambda_\theta = \frac{1}{\gamma^2} L^T L$. Here δ is fixed to provide the stochastic Dirichlet boundary conditions described in [CKS06], so that the prior variance is approximately equal across the domain. With reference

to (4.7), the solution to the linear system

$$Lx = \gamma w, \quad w \sim N(0, I), \quad (4.38)$$

approximates the Matérn 3/2 GP when it has the same boundary conditions and in the limit $l \rightarrow \infty$. It is therefore unsurprising that the precision matrix Λ_θ becomes poorly conditioned as the dimension number increases; the associated linear systems become more constrained. In fact, the condition number of the one-dimensional discrete Laplacian scales like $O(m^2)$ [Dem97].

For this example, a synthetic dataset of size $n = 12,000$ is generated as follows:

$$y_i = \sin(4s_i) + 4/5 \sin(8s_i) + \epsilon_i, \quad (4.39)$$

$$s_i \sim U(0, 1), \quad (4.40)$$

$$\epsilon_i \sim N(0, 0.2^2). \quad (4.41)$$

The dataset is illustrated in Fig. 4.4. The latent process is defined on a grid with $m = 10,000$ elements and is a matrix that performs linear interpolation between the grid nodes. For this example, all linear systems can be solved using a banded solver in linear time. There is also an explicit factorisation of the precision matrix Λ_θ , which allows for an assessment of the proposed method in the case that the latent precision is poorly conditioned. Inference is performed for the parameters $\theta = \{\gamma, \tau\}$ using log-uniform priors for each.

The preliminary runs revealed that Λ_θ is very poorly conditioned for just moderately sized m . For example, the condition number of the latent process described above is 3.7×10^{15} . Note that conditioning is more of an issue for models specified by their precision, because iterative methods must be performed directly on the precision matrix and the observation noise is not added to the diagonal. To demonstrate the severity of the problem, the number of conjugate gradient iterations against the matrix size is shown in Fig. 4.5, with and without preconditioning. In the case that an incomplete Cholesky factorisation is available, the conjugate gradient method converges almost immediately. To address the concerns over conditioning, each linear system in the rational approximation in (4.24) is solved separately with its own incomplete Cholesky decomposition preconditioner. Note that this preconditioning strategy is not possible with a multi-shift solver. Indeed, the increased computational costs of solving the N non-interacting linear systems in the rational approximation in (4.24) can be side-stepped via parallelisation.

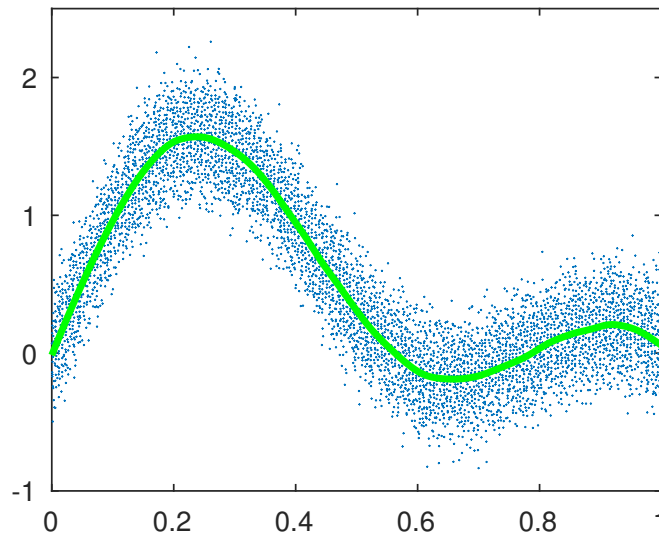


Figure 4.4: 12,000 data points, for example 3, which was generated from the function $f(x) = \sin(4x) + 4/5 \sin(8x)$ plus centred Gaussian noise. The green curve indicates the predictive mean for the model when the parameters are fixed at their maximum likelihood values.

If the conditioning of the original matrix is poor, it may not be possible to find a suitable preconditioner, for example, from $m = 3 \times 10^4$ onwards standard routines in MATLAB failed to compute the incomplete Cholesky due to numerical underflows.

A comparison is performed for four different approaches to sample from the marginal posterior of this model: MCMC with determinant computations, determinant-free MCMC without using rational approximations (by exploiting the fact that L is known), and determinant-free MCMC using rational approximations with 10 and 20 terms, respectively. The Markov chains are run for a length of 10,000 and the MCMC output is presented in Table 4.4. It can be seen that the memory intensive Cholesky-based approach outperforms the determinant-free approach. This is unsurprising, given that the band-diagonal structure of the precision is preserved in the sparse Cholesky factors. This is generally not the case, especially for more complex GMRF models, as already seen for the random pattern precision matrices example. Whilst the determinant-free MCMC suffers from the usual reduction in ESS, it appears to be of a similar magnitude to that observed for the well-conditioned examples. The mixing properties of the scheme do not appear to be significantly affected by the conditioning of the system, at least in the presence of observation noise. It can be seen that the method appears to perform just as well with 10 terms in the rational approximation.

Parameter	$\ln \tau$	$\ln \gamma$
Standard MCMC		
Mean	3.213	-12.73
Standard deviation	0.009	0.105
ESS	1111	980
Time (s)	103	103
ESS/Time (s^{-1})	10.8	9.515
Det-free MCMC - no rational approximations		
Mean	3.217	-12.49
Standard deviation	0.013	0.175
ESS	582	597
Time (s)	373	373
ESS/Time (s^{-1})	1.562	1.603
Det-free MCMC - with rational approximations (10 terms)		
Mean	3.230	-12.75
Standard deviation	0.013	0.145
ESS	664	703
Time (s)	1199	1199
ESS/Time (s^{-1})	0.554	0.586
Det-free MCMC - with rational approximations (20 terms)		
Mean	3.211	-12.71
Standard deviation	0.013	0.151
ESS	600	663
Time (s)	2279	2279
ESS/Time (s^{-1})	0.263	0.291

Table 4.4: A comparison of different MCMC schemes for the Gaussian Markov random field example. The comparison includes MCMC with determinants; MCMC without determinants and without rational approximations, MCMC without determinants and with rational approximations with 10 and 20 terms, respectively. Each column shows the MCMC output for the respective parameter. Each Markov chains is run for a length of 10,000. It can be seen that the standard Cholesky-based approach outperforms the determinant-free method due to the band-diagonal structure of the precision matrix.

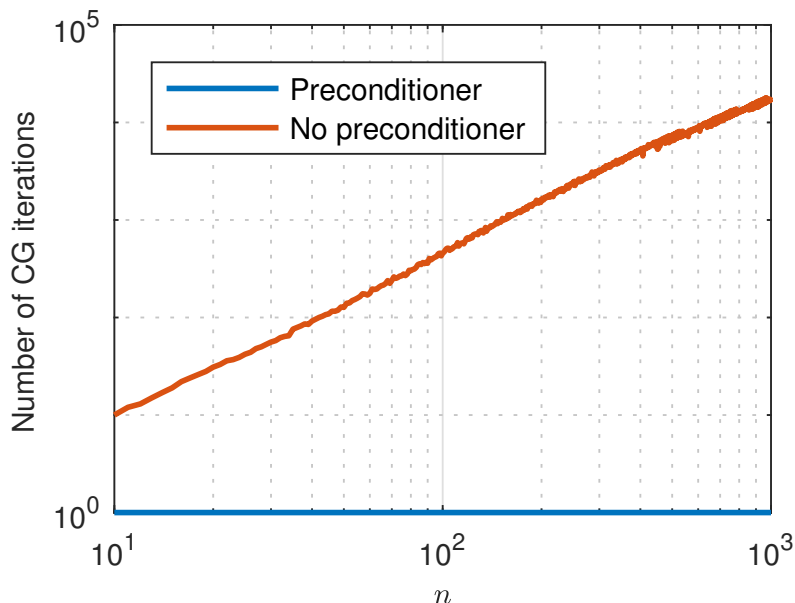


Figure 4.5: Matrix size n against the number of conjugate gradient iterations until convergence (within a tolerance of 1×10^{-5}) to solve the first term in the rational approximation of $(L^T L)^{-1/2}w$, with no preconditioning and using an incomplete Cholesky preconditioner.

4.6 Discussion and conclusions

A new MCMC scheme has been proposed and investigated for asymptotically exact Bayesian inference in high-dimensional Gaussian models. The proposed scheme is a particular implementation of the Metropolis-Hastings algorithm, and it has been shown that auxiliary random variables can be used to replace the determinant computation that arises in the Metropolis-Hastings acceptance probability calculation when the covariance matrix is specified by unknown model parameters. It was found that the Markov chains for the proposed method mix more slowly per iteration than for a standard MCMC Cholesky-based approach, due to the additional auxiliary variables, but can be much cheaper per iteration. The new method can be faster because it exploits fast matrix-vector operations and it never creates dense matrices. It can be said that the newly proposed approach allows for MCMC methods to be used for Bayesian inference for a larger class of linear Gaussian models; these are models whose Cholesky-factors cannot be obtained, yet the system is well-conditioned enough for matrix-vector products to be performed reliably.

In practice, latent functions require a degree of smoothness, which forces the smallest eigenvalue of the covariance matrix to become close to zero as the system size increases. This means that

large systems are poorly-conditioned. The irony here is that a system being poorly conditioned implies that it is almost low rank. Low rank systems are more constrained than an arbitrary process, and so should be easier to represent. For the crime example, some of the smoothness was removed from the process so that the GP described coloured noise with a relatively short length scale. As a result, the model had a well-conditioned covariance matrix. It is concluded that such a decomposition may be necessary to make full use of iterative methods, as these methods do not work well with extremely poorly-conditioned precision matrices as found in larger GMRF models.

For future work, it remains to identify ways to exploit low rank structure more generally. A careful model specification is required so that the Gaussian model is reasonably conditioned so that iterative methods can be used. The proposed scheme may prove useful in settings where exact simulation-based inference is currently impossible. It also remains to use the methodology for parameter estimation in more complex SPDE models that have a Gaussian representation. It may also be beneficial to use the augmented distribution in (4.32) to obtain more efficient variational approximations of the target distribution.

Chapter 5

Uncertainty quantification in urban simulations

The building of mathematical and computer models of cities has a long history. The core elements are models of flows (spatial interaction) and the dynamics of structural evolution. Standard practice has been to calibrate the spatial interaction models independently and to explore the dynamics through simulation.

In this chapter, a stochastic model of urban structure is developed to formally account for uncertainty arising from less predictable events. There are three significant contributions that are presented in this section. First, it is shown that the structural variables can be represented by a single potential function and a system of stochastic differential equations (SDEs) is proposed to model their evolution. Second, it is shown that the stationary distribution of the SDEs can be viewed as a probabilistic model of urban structure due to a connection with entropy maximisation. Third, it is shown that the parameters of the spatial interaction model can be estimated from the structure alone, in lieu of flow data, using the Bayesian inferential framework. The posterior distribution is doubly-intractable and poses significant computational challenges that are overcome using Markov chain Monte Carlo (MCMC) methods. It is demonstrated that the SDE model can be used to obtain uncertainty estimates for the evolving structural variables. The new methodology is demonstrated with a case study of the London retail system. Further work is suggested to improve the SDE model and to apply the methodology to larger-scale systems.

The work in this chapter was performed in collaboration with Mark Girolami, Grigorios Pavliotis

and Alan Wilson, and parts have been published in [EGPW18].

5.1 Context and literature review

The task of understanding the inner workings of cities and regions is a major challenge for contemporary science [Bat08, Bat13]. The key features of cities and regions are activities at locations, flows between location and the structures that facilitates these activities [DW15]. It is understood that cities and regions are complex systems, and that an emergent structure arises from the actions of many interacting individuals. The flows between locations arise from the choices of these individuals, and an understanding of the choice mechanism is important for planning and decision making. Economists have long supported the idea that consumer choices are derived from utility, a measure of net benefit, although preferences can only be measured indirectly by the phenomena they give rise to [Mar20].

The flows between any two locations concern a vast number of individuals and can be represented by averaging using Boltzmann's ideas from statistical mechanics [Wil67]. The resulting models are known as spatial interaction models in the urban geography literature. Economists later provided their own explanation, which describe a discrete choice mechanism using the idea of random utility maximisation [McF73]. These models have received considerable attention in the econometrics literature. Unfortunately, the direct calibration of spatial interaction models requires large volumes of flow data that is difficult to obtain, and may require an extensive survey that suffers from sampling biases.

On the other hand, the urban structure facilitating activities can be straight-forward to measure. The evolution of urban structure can be described by a system of coupled first-order ordinary differential equations that are related to the competitive Lotka-Volterra models in ecology [HW78]. The conventional Harris and Wilson model in [HW78] is obtained by combining Lotka-Volterra models with statistical averaging procedures, after having expressed the flows in terms of the evolving structure and spatial interaction. As it tends to be more feasible to observe the emergent structure, for example, configurations of floorspace dedicated to retail activity, this work is largely motivated by the existing models of urban structure [HW78, Wil00, Wil11, Wil12, DW15]. By adopting a similar approach, the flows between locations are missing data.

There is an urgent need to provide an improved modelling capability that captures the stochastic nature and uncertainty associated with the evolution of urban structure. A shortcoming of the Harris and Wilson model is that it is deterministic and converges to a fixed point that depends on the initial conditions. In reality, the evolution of urban structure would be accompanied by fluctuations arising from less predictable events [Vor85, Wil08, DW15]. In this chapter, a continuous-time stochastic term is included in the model to reformulate the Harris and Wilson model as a well-posed system of ergodic SDEs [KP13]. An associated Bayesian inference methodology is provided for parameter estimation and model calibration for the newly proposed model.

The existing methodology provides equilibrium structures as given by the fixed points of an ordinary differential equation (ODE) model. A stochastic analogue that represents the uncertainty in equilibrium structures is the stationary probability distribution of an ergodic SDE model. The structural variables can be represented by a potential function, and the stationary distribution of the SDE model is available in closed-form via the steady-state Fokker-Planck equation. The Boltzmann-Gibbs distribution one obtains is related to an idea in statistical mechanics known as the principle of maximum entropy [Jay57], which has been applied to variety of phenomena [BCG⁺12, TMM⁺15] including the modelling of urban flows [Wil67, Wil10, CNP18]. The Boltzmann-Gibbs distribution provides a probabilistic model of urban structure, and the potential function can be interpreted as the constraints on consumer welfare, running costs and a first-order moment at statistical equilibrium. This approach is novel, as the connection between dynamical systems and entropy maximisation described in [Mac03a, LM13] has not yet been explored in the context of urban modelling [Wil08, Wil09]. Whilst it is mathematically convenient to assume that the system is close to a fixed point or statistical equilibrium, there are likely to be economic, social or political events that prevent this from being the case. To relax the assumption of equilibrium, one would need to study the filtering or smoothing problem, which would be a computationally intensive task that assumes the availability of high-frequency data [LSZ15]. The assumption of statistical equilibrium provides a stochastic generalisation of the existing methodology.

To this end, the stationary distribution of the SDE model is used to represent the uncertainty in equilibrium structures as given by the existing methodology. A Bayesian approach to parameter estimation is adopted, in which the Boltzmann-Gibbs distribution forms an integral part of the

(assumed) data generating process. There are significant computational challenges, as there is an intractable term in the density of the Boltzmann-Gibbs distribution that depends on the model parameters. The intractable term must be taken into consideration when using MCMC to explore the posterior distribution. The proposed approach is applicable to a wide range of urban and regional systems, and is demonstrated with a case study of the London (UK) retail system.

5.2 The Boltzmann Lotka-Volterra methodology

The Boltzmann Lotka-Volterra (BLV) methodology is well established in urban and regional geography [HW78, Wil08, DW15]. This section describes the methodology that underpins the Harris and Wilson model. It is assumed that urban structure is characterised by a time-dependent vector of sizes $W(t) = \{W_1(t), \dots, W_M(t)\} \in \mathbb{R}_{>0}^M$, for $t \geq 0$. An aggregate model of the London (UK) retail system is used to demonstrate the ideas, however, the methodology is general and has wider applications, for example, in archaeology, logistics, healthcare and crime to name a few [DW15]. In the proceeding section, a stochastic reformulation of the Harris and Wilson model is provided to formally account for the uncertainty associated with the evolution of urban structure.

5.2.1 Spatial interaction models (Boltzmann)

In urban and regional geography, spatial interaction models describe the flow of quantities – such as goods, services or people – between two regions that are each characterised by their spatial location. A widely known type of spatial interaction model is the gravity model, which is based on an analogy with Newtonian physics [HF84]. It is assumed that each flow is determined by the attraction between the origin and the destination pair, however, the attraction is reduced by their spatial separation. Gravity models can require additional constraints, for example, to ensure that the total flows equal a known number. These constraints can be enforced with the inclusion of so-called balancing factors that play a similar role to the partition functions encountered in statistical mechanics. A more powerful modelling framework has been obtained by a thermodynamic analogy [Wil67]. This is the modelling framework described here.

It is assumed that an urban system is characterised by N origin zones and M destination zones, with the spatial location of each zone known. The flow of quantities from origin zone i to destination zone j is denoted by T_{ij} , and the configuration of flows $\{T_{ij}\}$ forms an $N \times M$ matrix. It is assumed that a flow T_{ij} is characterised by three parts: the origin-effects O_i , the destination effects W_j and the cost or inconvenience associated with the flow c_{ij} . The general form of model is

$$T_{ij} = kO_iW_jf(c_{ij}), \quad (5.1)$$

where $f(c_{ij})$ is a function that decreases as c_{ij} increases and k is a scaling constant. A single flow is depicted in Fig. 5.1.

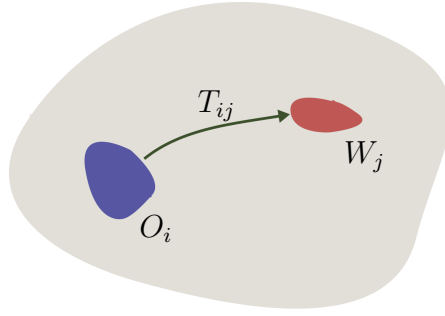


Figure 5.1: Illustration of a single flow in a urban or regional system. O_i are the origin effects, W_j are the destination effects and T_{ij} is the flow between the two locations.

In practice, there are a number of constraints that (5.1) does not satisfy. For example, for a model of retail activity, it is appropriate to constrain the origin-effects to be equal to the spending power of the respective zone. Mathematically, this means that the outflows for each origin zone is given by

$$\sum_{j=1}^M T_{ij} = O_i, \quad \text{for } i = 1, \dots, N. \quad (5.2)$$

For a model of retail activity, it is assumed that there is an average trip cost and an average trip benefit. These two constraints are expressed as

$$\sum_{i=1}^N \sum_{j=1}^M T_{ij} c_{ij} = C, \quad (5.3)$$

and

$$\sum_{i=1}^N \sum_{j=1}^M T_{ij} X_j = X, \quad (5.4)$$

respectively. Here, c_{ij} is the previously defined cost or inconvenience associated with a flow

from i to j , and X_j is the attractiveness (an unscaled measure of benefit) of the destination zone j . It is assumed that the attractiveness variables are related to the structural variables by $X_j = \ln W_j$. The constraints in (5.3) and (5.4) can be enforced using Boltzmann's ideas from statistical mechanics.

It is assumed that there are T units of flow in total, with each unit assigned to an (i, j) pair and with T being a large number. Boltzmann showed that when assuming each particular assignment is equally probable, there is one configuration $\{T_{ij}\}$ that is overwhelmingly the most likely to occur. This is the assignment that maximises the entropy function

$$H = - \sum_{i=1}^N \sum_{j=1}^M T_{ij} \ln T_{ij}, \quad (5.5)$$

subject to what is already known about the system. By maximising the entropy function in (5.5), subject to the constraints in (5.2)-(5.4), one obtains the following model of flows:

$$T_{ij} = A_i O_i W_j^\alpha e^{-\beta c_{ij}}. \quad (5.6)$$

The flows can be written as

$$W_j^\alpha e^{-\beta c_{ij}} = \exp(U_{ij}), \quad (5.7)$$

where each $U_{ij} = \alpha X_j - \beta c_{ij}$ is interpreted as a utility, and $\alpha > 0$ and $\beta > 0$ are the attractiveness and cost scaling parameters, respectively. α describes the extent that larger zones are preferred, for example, due to an improved range of choice or economies of scale. β describes the extent that local zones are preferred, for example, due to the inconvenience associated with travel. Here, the balancing factor A_i is given by

$$A_i = \frac{1}{\sum_{k=1}^M W_k^\alpha e^{-\beta c_{ik}}}, \quad (5.8)$$

which ensures that the flows are normalised according to (5.2). The inward flows for each destination are calculated by the summation

$$D_j = \sum_{i=1}^N T_{ij} = \sum_{i=1}^N O_i \frac{W_j^\alpha e^{-\beta c_{ij}}}{\sum_{k=1}^M W_k^\alpha e^{-\beta c_{ik}}}, \quad (5.9)$$

which is known as an activity location model, since D_j is a measure of activity at j .

Spatial interaction models are usually calibrated via least-squares optimisation, which amounts

to maximum likelihood estimation when using a Gaussian likelihood [Wil71]. It is also possible to use a Poisson likelihood [FL88]. In both cases, flow data is required to calibrate the spatial interaction model.

5.2.2 Structural evolution (Lotka Volterra)

So far it has been considered that the flows occur according to fixed values of W_j , O_i and c_{ij} . In practice, zones that have an inflow greater than their notional capacity will grow, whereas zones that are underutilised will shrink towards a more sustainable level. The assumption that zones will resize until a break-even point is achieved is justified by including the cost of capital in the costs that determine the notional capacity.

A suitable model for the evolution of the sizes $W(t) \in \mathbb{R}_{>0}^M$, for $t \geq 0$, is given by the Harris and Wilson model [HW78]. The model is a system of ODEs that are coupled by the spatial interaction model in (5.9)¹:

$$\begin{aligned} \frac{dW_j}{dt} &= \epsilon W_j (D_j - \kappa W_j), \\ W_j(0) &= w_j, \end{aligned} \tag{5.10}$$

for $j = 1, \dots, M$. Here, $\epsilon > 0$ is the responsiveness parameter and $\kappa > 0$ is the cost per unit size. The multiplicative responsiveness represents that larger zones are faster at responding to change than smaller zones are. It is assumed that the spatial interaction model is updated immediately for any change in W ; the spatial interaction model describes fast dynamics and the logistic growth equations describe the slow dynamics. The multiplicative responsiveness in (5.10) can be replaced with a constant responsiveness to accelerate the convergence to equilibrium, however, the former is considered a more realistic modelling assumption [DW11, DW15].

It is known that the system of ODEs will converge to a fixed point that satisfies the following system of equations:

$$\sum_{i=1}^N O_i \frac{W_j^\alpha e^{-\beta c_{ij}}}{\sum_{k=1}^M W_k^\alpha e^{-\beta c_{ik}}} = \kappa W_j, \quad j = 1, \dots, M. \tag{5.11}$$

It is known the system of equations has a unique positive solution for $0 < \alpha \leq 1$, however, there may be multiple positive solutions for $\alpha > 1$ [RV83, Vor85]. The stability of the system can be

¹For simplicity, time-dependence is not made explicit when describing differential equations.

studied via the Lyapunov function

$$V(W) = -\alpha^{-1} \sum_{i=1}^N O_i \ln \sum_{j=1}^M W_j^\alpha e^{-\beta c_{ij}} + \kappa \sum_{j=1}^M W_j + \text{const}, \quad (5.12)$$

which satisfies

$$\frac{dV}{dt} = \sum_{j=1}^M \frac{dV}{dW_j} \frac{dW_j}{dt} = -\epsilon \sum_{j=1}^M (D_j - \kappa W_j)^2 \leq 0, \quad (5.13)$$

with equality if and only if the fixed point condition in (5.11) holds [Vor85]. For $0 < \alpha \leq 1$, the uniqueness of the fixed point follows by the convexity of V , which can be shown by writing the Hessian of V as the sum of two positive definite matrices:

$$\text{Hess}(V(W)) = \alpha H_1 + (1 - \alpha) H_2, \quad (5.14)$$

where the elements of H_1 and H_2 are given by

$$(H_1)_{nm} = \sum_{i=1}^N O_i \frac{W_n^{\alpha-1} e^{-\beta c_{in}}}{\sum_{k=1}^M W_k^{\alpha-1} e^{-\beta c_{ik}}} \frac{W_m^{\alpha-1} e^{-\beta c_{im}}}{\sum_{k=1}^M W_k^{\alpha-1} e^{-\beta c_{ik}}}, \quad (5.15)$$

and

$$(H_2)_{nm} = \delta_{nm} \sum_{i=1}^N O_i \frac{W_n^{\alpha-2} e^{-\beta c_{in}}}{\sum_{k=1}^M W_k^{\alpha} e^{-\beta c_{ik}}}, \quad (5.16)$$

respectively.

The Harris and Wilson model can be calibrated to measurement data via least-squares optimisation or an R^2 analysis. It is common practice to assume an initial condition before running the ODE model until it converges to a fixed point. It is possible to calibrate the model by observing the structural variables [DW11, DW15] or the flow data [CNP18]. For $\alpha > 1$, the calibration procedure may be sensitive to the choice of initial condition due to non-uniqueness of the fixed point.

5.2.3 Case study: the London retail system

The area of retail floorspace for London town centres for 2008 is obtained from a London Town Centre Health Check report prepared by the Greater London Authority [Aut09]. There are 227 town centres, although only those classified as international, metropolitan and major town centre classifications are used for the study, giving $M = 49$ town centres. The remaining town

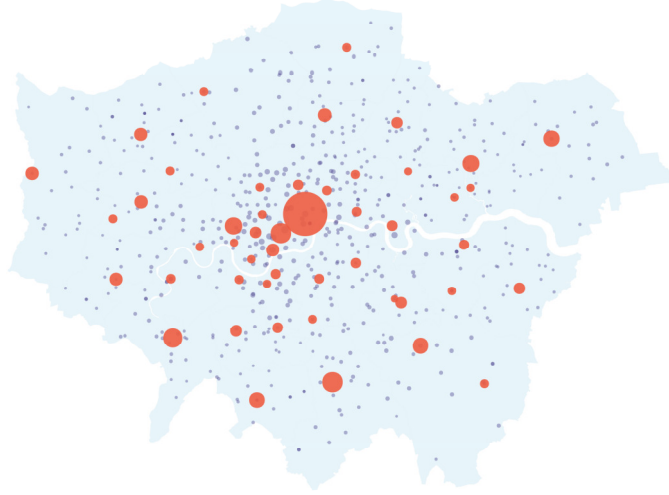


Figure 5.2: Visualisation of the observation data Y (red) over the map of London. The red markers indicate the destination zones, which are the 49 town centres, and the blue markers indicate the origin zones, which are the 625 residential wards. The sizes of the markers are given by the respective Y and O values, and each zone is plotted at its longitude-latitude coordinate.

centres are mostly district town centres that have a relatively high concentration of convenience goods and have a more localised catchment; these would be better modelled separately. The destination sizes W_j are taken to be the floorspace measurements for retail activity. The origin quantities are determined from ward-level household and income estimates, for which there are $N = 625$ wards, published by the Greater London Authority [Aut16b, Aut15]. The origin quantities O_i are taken to be the spending powers equal to the population size multiplied by the average income. The data is presented in Fig. 5.2, over the map of London using boundary data from [Aut16a].

The two key unknown parameters are α and β ; the attractiveness and cost scaling parameters, respectively. Both parameters are expected to be positive, however, these quantities describe consumer behaviour and are difficult to ascertain. It is assumed that the practitioner has a reasonable way of determining the remaining parameters ϵ and κ , and this has been standard practice in the urban modelling literature so far.

5.3 Stochastic modelling of urban structure

This section provides a stochastic generalisation of the Harris and Wilson model. The need for an element of stochasticity to account for less predictable events has long been recognised in the literature. For example, the case is made in [Wil08] and some ideas have been proposed in [Vor85, DW15]. In this section, it can be more natural to represent the structural variables $W(t) = \{W_1(t), \dots, W_M\} \in \mathbb{R}_{>0}^M$ by the attractiveness variables $X(t) = \{X_1(t), \dots, X_M\} \in \mathbb{R}^M$ where each $W_j = \exp(X_j)$. In this way, there are no positivity issues when developing a stochastic model. First, a stochastic reformulation of the Harris and Wilson model is provided. Second, the connection between the SDE model and entropy maximisation is made.

5.3.1 A stochastic reformulation of the Harris and Wilson model

This section considers diffusion processes, which are the continuous-time analogue of a discrete-time system with a constant stochastic perturbation [LM13]. Diffusion processes provide a sufficiently broad class of processes for a stochastic urban model and, due to the Markov property, are well-understood. For instance, there are results over existence, uniqueness and well-posedness of their solutions, and the probability distribution of the process solves a partial differential equation known as the Fokker-Planck equation. To this end, a diffusion processes is a SDE of the form

$$dX(t) = b(X(t), t)dt + \sigma(X(t), t) \circ dB(t), \quad (5.17)$$

where $X(t) \in \mathbb{R}^M$, $b : \mathbb{R}^M \rightarrow \mathbb{R}^M$ is a smooth vector-valued function and $\sigma : \mathbb{R}^M \rightarrow \mathbb{R}^{M \times K}$ is a smooth matrix-valued function. Here, $B(t)$ is a standard K -dimensional Brownian motion and the SDE is interpreted in the Stratonovich sense. The Stratonovich interpretation is obtained from a smooth approximation of white noise², which is known as the Wong–Zakai theorem [WZ65, Pav16]. The Stratonovich interpretation is appropriate for the dynamical system of interest, as it is stable with respect to perturbations in the noise term. The reader is referred to Appendix A for further preliminaries and background; the key modelling aspects are discussed here.

Stochastic models have been informally introduced in [DW15], by randomly perturbing an Euler-

²At least for the one-dimensional case.

discretisation of the ODE model. A more formal SDE approach has already been considered in [Vor85], however, the proposed model is restricted to the case $0 < \alpha \leq 1$, and the process converges (almost surely) to the unique fixed point satisfying (5.11), which does not account for the uncertainty in equilibrium structures. The restriction to $0 < \alpha \leq 1$ is a concern because existing retail models have been calibrated with $\alpha > 1$ [DW11, DW15].

Instead, the following system of SDEs with Stratonovich noise is considered here:

$$\begin{aligned} dW_j &= \epsilon W_j (D_j - \kappa W_j + \delta) dt + \sigma W_j \circ dB_j, \\ W_j(0) &= w_j, \end{aligned} \tag{5.18}$$

for $j = 1, \dots, M$. The parameter $\sigma > 0$ controls the level of noise in the system. The multiplicative noise represents fluctuations in the growth or contraction rate given by $D_j - \kappa W_j$. A heuristic interpretation of the SDE is that, over a short time δt , the rate $D_j - \kappa W_j$ is randomly perturbed by Gaussian noise with a standard deviation of $\sigma\sqrt{\delta t}$. The noise term represents fluctuations in growth rates arising from less predictable events that are not represented by the ODE model. The specification of multiplicative noise preserves the positivity of each W_j , and can be justified by the idea that larger zones have a greater potential for unpredictable change [DW15].

It is more convenient to study the properties of (5.18) with the change of variables $X_j = \ln W_j$, so that urban structure is represented in terms of attractiveness rather than size. This is a change of variables that transforms multiplicative noise into additive noise, known as the Lamperti transformation [Pav16]. The system becomes

$$\begin{aligned} dX_j &= \epsilon (D_j - \kappa e^{X_j} + \delta) dt + \sigma dB_j, \\ X_j(0) &= x_j, \end{aligned} \tag{5.19}$$

for $j = 1, \dots, M$. A useful observation is that (5.19) can be rewritten as a gradient system. This done by introducing the potential function $V : \mathbb{R}^M \rightarrow \mathbb{R}$ and its gradient $\nabla V : \mathbb{R}^M \rightarrow \mathbb{R}^M$ by

$$\epsilon^{-1} V(x) = -\alpha^{-1} \sum_{i=1}^N O_i \ln \sum_{j=1}^M W_j^\alpha \exp(\alpha x_j - \beta c_{ij}) + \kappa \sum_{j=1}^M \exp(x_j) - \delta \sum_{j=1}^M x_j. \tag{5.20}$$

so that the system can be written in vector-form as

$$\begin{aligned} dX &= -\nabla V(X)dt + \sqrt{2\gamma^{-1}}dB, \\ X(0) &= x. \end{aligned} \tag{5.21}$$

Here, the noise strength σ has been replaced by an ‘inverse-temperature’ parameter $\gamma = 2\sigma^{-2}$ for notional convenience. Equation (5.21) is the Langevin equation in the overdamped limit. The Langevin equation is used in the physical sciences to model interacting particles that are subjected to potential forces and stochastic fluctuations in the limit of no inertial effects [LM15, Pav16]. For Langevin dynamics, the strength of the fluctuations is proportional to the square-root of the temperature of the system. This is why γ is referred to as an inverse temperature.

The probability density of $X(t)$, denoted $\rho(x, t)$, evolves in time according to the Fokker-Planck equation (FPE). The FPE may be written as

$$\begin{aligned} \frac{d\rho(x, t)}{dt} &= -\nabla \cdot J(x, t), \\ J(x, t) &= -\nabla V(x)\rho(x, t) - \gamma^{-1}\nabla\rho(x, t), \\ \rho(x, 0) &= \rho_0(x). \end{aligned} \tag{5.22}$$

Whilst (5.22) is very challenging to solve, especially in higher-dimensions, a steady-state solution occurs when the probability flux $J(x, t)$ vanishes, which happens for

$$\rho_*(x) = \frac{1}{Z}e^{-\gamma V(x)}. \tag{5.23}$$

This is the density of a Boltzmann-Gibbs measure. For the measure to be appropriately defined, the potential function must be confining so that $\lim_{|x| \rightarrow \infty} V(x) = \infty$ and

$$Z = \int_{\mathbb{R}^M} e^{-\gamma V(x)} dx < \infty, \tag{5.24}$$

for all $\gamma > 0$. A sufficient condition is that there exists $b \in (0, 1)$ and $C > 0$ such that

$$\liminf_{|x| \rightarrow \infty} (b|\nabla V|^2 - \gamma^{-1}\Delta V) \geq -C, \tag{5.25}$$

as proved by Lemma 3.14 in [MS14]. The condition is satisfied when V grows at least linearly but no faster than exponentially at infinity. The following bound on the softmax function [NS16]:

$$\max_{0 \leq j \leq M} x_j \leq \log \sum_{j=1}^M \exp(x_j) \leq \max_{0 \leq j \leq M} x_j + \ln M. \quad (5.26)$$

means that $\delta > 0$ is a requirement. The corresponding FPE has a vanishing boundary condition.

The process satisfying (5.21) may also be restricted to a bounded domain $\mathcal{X} \subset \mathbb{R}^M$ with a smooth boundary $\partial\mathcal{X}$, for example, via a reflecting process on $\partial\mathcal{X}$. Reflecting SDEs are formally defined in [LS84]. The domain $\mathcal{X}$ can be chosen to contain a metastable state of V so that one effectively studies the quasi-stationary distribution of a trapped process [CMSM12, TLL19]. In this way, it is possible to alleviate the multimodality caused by an excessively complex potential function. The simplification may be justifiable due to prior knowledge of the system being studied. The corresponding FPE has a no-flux boundary condition.

Properties of the Boltzmann-Gibbs measure

The Boltzmann-Gibbs measure is unimodal for $0 < \alpha \leq 1$ but can be multimodal for $\alpha > 1$ by a similar argument to the one presented in Section 5.2.2. There must be at least one mode as V is coercive [NW06]. A mode can only occur when $\nabla V(x) = 0$, which is equivalent to the condition

$$\sum_{i=1}^N O_i \frac{W_j^\alpha \exp(-\beta c_{ij})}{\sum_{k=1}^M W_k^\alpha \exp(-\beta c_{ik})} + \delta = \kappa W_j, \quad j = 1, \dots, M. \quad (5.27)$$

Clearly as $\delta \rightarrow 0$, (5.27) approaches the equilibrium condition given by (5.11). Requiring $\delta > 0$ means that each W_j solving (5.27) must be strictly positive and bounded below by $W_{\min} = \delta/\kappa$. The potential function is illustrated in Fig. 5.3 for a 2-dimensional example for various values of α with the remaining parameters held fixed. The Boltzmann-Gibbs measure is noticeably multimodal for $\alpha \geq 1.3$. The multimodality occurs because only one of the two zones carries most of the activity.

If $x_* = \min_{x \in \mathbb{R}^M} V(x)$ is a unique minima of V , then the Boltzmann-Gibbs measure collapses to a Dirac mass located at x_* in the noiseless limit as $\gamma \rightarrow \infty$. This is in contrast to the notion of a fixed point of the deterministic Harris and Wilson model, which depends on the initial condition. In other words, x_* is the solution to a global optimisation problem rather than a local one.

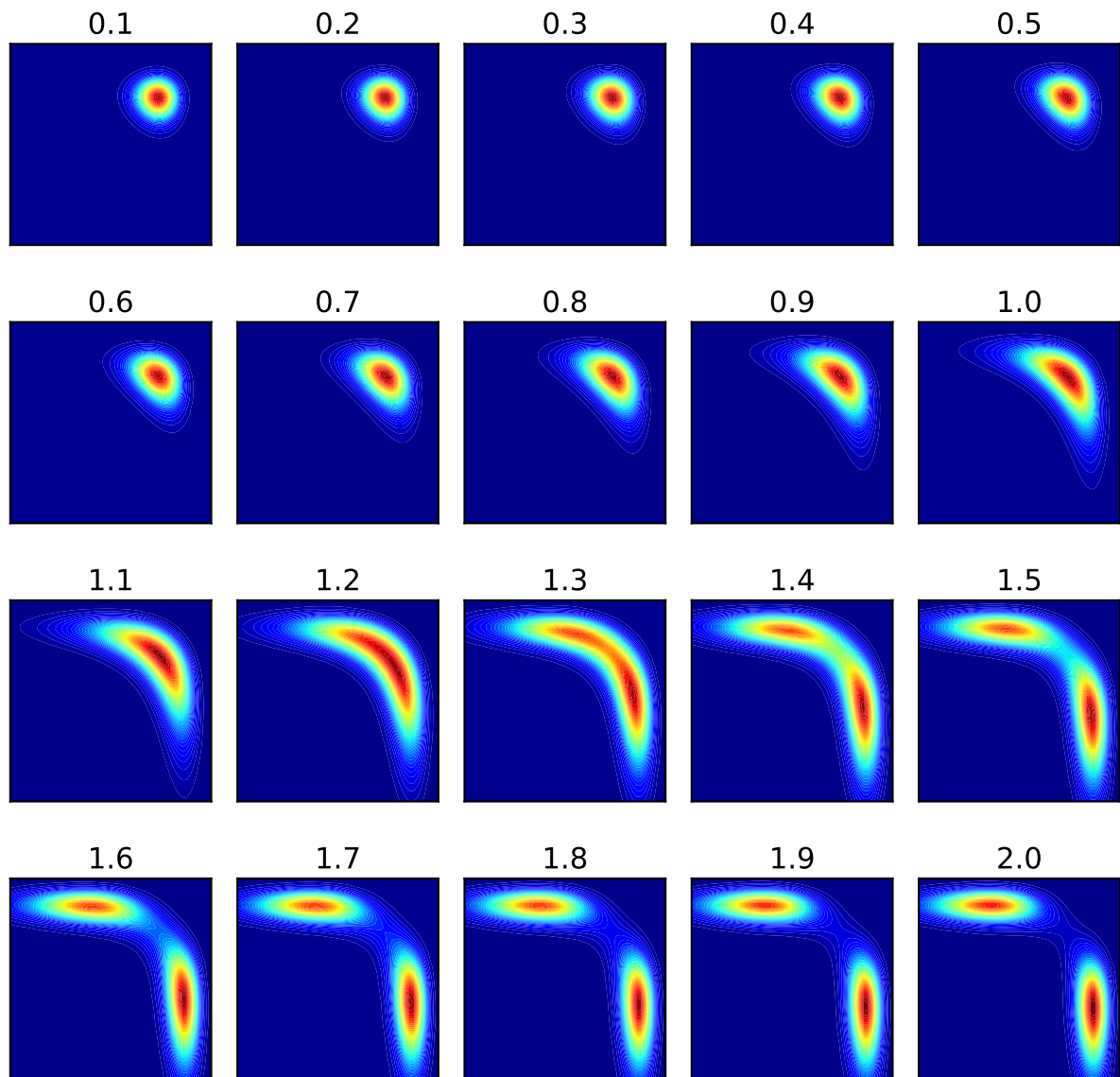


Figure 5.3: The potential function for a 2-dimensional system with $x = (x_1, x_2)^T$ for various values of α . The x -axis represents the x_1 value and the y -axis represents the x_2 value. The potential function noticeably has multiple minima for $\alpha \geq 1.3$.

5.3.2 The principle of maximum entropy

The Boltzmann-Gibbs measure in (5.23) can also be obtained from a maximum entropy argument. The advantage of this view is that the terms in the potential function can be interpreted as economic constraints on the stationary distribution of (5.21).

The idea is to introduce an entropy function that represents the uncertainty in the system, and to maximise the entropy function subject to what is known about the system. The entropy function is taken to be the expectation of the negative index of probability. It is defined as

$$H(\rho) = - \int_{\mathbb{R}^M} \rho(x) \ln \rho(x) dx. \quad (5.28)$$

The integral is only taken over the positive values of ρ , since $\lim_{\rho \rightarrow 0} \rho \ln \rho = 0$. It is assumed that there are a finite number of constraints of the form

$$\langle Q_i \rangle = \int_{\mathbb{R}^M} Q_i(x) \rho(x) dx = C_i, \quad (5.29)$$

where each C_i is a real number and each Q_i is a measurable function representing a constrained quantity. Then provided that all integrals are finite, the unique probability density function that maximises the entropy function is the Boltzmann-Gibbs measure whose potential function is linear in the constrained quantities, as shown by Proposition 9.1.2 in [LM13].

It remains to specify suitable constraints for urban systems. An attempt has already been made to represent the structural variables by a potential function to obtain a model analogous to the Ising model in statistical physics [Wil08, Wil09]. The constraint that has already been considered is on the expected deviation from the equilibrium condition in (5.11), which amounts to studying the potential function

$$V(x) = \sum_{j=1}^M (D_j - \kappa W_j). \quad (5.30)$$

Whilst (5.30) is expected to be close to zero, it can only represent a constraint on the total capacity of the system because the sum over D_j is constant for a singly-constrained system. It is not restrictive enough.

Instead, the potential function in (5.20) is written as

$$\epsilon^{-1}V(x) = Q_{\text{welfare}}(x) + \kappa Q_{\text{capacity}}(x) + \delta Q_{\text{attractiveness}}(x), \quad (5.31)$$

and the constraints are given by:

$$\langle Q_{\text{welfare}} \rangle = C_{\text{welfare}}, \quad (5.32)$$

$$\langle Q_{\text{capacity}} \rangle = C_{\text{capacity}}, \quad (5.33)$$

$$\langle Q_{\text{attractiveness}} \rangle = C_{\text{attractiveness}}. \quad (5.34)$$

The Boltzmann Gibbs measure given by (5.23) is the distribution that maximises the Boltzmann-Gibbs entropy in (5.28), subject to the constraints (5.32)-(5.34). This is a consequence of Gibb's inequality, which may be written as

$$-\int_{\mathbb{R}^M} \rho(x) \ln \rho(x) dx \leq -\int_{\mathbb{R}^M} \rho(x) \ln \rho_*(x) dx, \quad (5.35)$$

with equality if and only if $\rho = \rho_*$ [LM13]. It is assumed that both integrals are finite. An application of Gibbs' equality provides an upper bound on the entropy function:

$$\begin{aligned} H(\rho) &\leq -\int_{\mathbb{R}^M} \rho(x) \ln \rho_*(x) dx, \\ &= \int_{\mathbb{R}^M} [\ln Z + \gamma V(x)] \rho(x) dx, \\ &= \ln Z + \gamma \int_{\mathbb{R}^M} Q_{\text{welfare}}(x) \rho(x) dx + \gamma \kappa \int_{\mathbb{R}^M} Q_{\text{capacity}}(x) \rho(x) dx + \gamma \delta \int_{\mathbb{R}^M} Q_{\text{attractiveness}}(x) \rho(x) dx, \\ &= \ln Z + \gamma C_{\text{welfare}} + \gamma \kappa C_{\text{capacity}} + \gamma \delta C_{\text{attractiveness}}, \end{aligned} \quad (5.36)$$

with equality if and only if $\rho = \rho_*$. Therefore the entropy function is only maximized for $\rho = \rho_*$. The constraints are considered in more detail under the following headings.

Total welfare constraint

The first term in (5.20) relates to the constraint in (5.32), which can be written as

$$\left\langle \alpha^{-1} \sum_{i=1}^N O_i \ln \sum_{j=1}^M \exp(U_{ij}) \right\rangle = C_{\text{welfare}}. \quad (5.37)$$

where the deterministic utility functions is defined by

$$U_{ij}(x_j) = \alpha x_j - \beta c_{ij}. \quad (5.38)$$

The welfare measure can be interpreted by using the random utility maximisation framework [McF73, Wil77]. The stochastic utility function for a possible choice being made from i to j is defined by:

$$\tilde{U}_{ij}(x_j) = U_{ij}(x_j) + \epsilon_{ij}, \quad (5.39)$$

where each of the ϵ_{ij} are independent and identically distributed Gumbel random variables. With this choice of random variables, it can be shown that

$$P\left[\cap_{k \neq j} \left\{U_{ij}(x_j) > U_{ik}(x_k)\right\}\right] = \frac{\exp(U_{ij}(x_j))}{\sum_{k=1}^M \exp(U_{ik}(x_k))}. \quad (5.40)$$

The expected utility attained for a unit flow $E[\max_{1 \leq j \leq M} \tilde{U}_{ij}(x_j)]$ is available in closed-form, and is given by

$$E\left[\max_{1 \leq j \leq M} \tilde{U}_{ij}(x_j)\right] = \ln \sum_{j=1}^M \exp(U_{ij}(x_j)) + c, \quad (5.41)$$

where c is the Euler-Mascheroni constant [Wil77, Tra09]. Lastly, by summing over each flow and one obtains

$$\alpha^{-1} \sum_{i=1}^N O_i E\left[\max_{1 \leq j \leq M} \tilde{U}_{ij}(x_j)\right] = Q_{\text{welfare}}(x) + \text{const}, \quad (5.42)$$

where the constant term can be discarded. The scaling factor of α^{-1} is mathematically necessary for the gradient system in (5.21) to be consistent with (5.19) and (5.9).

Total capacity constraint

The second term in (5.20) relates to the constraint in (5.33), which can be written as

$$\left\langle \sum_{j=1}^M W_j \right\rangle = C_{\text{capacity}}. \quad (5.43)$$

Each $W_j(x_j) = \exp(x_j)$, as before. This constraint reflects that each zone can only operate at a finite capacity, for example, land is a scarce resource that must be put to good use. The constraint on capacity was initially represented by the logistic growth equations in (5.10). An appropriate specification must satisfy $\lim_{x_j \rightarrow \infty} V(x) = \infty$.

Total attractiveness constraint

The third term in (5.20) relates to the constraint in (5.34), which can be written as

$$\left\langle \sum_{j=1}^M X_j \right\rangle = C_{\text{attractiveness}}. \quad (5.44)$$

The purpose of this constraint is to ensure that no zone is so unattractive that it bears no activity. For example, there are mechanisms such as continued investment or government initiatives that prevent this from happening. It prevents zones from collapsing due to a lack of activity. An appropriate specification must satisfy $\lim_{x_j \rightarrow -\infty} V(x) = \infty$. The total attractiveness constraint is mathematically necessary to ensure that V is a confining potential.

5.3.3 The thermodynamics of cities and regions

It has been shown that at statistical equilibrium, when $\rho = \rho_*$, the Boltzmann-Gibbs entropy of the attractiveness variables is maximised subject to the equality constraints on the expected total welfare, expected total capacity and expected total attractiveness. This is not by coincidence, and the connection between stochastic dynamical systems and entropy maximisation has been studied extensively in the mathematical physics literature [JKO98, MV99, Mac03b, Pav16]. These ideas have not been recognised in the urban modelling literature.

When the structural variables evolve in time according to (5.21), the system is analogous to a thermodynamic system in that there exists an entropy function that is increasing and converges to a unique limit attained if and only if $\rho = \rho_*$. The relevant form of entropy is the conditional entropy

$$H_c(\rho|\rho_*) = - \int_{\mathbb{R}^M} \rho(x) \ln \left[\frac{\rho(x)}{\rho_*(x)} \right] dx, \quad (5.45)$$

where ρ_* is the density of the Boltzmann-Gibbs measure in (5.23). The conditional entropy can

be rewritten as

$$H_c(\rho|\rho_*) = H(\rho) + \int_{\mathbb{R}^M} \rho(x) \ln \rho_*(x) dx, \quad (5.46)$$

$$= H(\rho) - \gamma \int_{\mathbb{R}^M} V(x) \rho(x) dx - \ln Z, \quad (5.47)$$

which is equal to the Boltzmann-Gibbs entropy plus a Lagrange multiplier that enforces the constraints on the maximum entropy distribution ρ_* . By Gibbs' inequality in (5.35), the conditional entropy is bounded above by zero and attains its maximum value when $\rho = \rho_*$. It is shown in Appendix A that

$$\frac{dH_c}{dt} \geq 0, \quad (5.48)$$

with equality if and only if $\rho = \rho_*$. It is also shown in Appendix A that, due to the Poincaré inequality implied by (5.25), $\rho(\cdot, t)$ converges to ρ_* at rate that is exponentially fast provided that $\rho_0 \in L^2(\mathbb{R}^M; \rho_*)$. In this case

$$\lim_{t \rightarrow \infty} H_c(\rho(\cdot, t)|\rho_*) = 0. \quad (5.49)$$

5.4 Bayesian parameter estimation

This section addresses the inverse problem: the task of determining the model parameters from an observed configuration of sizes. The inverse problem is primarily motivated by the need to calibrate models for urban simulations.

5.4.1 Existing approaches

The two key unknowns are α and β . It is assumed that the practitioner is able to determine the remaining parameters and this has been standard practice in the urban modelling literature so far. In principle, α and β can be estimated from flow data by calibrating a spatial interaction model. This can approach can be impractical as flow data is difficult to obtain and requires an expensive survey that would suffer from sampling biases. In [DW11, DW15], the values of α and β were determined via least squares minimisation by comparing the predicted sizes with the observed sizes. The predicted sizes are obtained by running the ODE model in (5.10) until convergence, using a pre-specified initial condition. This approach does not account for the

uncertainty in the dynamics nor the observation data. Instead a Bayesian approach is proposed here.

5.4.2 Proposed Bayesian hierarchical model

It is assumed that an observed vector of sizes $y \in \mathbb{R}_{>0}^M$ is related to some latent sizes $w \in \mathbb{R}_{>0}^M$ and multiplicative noise $e \in \mathbb{R}_{>0}^M$ by

$$\ln y = \ln w + \ln e. \quad (5.50)$$

The specification of multiplicative noise is appropriate since all measurements are positive and there is more scope for error when measuring larger zones. As before, it is natural to work in terms of log-sizes $x = \ln w \in \mathbb{R}^M$, and it is assumed that $x \sim \rho_*$ is a realisation of the Boltzmann-Gibbs measure in (5.23). The latent variables x depend on the model parameters $\theta = \{\alpha, \beta\} \in \mathbb{R}_{>0}^2$, which is summarised by a single variable for notational convenience. It is further assumed that $\ln e \sim N(0, \Sigma)$ is a realisation of Gaussian noise for some symmetric positive definite covariance matrix $\Sigma \in \mathbb{R}^{M \times M}$. The noise term is assumed to be independent and identically distributed so that $\Sigma = \lambda^2 I$ with $\lambda > 0$.

The prior distribution on the model parameters is denoted $\pi(\theta)$. The prior distribution for the latent variables is denoted $\pi(x|\theta)$ and is equal to $\rho_*(x)$ in (5.23). It is repeated here to make the θ -dependence explicit in the notation:

$$\pi(x|\theta) = \frac{1}{z(\theta)} \exp(-\gamma V_\theta(x)), \quad z(\theta) = \int_{\mathbb{R}^d} \exp(-\gamma V_\theta(x)) dx. \quad (5.51)$$

It is emphasised that $\pi(x|\theta)$ is only known up to a normalising constant $z(\theta)$ that is a function of the unknown model parameters θ . The likelihood function $\pi(y|x)$ is the Gaussian density as specified by (5.50). The joint posterior density then has the form

$$\pi(x, \theta|y) \propto \pi(\theta) \frac{1}{z(\theta)} \exp(-\gamma V_\theta(x)) \pi(y|x). \quad (5.52)$$

It is ‘doubly-intractable’ in that both the normalising constant of (5.52) and the function $z(\theta)$ are both unknown. The θ -dependence for the $z(\theta)$ -term poses significant computational challenges as the joint posterior density cannot be evaluated at all, not even up to an irrelevant multiplicative constant.

The following subsections discuss some possible methodologies to estimate posterior quantities of the form

$$E[\phi(x, \theta)|y] := \int_{\mathbb{R}_{>0}^2} \int_{\mathbb{R}^d} \phi(x, \theta) \pi(x, \theta|y) dx d\theta, \quad (5.53)$$

where $\phi(x, \theta)$ is an integrable function of interest.

5.4.3 Parameter estimation in the noiseless limit

The inference procedure can be simplified substantially by calibrating an approximate model in the noiseless limit $\gamma \rightarrow \infty$ for a bounded domain $\mathcal{X} \subseteq \mathbb{R}^d$. This is because of the following asymptotic result, known as a saddle point approximation:

$$\int_{\mathcal{X}} f(x) e^{\gamma V_{\theta}(x)} dx = (2\pi)^{M/2} e^{\gamma V(m_{\theta})} |\nabla \nabla V_{\theta}(m_{\theta})|^{-1/2} (f(m_{\theta}) + O(\gamma^{-1})), \quad (5.54)$$

where m_{θ} is the single minimum of V_{θ} on $\mathcal{X}$. Heuristically, it can be justified by Taylor-expanding V_{θ} around its global minima m_{θ} and noting that the most significant contributions to the integral are from this neighbourhood.

The global minima can be found via Newton-based optimisation with a suitable initial condition [NW06]. It is assumed that $\mathcal{X}$ is chosen such that V_{θ} has a single minima in the interior of $\mathcal{X}$. This is possible in all but special cases. As noted in Section 5.3.1, V_{θ} may have multiple minima when $\alpha \geq 1$ and therefore the optimisation procedure may need to be performed with multiple initialisations.

In the limit $\gamma \rightarrow \infty$ the marginal likelihood becomes

$$p(y|\theta) = \int_{\mathcal{X}} p(y|x) p(x|\theta) dx = \frac{\int_{\mathcal{X}} p(y|x) e^{\gamma V_{\theta}(x)} dx}{\int_{\mathcal{X}} e^{-\gamma V_{\theta}(x)} dx} = \mathcal{N}(y|m_{\theta}, \Sigma). \quad (5.55)$$

Posterior quantities of the form can be estimated via a 2-dimensional quadrature rule. Alternatively the posterior can be approximated by the maximum-a-posteriori values for α and β by solving the optimisation problem with a 2-dimensional grid search. In the case that uniform priors are placed on α and β , the optimisation problem is a least-squares optimisation problem

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} |y - m_{\theta}|^2. \quad (5.56)$$

This is particularly convenient as the solution does not depend on λ . In this way, the uncertainty

in equilibrium structures is effectively represented by the Gaussian likelihood rather than a softening of the Harris and Wilson model.

5.4.4 Markov chain Monte Carlo with Laplace approximations

As suggested in [GCS⁺14], a deterministic approximation can be used to estimate the unknown $z(\theta)$ term. Following on from the asymptotic result in (5.54), $z(\theta)$ can be estimated by

$$z(\theta) \approx (2\pi)^{M/2} e^{\gamma V(m_\theta)} |\nabla \nabla V_\theta(m_\theta)|^{-1/2}. \quad (5.57)$$

The required derivatives for the curvature term are given by

$$\frac{\partial^2 V_\theta}{\partial x_n \partial x_m} = \delta_{nm} \kappa e^{x_n} - \alpha \sum_{i=1}^N O_i \Lambda_{in} (\delta_{nm} - \Lambda_{im}), \quad (5.58)$$

where Λ_{ij} is the softmax function:

$$\Lambda_{ij} = \frac{\exp(\alpha x_j - \beta c_{ij})}{\sum_{k=1}^M \exp(\alpha x_k - \beta c_{ik})}. \quad (5.59)$$

The approximation is accurate in the limit $\gamma \rightarrow \infty$, and is therefore well-suited to regimes in which the system behaviour is dominated by the potential effects rather than stochastic fluctuations [But07]. These regimes are of more interest for urban modelling, as the system responses are more predictable.

Having specified a deterministic approximation of $z(\theta)$, standard MCMC methods can be used to estimate the integral in (5.53). Since a Laplace approximation yields a biased estimate of $z(\theta)$, the MCMC scheme will not converge to the correct stationary distribution and the resulting estimates will not be consistent. The scheme is expected to be acceptable for low-noise regimes, however, there is a limited understanding of the approximation error incurred.

A block Gibbs scheme is proposed, in which the latent variables and model parameters are updated in two separate steps. This is because it would otherwise be difficult to draw joint proposals that are likely to be accepted. The MCMC scheme is also easier to tune this way. The proposed scheme is as follows.

Latent variable update. Hamiltonian Monte Carlo (HMC) is used for the x -updates to suppress random walk behaviour. This requires drawing Gaussian auxiliary variables $p \sim N(0, I)$ and updating (x, p) via simulated Hamiltonian dynamics to obtain (x', p') . The newly proposed pair is accepted with probability:

$$a_x(x', p'|x, p) = \min \left\{ 1, \frac{\pi(y|x', \theta) \exp \left(-\gamma V_\theta(x') - \frac{1}{2}|p'|^2 \right)}{\pi(y|x, \theta) \exp \left(-\gamma V_\theta(x) - \frac{1}{2}|p|^2 \right)} \right\}. \quad (5.60)$$

Model parameter update. Random walk Metropolis with reflecting boundaries are used for the θ -updates. The Metropolis-Hastings acceptance probability is given by:

$$a_\theta(\theta'|\theta) = \min \left\{ 1, \frac{\pi(y|x', \theta') z(\theta) \exp \left(-\gamma V_{\theta'}(x) \right) \pi(\theta')}{\pi(y|x, \theta) z(\theta') \exp \left(-\gamma V_\theta(x) \right) \pi(\theta)} \right\}. \quad (5.61)$$

5.4.5 Markov chain Monte Carlo with unbiased estimates

To obtain more accurate posterior summaries, especially in the event that the saddle point approximation performs poorly, a consistent estimator of (5.53) can be constructed using the techniques discussed in Section 2.4.2. The estimator is of the form

$$E[\phi(x, \theta)|y] = \lim_{n \rightarrow +\infty} \frac{\sum_{i=1}^n \Omega^{(i)} g(x^{(i)}, \theta^{(i)})}{\sum_{k=1}^n \Omega^{(k)}}. \quad (5.62)$$

The suitability of the scheme is dependent on being able to obtain precise importance sampling estimates of $z(\theta)$, which is challenging for low-noise regimes due to the concentration of measure. This is because negative values of $\Omega^{(i)}$ arise from high variability in the $z(\theta)$ estimates and have the effect of increasing the variance of the estimator given by (5.62). Moreover, high variability in the $z(\theta)$ estimates results in high rejection rates for the pseudo-marginal MCMC scheme.

An unbiased estimate of $z(\theta)$ is given by averaging over a batch of importance weights. The importance weights are evaluations of

$$w(x) = \exp \left(-\gamma V_\theta(x) \right) / q(x), \quad (5.63)$$

at locations drawn from a proposal distribution with density $q(x)$. Having obtained a set of importance weights, one can use the increasing averages estimator in [WM17], given by

$$\mathcal{V}_i = \frac{i+1}{\sum_{k=0}^i w(\hat{X}^{(k)})}, \quad \hat{X}^{(k)} \sim q, \quad (5.64)$$

so that the random truncated series

$$S = \mathcal{V}_0 + \sum_{i=1}^K \frac{\mathcal{V}_i - \mathcal{V}_{i-1}}{\Pr(T \geq i)}, \quad (5.65)$$

gives an unbiased estimate of $1/z(\theta)$ for the reasons discussed in Section 2.4.2.

A Markov-chain can be obtained using a similar block Gibbs scheme to the one used before: the acceptance probability for the x -updates is given by (5.60), however, the acceptance probability in (5.61) is

$$a_\theta(\theta'|\theta) = \min \left\{ 1, \frac{\pi(y|x', \theta')|S'| \exp(-\gamma V_{\theta'}(x))\pi(\theta')}{\pi(y|x, \theta)|S| \exp(-\gamma V_\theta(x))\pi(\theta)} \right\}. \quad (5.66)$$

With reference to the estimator in (5.62), the weights $\Omega^{(i)}$ are given by $\Omega^{(i)} = \text{sgn}(S^{(i)})$. As already mentioned, it is necessary to cache the value of $S^{(i)}$ at each iteration as part of the pseudo-marginal MCMC scheme.

5.4.6 Implementation details

Modelling aspects

Weakly-informative uniform priors are specified for α and β , which are restricted to the interval $[0, 2]$ after a suitable rescaling of β as determined by a preliminary study. This is the approach adopted in [DW11, DW15]. This setting allows for a comparison with the R^2 analysis performed for the deterministic Harris and Wilson model in [DW11, DW15].

Ideally, uninformative priors would be placed on all parameters that specify the Boltzmann-Gibbs measure. The first problem is that the resulting inferences are likely to be unacceptable for model calibration purposes. For example, a model that is primarily driven by stochastic fluctuations is of limited use for studying urban systems. The second difficulty is that there are computational issues associated with the choice of noise regime. The Laplace approximation performs well for a low noise regime, however, the importance sampling estimates are much more difficult to obtain due to the concentration of measure. A strongly informative prior over γ is therefore necessary, at least for model calibration. For simplicity, γ is specified to reflect the desired level of noise and ϵ is set to one as it is fully subsumed by γ .

Following the urban modelling literature, the value of κ can be specified by a supply and

demand argument. In the noiseless limit $\gamma \rightarrow \infty$, the equilibrium condition in (5.27) applies, and summing over each equation yields

$$W_{\text{total}} = \frac{1}{\kappa} \left(\sum_{i=1}^N O_i + \delta M \right), \quad (5.67)$$

where W_{total} is the total capacity of the system. Equation (5.27) also provides a lower bound for the smallest zone

$$W_{\min} = \delta / \kappa, \quad (5.68)$$

which can be used to determine a suitable value of δ .

Computational aspects

The step size parameter for the θ -updates is tuned to obtain an acceptance rate in the range 30% – 70%, and the step size and number of steps for the x -updates is tuned to obtain an acceptance rate of at least 90%. The θ -updates require either the global minima of V_θ on $\mathcal{X}$, for the saddle point approximation in (5.57), or consistent estimates of $1/z(\theta)$, for the pseudo-marginal MCMC scheme. When attempting to find the global minima of V_θ , M runs of the L-BFGS algorithm are performed but each with a different but deterministically chosen initial condition. To obtain consistent estimates of $1/z(\theta)$, annealed importance sampling (AIS) with HMC transition kernels is used. AIS is initialised by drawing from the log-gamma distribution equal to $\pi(x|\theta)$ in the limit $\alpha, \beta \rightarrow 0$. The random stopping time T in (5.65) is specified by the survival function $\Pr(T \geq k) \propto k^{-1.1}$. Each unbiased estimate of $1/z(\theta)$ requires $T + 1$ runs of AIS, with a new realisation of T each time. Running AIS a large number of times is a computationally intensive task, however, the estimates can be obtained in parallel. This is done using the multiprocessing library in Python.

5.5 Case study

In this section, the proposed methodology is applied to the London retail system, following [DW11, DW15]. In the context of retail, the size variables represent the area of space designated to retail activity, and the utility function in (5.38) is justified by the benefit consumers gain from the improved range of choice and economies of scale that is offset by the cost

of travel. The parameter estimation problem is of particular interest in the context of retail, as the flow data is difficult to obtain and the attractiveness variables are not directly observable.

5.5.1 Preliminary studies

First, a preliminary study is performed for a simplified model with no observation noise $\lambda \rightarrow 0$, in which case the θ -marginal of (5.52) is

$$\pi(\theta|y) \propto \pi(\theta) \frac{1}{z(\theta)} \exp(-\gamma V_\theta(x)). \quad (5.69)$$

With this simplification, it is possible to evaluate and visualise the posterior probabilities over a grid of α and β values. This is done over a 100×100 grid for $\gamma = 10^2$ and $\gamma = 10^4$, representing high-noise and low-noise regimes, respectively. Using the justification provided in the previous section, the values $\delta = 0.006$ and $\kappa = 1.3$ are used. $z(\theta)$ is estimated with a saddle point approximation, as it was not possible to obtain reliable $z(\theta)$ estimates for high values of γ using importance sampling techniques with a reasonable computational effort. The results are presented in Fig. 5.5, in which the scales indicate that the model with high-noise provides the better explanation of the data. The best fit for the high-noise regime is $\alpha = 0.90$ and $\beta = 0.46$ and the best fit for the low-noise regime is $\alpha = 1.18$ and $\beta = 0.28$. As expected, the low-noise regime suggests stronger attractiveness effects as the model with a higher level of noise is able to explain variation by stochastic growth. The α and β values are positively correlated; this is because of the offsetting effects in the utility function in (5.38).

A comparison can be performed with the R^2 analysis for the deterministic version of the Harris and Wilson model in [DW15]. The predicted value W_{Pred} is taken to be the equilibrium satisfying (5.27), obtained from the ODE model given by (5.21) but by letting $\sigma \rightarrow 0$ and using the initial condition $w_0 = y$. The R^2 value is defined as $R^2 = 1 - SS_{\text{res}}/SS_{\text{tot}}$, where $SS_{\text{res}}/SS_{\text{tot}}$ is the ratio of the variance of the residuals $y - W_{\text{Pred}}$ and the variance of the observed y . Maximising the R^2 values is effectively a least-squares optimisation procedure. The R^2 analysis yield a best fit of $\alpha = 1.36$ and $\beta = 0.42$, which is agreeable with the findings for the low-noise regime in Fig. 5.5. There are also some strong similarities between the profile of posterior probabilities and the profile of the R^2 values. First, both approaches find that the poorest fit is for a regime in which α is too high and β is too low; these values result in most

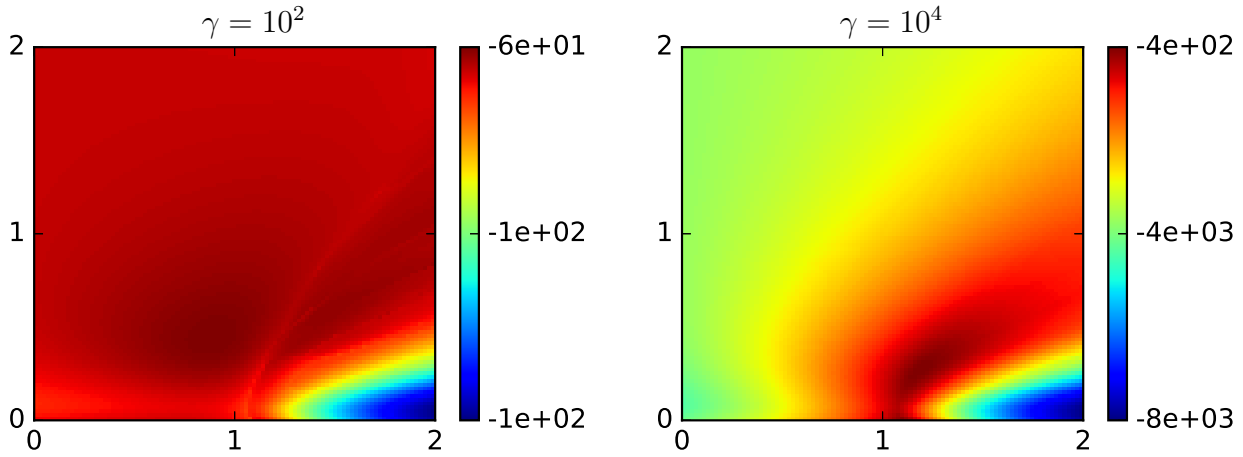


Figure 5.4: Evaluations of the log-density of the Boltzmann-Gibbs measure over a grid of 100×100 values of α and β for a regime with high-noise (left) and a regime with low-noise (right).

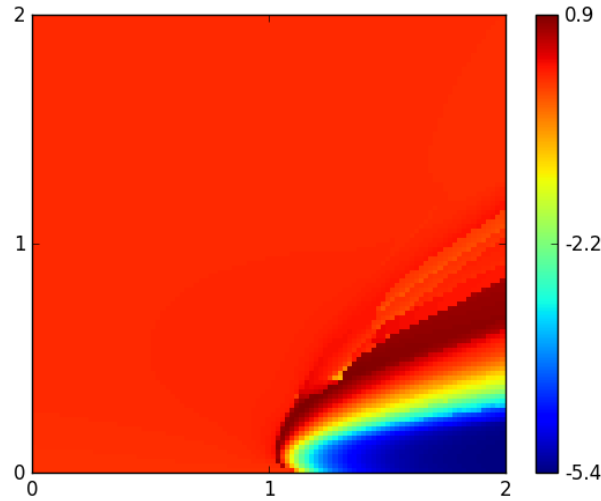


Figure 5.5: Evaluations of the R^2 value over a 100×100 grid of α and β values for the deterministic model.

activity taking place in a single zone. Second, both approaches agree in that a good fit can be found in the range $1 < \alpha < 2$.

5.5.2 Prior modelling

First, the latent variables are drawn from the prior distribution $\pi(x|\theta)$ to verify the suitability of the modelling. For illustrative purposes, α values are considered across the range $[0, 2]$, and $\beta = 0.5$ is held fixed to a reasonable value.

To represent the noiseless and low-noise regimes, configurations of the global minima of $V_\theta(x)$

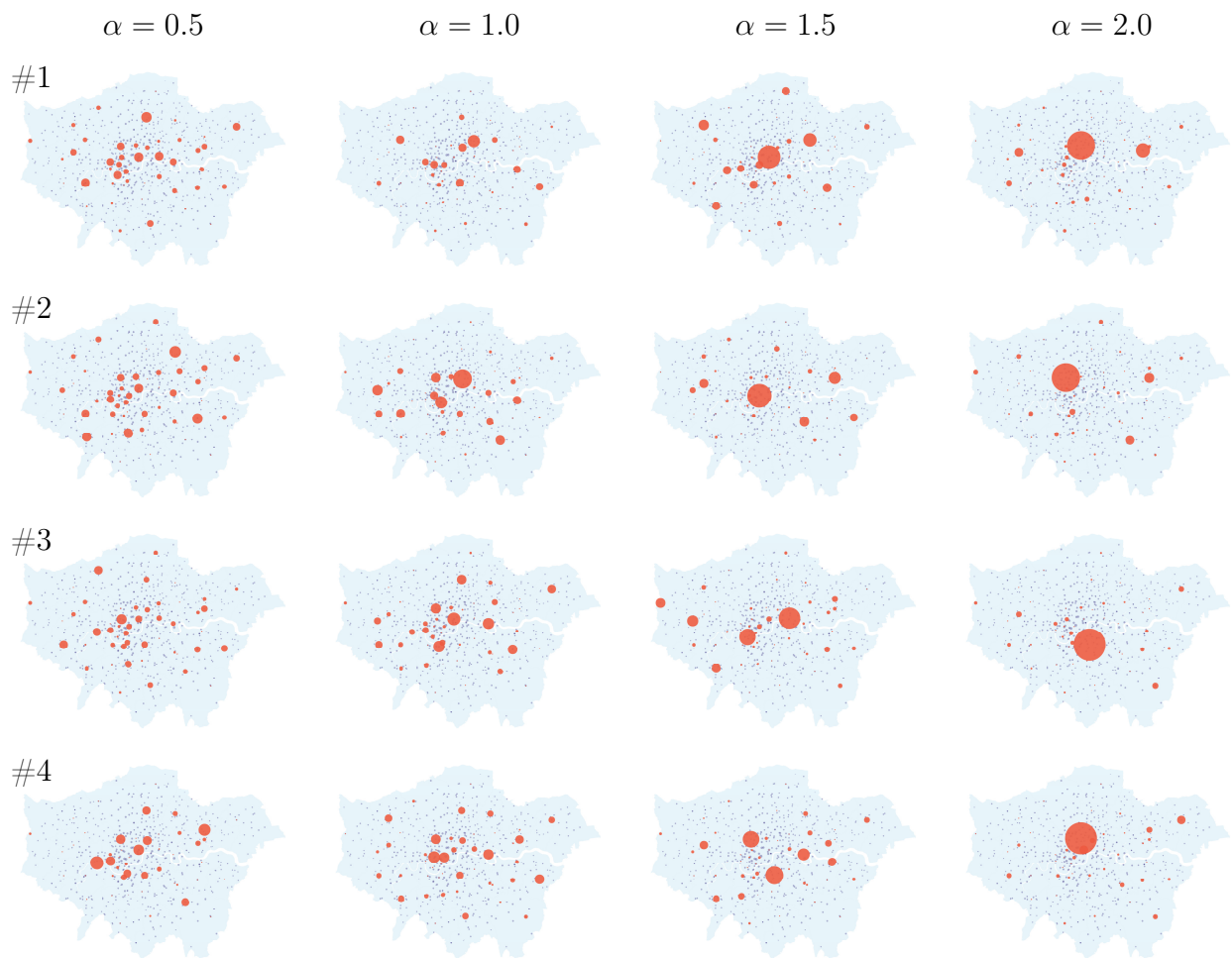


Figure 5.6: Approximate draws of the latent variables from $\pi(x|\theta)$ for a high-noise regime with $\gamma = 10^2$, obtained by running a Markov chain of length 10,000 using HMC combined with parallel tempering. Each row shows 4 randomly selected states from the Markov chain with α as specified and $\beta = 0.5$.

obtained from numerical optimisation. This is appropriate because there is little variation between samples. For the regime with high-noise, the approximate draws are obtained by running a Markov chain of length 10,000 using HMC combined with parallel tempering with 5 auxiliary distributions. The results are in Fig. 5.6-5.7, respectively. It can be seen that higher values of α and lower values of β create a sparse structure in that all activity takes place in very few zones. Conversely, lower values of α and higher values of β lead to a more dense structure.

5.5.3 Bayesian parameter estimation

This subsection investigates the parameter estimation procedures described in the previous for the observation model in (5.50) that accounts for observation noise in the data. For illustrative

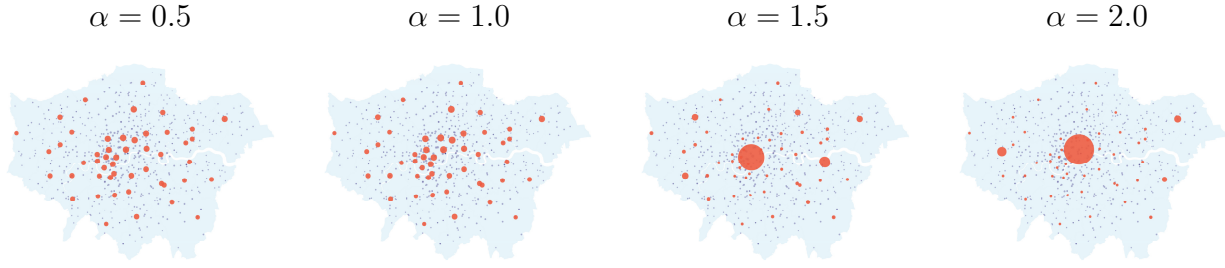


Figure 5.7: Global minima of the latent variables from $\pi(x|\theta)$, obtained by running the L-BFGS algorithm for M different initial conditions. These configurations are representative of draws from $\pi(x|\theta)$ in a low-noise regime with $\gamma \gg 1$.

purposes λ is held fixed to $\lambda = 0.1$ so that the relative noise for a zone of size $1/M$ is 3%. Relative noise here is defined as $\lambda/\log M$. It is held fixed to a low level to ensure that the likelihood is sufficiently informative for inferences to be made. Whilst an improved specification of observation noise from a preliminary study would lead to more dependable inferences, the arguments and methodology would remain the same.

Noiseless limit

First the α and β values are inferred by considering the noiseless limit $\gamma \rightarrow \infty$. For simplicity, uniform priors are placed on α and β so that a direct comparison can be performed with the preliminary R^2 analysis. The MAP estimates were found to be $\hat{\alpha} = 1.14$ and $\hat{\beta} = 0.54$. The results slightly differ from the R^2 analysis because the model performs least squares optimisation of the log-sizes rather than the sizes. Again, as shown for the R^2 analysis, it can be seen that the likelihood is maximised in the region $1 < \alpha < 2$. By plotting the residuals it is seen that they are approximately Gaussian, reaffirming the suitability of the approach and also showing that the model is sufficiently flexible to obtain a good fit.

Low noise regime

For the low noise regime, Markov chains of length 20,000 are obtained using the Metropolis-within-Gibbs scheme using Laplace approximations. Owing to the concentration of measure, it was not possible to obtain sufficiently precise importance sampling estimates to allow the debiasing scheme to be used. Therefore, a Laplace approximation of the unknown normalising constant was obtained instead. The empirical autocorrelation for α and β is below 0.2 after 25

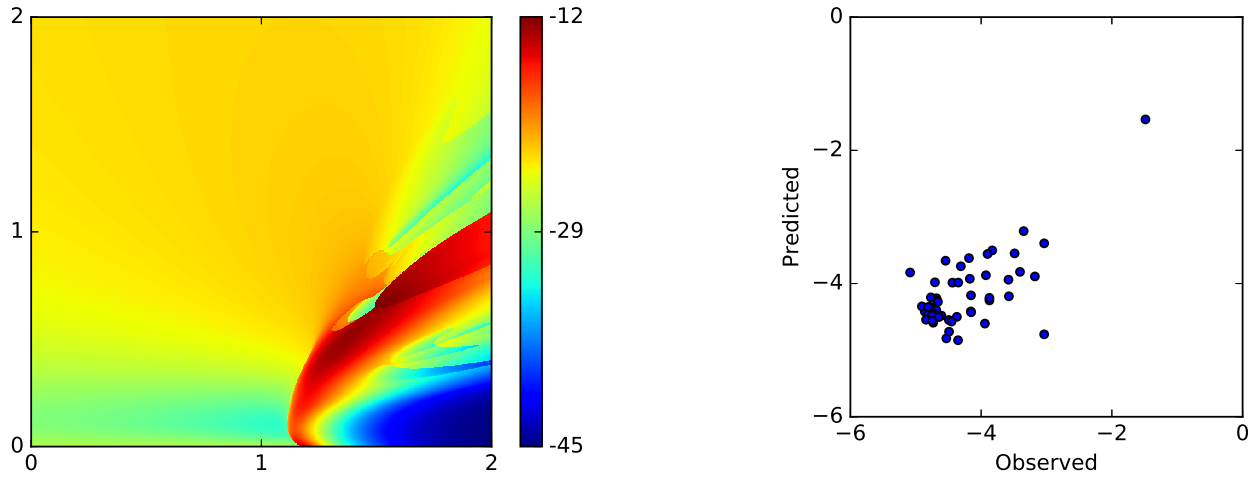


Figure 5.8: Model calibration for the deterministic model: log-likelihood values (left) and a scatter plot of actual against observed values (right).

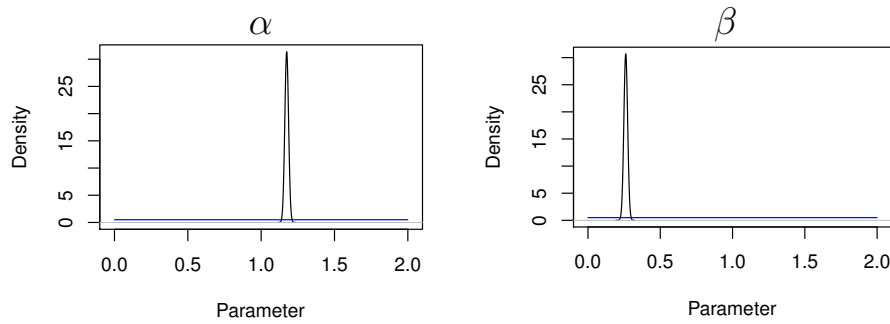


Figure 5.9: Posterior marginal density estimates for α and β for the low-noise regime ($\gamma = 10^4$). The smooth density estimates were obtained by applying (5.53) to a Gaussian kernel. The blue line indicates the uniform prior density.

steps, which is sufficient for the required posterior summaries of α and β values.

The estimated smoothed density estimates for α and β are presented in Fig. 5.9, and plots of the latent sizes showing the posterior mean plus or minus three standard deviations are presented in Fig. 5.10 alongside plots of the expected residuals and observation data. These were obtained using kernel density estimation with a Gaussian kernel. The posterior-marginals of α and β give a mean of 1.17 ± 0.01 and 0.26 ± 0.01 , respectively. The inferred values for the low-noise regime are in line with Fig. 5.5, although appear softer. The posterior summaries suggest that attractiveness and inconvenience effects are present in the data. The plots of the expected residuals and observation data suggest that the model also provides a reasonable fit to the data, however, the dispersion in the plotted quantities and possibly a degree of heteroscedasticity.

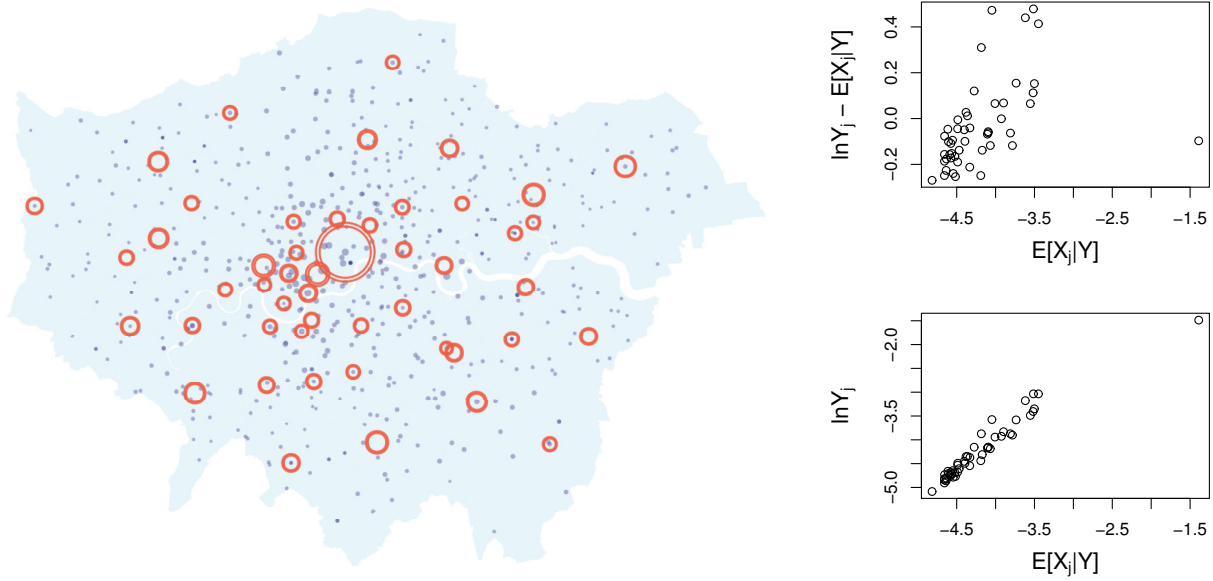


Figure 5.10: Visualisation of the posterior latent variables X for the low-noise regime ($\gamma = 10^4$). Left: The outer and inner rings show the posterior mean plus or minus three standard deviations, respectively. Top-right: Expected attractiveness against the expected residual. Bottom-right: Expected attractiveness against observed value.

High noise regime

For the high-noise regime, the pseudo-marginal MCMC methodology can be used as it is possible to obtain sufficiently reliable importance sampling estimates of the unknown normalising constants. The AIS estimates each used 10 particles and 50 equally spaced tempering parameters. Again, the empirical autocorrelation for α and β is below 0.2 after 25 steps. For pseudo-marginal MCMC scheme, 88% of the signs are positive, which is acceptably high for the scheme to

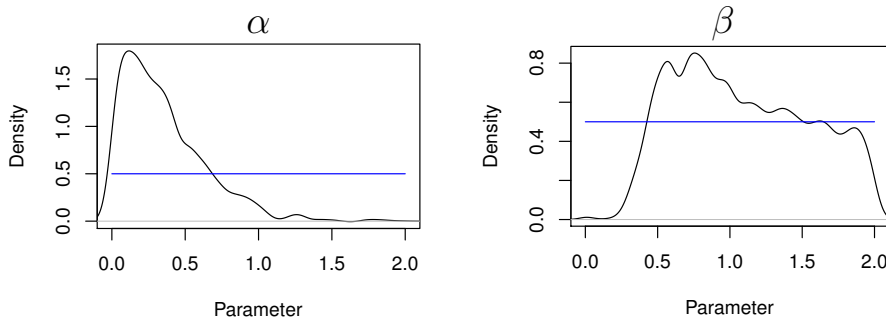


Figure 5.11: Posterior marginal density estimates for α and β for the high-noise regime ($\gamma = 10^2$). The smooth density estimates were obtained by applying (5.53) to a Gaussian kernel. The blue line indicates the uniform prior density.

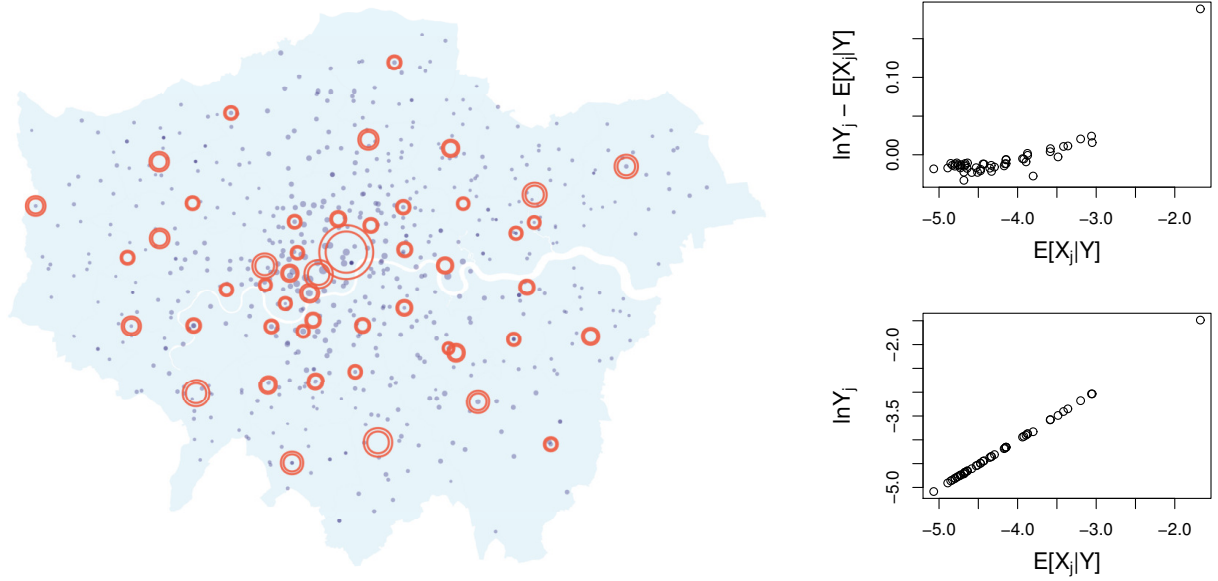


Figure 5.12: Visualisation of the posterior latent variables X for the high-noise regime ($\gamma = 10^2$). Left: The outer and inner rings show the posterior mean plus or minus three standard deviations, respectively. Top-right: Expected attractiveness against the expected residual. Bottom-right: Expected attractiveness against observed value.

be used. Plots of the smoothed density estimates for α and β for the high-noise regime are presented in Fig. 5.11. Plots of the latent sizes showing the posterior mean plus or minus three standard deviations are presented in Fig. 5.12 alongside plots of the expected residuals and observation data. The posterior-marginals of α and β give mean plus or minus one standard deviation estimates of 0.35 ± 0.28 and 1.09 ± 0.46 , respectively, which appear reasonable in light of the analysis in Fig. 5.5. The plots of the expected residuals and observation data suggest that the model provides a reasonable fit to the data, and that the assumption of homogeneous observation noise is reasonable. This is as expected as the high-noise model provides a more flexible model. After taking into account the observation noise, a weaker attractiveness effect was observed.

The model with more noise favours the simpler explanation that most variation is due to stochastic growth, whereas the low-noise model is more constrained. An SDE model that is predominately driven by stochastic effects may be of less use for exploring possible urban futures. There is noticeably more uncertainty in the latent variables and the model parameters for the high-noise regime, as there are more possible explanations for the observation data. The uncertainty in the α and β estimates is so great for the high-noise regime that limited insights

are gained for the purposes of model calibration. It is concluded that strong assumptions are required in the prior modelling in order to be able to exploit known structure in the data generating process. The required assumptions can be made, for example, through the prior modelling of α and β or by specifying a high value of γ . On the other hand, the low-noise regime results in very confident posteriors. Although the resulting inferences are consistent with previous studies, care must be taken to avoid being overconfident in a particular model by not adequately accounting for uncertainty in the model.

5.5.4 Forward evolution of urban structure

Lastly, the system response is investigated under two different noise regimes for $\sigma = 0.02$ and $\sigma = 0.04$, respectively. For each, 1,000 path trajectories were simulated according to (5.19). Each trajectory was simulated using the tamed Euler method [HJK12, WG13]. The tamed Euler method is stable for SDEs that have non-globally Lipschitz drift coefficients and satisfy a one-sided growth condition, such as the one being considered. For demonstration purposes, the system is initialised at the calibrated value in the noiseless limit, so that $x_0 = m_\theta$ and $\rho(x, 0) = \delta(x - x_0)$ with θ held at the maximum a posteriori value $\hat{\theta}$. Monte Carlo expectations have also been used to estimate the 5th and 95th percentiles of the sizes at each given time. The results are shown in Fig. 5.13. It is seen that for $\sigma = 0.02$, the urban system does not escape the potential well it is initialised within during the finite simulation time. On the other hand, for $\sigma = 0.04$, the system occasionally escapes the first potential well during the finite simulation time, and there is much greater uncertainty at $t = 10$.

5.6 Discussion and conclusions

A novel stochastic model has been developed to simulate realistic configurations of urban and regional structure, and a new probabilistic model of urban structure has been proposed. The stochastic model is a substantial improvement on existing deterministic models as it fully addresses the uncertainty associated with the evolution of urban structure. The probabilistic model of urban structure can be derived by constrained entropy maximisation using economic constraints on total welfare, total capacity and total attractiveness. It is the stationary distribution of the newly proposed SDE model.

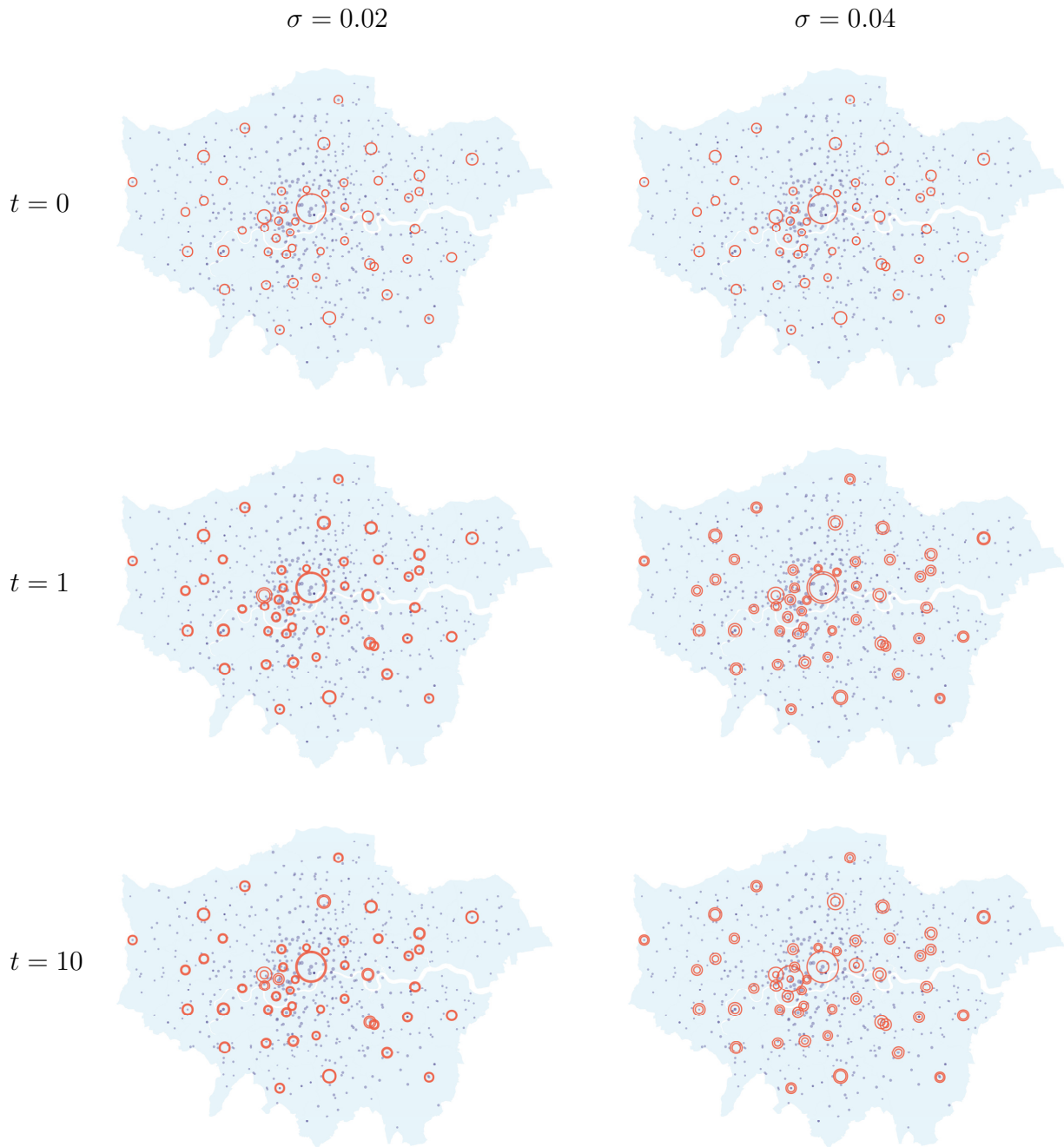


Figure 5.13: Forward evolution of the London retail system for two illustrative noise regimes. Each column shows a different noise regime, and each row shows a different time. The inner and outer red rings show 5% and 95% percentiles, respectively. The light blue circles show the residential areas.

It was shown that the Boltzmann-Gibbs measure can be used to simulate realistic configurations of urban structure using MCMC methods. It has been demonstrated that the model can be used to infer the components of a utility function from observed structure, thereby providing an alternative range of methods for model calibration. The key advantage of taking this approach is that there is no need to collect vast amounts of flow data. Whilst the proposed methodology has been presented in the context of consumer-led behaviour, the approach is applicable to other urban and regional settings such as archaeology, logistics, health care and crime to suggest a few.

The connection between the stochastic Harris and Wilson model and entropy maximisation can be explored by studying the time-dependent Fokker-Planck equation. For example, a conditional entropy is monotonically decreasing the dynamics converge exponentially fast to the Boltzmann-Gibbs measure. That said, given an equilibrium distribution obtained via a maximum entropy argument, there are infinitely many other stochastic processes that have the same stationary distribution. A much broader question is the inverse problem: for a given equilibrium distribution, which process best describes the dynamics? So far it has been considered that the first-order moment is constrained at equilibrium, however, enforcing the constraint for all times may lead to an improved model. The constrained Fokker-Planck equation is non-linear and non-local, and can undergo phase transitions [MB15, ENS17]. Another possible direction for further work is the investigation of mean field models [CCH14]. Both remain unexplored in the context of urban modelling.

There are many other specific areas for further research. For example, the deployment of the proposed methodology to large scale urban systems, for which there are further substantial computational challenges to overcome. The cost of a potential or gradient evaluation is $O(NM)$, however, increasing M means that the $z(\theta)$ estimates are more challenging to obtain owing to the curse of dimensionality. It remains to use the proposed methodology as part of a more realistic study with wider objectives. Lastly, it is emphasised that the methodology is only applicable to cross-sectional data. In practice, many applications of interest require processing time-series data that is highly correlated over time. In this setting, it would be necessary to solve the filtering or smoothing problem, and in doing so would also need to account for general trends and seasonality effects that are exogenous to the model.

Chapter 6

Summary and further work

The work in this thesis is motivated by the need for uncertainty quantification in mathematical models of complex phenomena. This thesis is concerned with the Bayesian framework owing to the philosophical and practical reasons discussed in Section 1.3. Bayesian principles have gained more popularity over the recent decades due to the advances in scientific computing and the increased availability of data. That said, there are computational and statistical challenges that continue to make Bayesian framework infeasible in certain cases. This thesis has provided a number of new approaches that improve upon the existing modelling capabilities. The focus has been on physical and socio-economic systems, for which there are practical applications in mind. A summary of the thesis contributions and findings is provided in Section 6.1 and some of the key limitations and areas for further work are identified in Section 6.2. Lastly, an outlook is provided in Section 6.3.

6.1 Summary of thesis

The Bayesian approach to modelling was described in Chapter 1, where it was argued that the numerical integration is the dominant task. Throughout this thesis, stochastic simulation techniques have been used to evaluate challenging integrals for their asymptotic properties. An up-to-date review of the relevant simulation methods was carried out in Chapter 2.

In Chapter 3, a new multiscale approach was proposed and investigated for the Bayesian estimation of a spatially varying parameter that is observed via a complex function. It was shown that wavelet basis functions can provide an effective representation of an unknown

spatially varying parameter, and the hierarchical structure of the basis functions can be used to improve the inference procedure. The hierarchical structure can also be used to suggest a parsimonious refinement of the model by excluding basis functions that are unlikely to improve the model fit. A number of models were proposed from coarse to fine scales that increase in complexity. The model fit can be assessed via the marginal likelihoods, which can be estimated as a by-product of the sequential Monte Carlo sampler. The proposed method was demonstrated for the task of permeability estimation in corner-to-corner groundwater flow.

A new Markov chain Monte Carlo (MCMC) scheme was proposed in Chapter 4 that is applicable to high-dimensional Gaussian models whose covariance is specified by the model parameters. A number of auxiliary random variables are introduced to replace the determinant term that arises in the Metropolis-Hastings acceptance probability. A number of empirical studies were performed. The Markov chains for the proposed method were found to mix more slowly per iteration than for a standard MCMC using a Cholesky-based approach, but can be much cheaper per iteration. Most importantly, the new method allows for a larger class of problems to be approached by MCMC methods.

Chapter 5 addresses the need for a stochastic generalisation of the Harris and Wilson model that captures the uncertainty associated with the evolution of urban structure. The Harris and Wilson model has been reformulated as a system of well-posed stochastic differential equations (SDEs), and the stationary distribution of the SDE model has been used as a probabilistic model of urban structure. The stationary distribution of the stochastic model is related to a maximum entropy argument as the conditional entropy of the attractiveness variables increases over time. An associated Bayesian inference methodology is proposed to estimate the unknown model parameters. Sampling from the posterior distribution is a non-trivial task and is a doubly-intractable problem. A number of approaches to inference were explored with a case study on the London retail system.

6.2 Limitations and further work

This thesis has provided a number of new approaches that improve upon the existing modelling capabilities. A common concern for each approach is scalability, owing to cost of evaluating the respective density functions and the need to explore a large state space. Scalability is a particular concern when working with large datasets and expensive computer models owing to the computational costs involved. Whilst a number of new Bayesian approaches have been proposed in this thesis, further work is required to apply the proposed methods to large-scale real-world problems with wider objectives. In practice, this may involve dealing with more complex datasets that are, for example, heterogeneous, multiscale and/or correlated in both space and time. The work in this thesis has been restricted to MCMC methods to obtain estimators that have the property of being asymptotically exact. In practice, it may be advantageous to use other approximations, such as variational inference, so that inference can be performed for larger systems. Several areas for further work are discussed below.

A new multiscale approach to infer a spatially varying parameter was proposed and investigated in Chapter 1. Whilst the new approach has been investigated using synthetic data – so that inferences can be compared with the ground truth – it remains to apply the new methodology using a real-world dataset. In doing so, care must be taken when specifying the hyperparameters of the unknown parameter. In practice, it is beneficial to use any available direct measurement data to help determine suitable hyperparameter values, for example, as done in [FL07]. Whilst multiscale methods can lead to an improved exploration of the parameter space, it can be necessary to reduce the computational burden at the coarser scales so that a more expensive forward solver can be used at the finest scale. For example, a coarser mesh could be used to approximate the forward model at the coarser scales [Kou09]. Alternatively, expensive forward solvers can be replaced by inexpensive surrogate models [MZ09, BZKL13, TB18] or posterior approximations can be used at the coarser scales [TBKZ16]. It also remains to explore the use of Bayes factors for model selection as part of other multiscale approaches.

A new MCMC method to infer the covariance parameters for a high-dimensional Gaussian models is studied in Chapter 2. The new method introduces auxiliary variables that replace the determinant term in the Gaussian probability density function. Iterative methods are used to avoid storing dense matrices and to exploit fast matrix-vector products. A key concern is that

the covariance matrix representing the latent process is sparse but also well conditioned. Since latent processes require a degree of smoothness, the covariance matrix for large systems tends to be poorly conditioned and almost rank deficient. It remains to explore ways to exploit the low rank structure more generally in larger datasets, for example, as done using a multiscale decomposition in [VV08]. It also remains to explore the methodology for parameter estimation in stochastic differential equation models that have a Gaussian representation, and also for Gaussian models that have a non-Gaussian likelihood. It may be beneficial to introduce auxiliary variables to obtain more efficient variational approximations. These approximations can potentially be combined with MCMC methods to preserve asymptotic exactness, for example, as done using psuedo-marginal MCMC in [FG14].

A new stochastic model of urban structure is proposed in Chapter 5. It is proposed that the stationary distribution of a system of SDEs is used for parameter estimation. Whilst assuming that the SDE model is at statistical equilibrium is a simplifying assumption, it can be justified in lieu of any additional knowledge about the system by the principle of maximum entropy. Further work is needed to address the more challenging non-equilibrium case. It would be necessary to solve the filtering or smoothing problem to account for time series data, which will contain general trends and seasonality effects that are exogenous to the SDE model. It may also be beneficial use likelihood free methods such as approximate Bayesian computation to address difficulties in specifying an appropriate likelihood function. There are infinitely many other stochastic processes that have the same stationary distribution. The much broader task of finding a process that best describes the trend to equilibrium remains an open problem. So far it has been considered that the first-order moment is constrained at equilibrium, however, enforcing a similar constraint for all times may lead to an improved model. The constrained Fokker-Planck in the context of urban modelling is currently unexplored.

6.3 Closing remarks

Vast efforts have been made over previous decades to develop mathematical models that describe phenomena of interest. More recently, datasets have become routinely available in almost all areas of science and engineering. The improved availability of data has been matched with enormous improvements in the computational tools available to statisticians. A research

challenge for the mathematical sciences is the development of principled methodologies that combine the available datasets with the existing mathematical models.

Despite the appeal of the Bayesian framework, in many cases Bayesian inference is a computationally challenging task. A key concern throughout this thesis has been exploring complex probability distributions. The computational challenges are exacerbated when the probability density functions involved are expensive to evaluate, typically due to the computer operations required by a mathematical model or its approximation, or due to working with a large number of observations. The development of efficient sampling methods and their application continues to be an active area of research. In fact, the inference methodology in Chapter 5 was only possible due to recent advances in Monte Carlo methods for doubly-intractable distributions. The development of more complex statistical models motivates further work in sampling methods.

With the arrival of larger datasets – or the ability to obtain these datasets – there is an inevitable dilemma over how much reliance should be placed on a mathematical model versus a statistical model. The principle of parsimony suggests a preference towards a simpler explanation of the data, however, computational matters should also be given due consideration. For example, a mathematical model might explain the hidden structure in the observation data, however, with sufficient data, it may also be possible identify a less expensive surrogate model that can be used instead. For example, a Gaussian process or neural network may be trained using simulation data to replace the fluid flow simulator in Chapter 3. The investigation of machine learning techniques to replace the retail models in Chapter 5 is an unexplored area of research. It may be beneficial to use a more flexible class of models to address the missing dynamics in the existing retail models. The modelling dilemma also arises for the Gaussian process models encountered in Chapter 4; the mean or covariance functions can be specified according to a mathematical model.

Additional complexity is arising from the need to process multiple (and larger) datasets that require a more comprehensive explanation. Notwithstanding the additional complexity, the task of modelling evermore complex datasets is likely to require simplifications that adequately represent the subject matter.

Bibliography

- [ABH03] Kim E. Andersen, Stephen P. Brooks, and Martin B. Hansen. Bayesian inversion of geoelectrical resistivity data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):619–642, 2003.
- [ACS17] Mihai Anitescu, Jie Chen, and Michael L. Stein. An inversion-free estimating equations approach for Gaussian process Models. *Journal of Computational and Graphical Statistics*, 26(1):98–107, 2017.
- [AEP13] Erlend Aune, Jo Eidsvik, and Yvo Pokern. Iterative numerical methods for sampling from high dimensional Gaussian distributions. *Statistics and Computing*, 23(4):501–521, 2013.
- [AKLR07] Jørg E. Aarnes, Vegard Kippe, Knut-Andreas Lie, and Alf B. Rustad. Modelling of multiscale structures in flow simulations for petroleum reservoirs. In *Geometric Modelling, Numerical Simulation, and Optimization*, pages 307–360. Springer, 2007.
- [AR09] Christophe Andrieu and Gareth O. Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37(2):697–725, 2009.
- [ASE14] Erlend Aune, Daniel P. Simpson, and Jo Eidsvik. Parameter estimation in high dimensional Gaussian distributions. *Statistics and Computing*, 24(2):247–263, 2014.
- [Aun12] Erlend Aune. *Computation and modeling for high dimensional Gaussian distributions*. PhD thesis, Norwegian University of Science and Technology, 2012.
- [Aut09] Greater London Authority. 2009 London town centre health check analysis report. Technical report, 2009.

- [Aut15] Greater London Authority. Household income estimates for small areas. data.london.gov.uk, 2015.
- [Aut16a] Greater London Authority. Statistical GIS boundary files for London. data.london.gov.uk, 2016.
- [Aut16b] Greater London Authority. Ward profiles and atlas. data.london.gov.uk, 2016.
- [AZ19] Steven Atkinson and Nicholas Zabaras. Structured Bayesian Gaussian process latent variable model: applications to data-driven dimensionality reduction and high-dimensional inversion. *Journal of Computational Physics*, 383(15), 2019.
- [Bat08] Michael Batty. The size, scale, and shape of cities. *Science*, 319(5864):769–771, 2008.
- [Bat13] Michael Batty. *The new science of cities*. Mit Press, 2013.
- [BBLG17] Michael Betancourt, Simon Byrne, Sam Livingstone, and Mark Girolami. The geometric foundations of Hamiltonian Monte Carlo. *Bernoulli*, 23(4A):2257–2298, 2017.
- [BCG08] Dominique Bakry, Patrick Cattiaux, and Arnaud Guillin. Rate of convergence for ergodic continuous Markov processes: Lyapunov versus poincaré. *Journal of Functional Analysis*, 254(3):727–759, 2008.
- [BCG⁺12] William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M. Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, 2012.
- [BCORR96] Luciane Bonet-Cunha, Dean S. Oliver, Richard A. Redner, and Albert C. Reynolds. A hybrid Markov chain Monte Carlo method for generating permeability fields conditioned to multiwell pressure data and prior information. In *SPE Annual Technical Conference and Exhibition*. Society of Petroleum Engineers, 1996.
- [BFG96] Zhaojun Bai, Gark Fahey, and Gene Golub. Some large-scale matrix computation problems. *Journal of Computational and Applied Mathematics*, 74(1):71–89, 1996.

- [BGHM95] Julian Besag, Peter Green, David Higdon, and Kerrie L. Mengersen. Bayesian computation and stochastic systems. *Statistical science*, 10(1):3–41, 1995.
- [BHK07] Wolfgang Bangerth, Ralf Hartmann, and Guido Kanschat. deal.II – a general purpose object oriented finite element library. *ACM Transactions on Mathematical Software*, 33(4):24/1–24/27, August 2007.
- [BHMM05] Franco Brezzi, Thomas J.R. Hughes, L. Donatella Marini, and Arif Masud. Mixed discontinuous Galerkin methods for Darcy flow. *Journal of Scientific Computing*, 22(1-3):119–145, 2005.
- [BJL⁺17] Alexandros Beskos, Ajay Jasra, Kody Law, Raul Tempone, and Yan Zhou. Multi-level sequential Monte Carlo samplers. *Stochastic Processes and their Applications*, 127(5):1417–1440, 2017.
- [BJMS15] Alexandros Beskos, Ajay Jasra, Ege A. Muzaffer, and Andrew M. Stuart. Sequential Monte Carlo methods for Bayesian elliptic inverse problems. *Statistics and Computing*, 25(4):727–737, 2015.
- [But07] Ronald W. Butler. *Saddlepoint approximations with applications*, volume 22. Cambridge University Press, 2007.
- [BZKL13] Ilias Bilonis, Nicholas Zabaras, Bledar A. Konomi, and Guang Lin. Multi-output separable Gaussian process: Towards an efficient, fully Bayesian paradigm for uncertainty quantification. *Journal of Computational Physics*, 241:212–239, 2013.
- [CCH14] José A. Carrillo, Young-Pil Choi, and Maxime Hauray. The derivation of swarming models: mean-field limit and Wasserstein distances. In *Collective dynamics from bacteria to crowds*, pages 1–46. Springer, 2014.
- [CGST11] Andrew Cliffe, Mike B. Giles, Robert Scheichl, and Aretha L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Computing and Visualization in Science*, 14(1):3, 2011.
- [Cho02] Nicolas Chopin. A sequential particle filter method for static models. *Biometrika*, 89(3):539–552, 2002.

- [Cho04] Nicolas Chopin. Central limit theorem for sequential Monte Carlo methods and its application to Bayesian inference. *The Annals of Statistics*, 32(6):2385–2411, 2004.
- [CJ08] Noel Cressie and Gardar Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226, 2008.
- [CKS06] Daniela Calvetti, Jari P. Kaipio, and Erkki Somersalo. Aristotelian prior boundary conditions. *International Journal of Mathematics and Computer Science*, 1:63–81, 2006.
- [CMSM12] Pierre Collet, Servet Martínez, and Jaime San Martín. *Quasi-stationary distributions: Markov chains, diffusions and dynamical systems*. Springer Science & Business Media, 2012.
- [CNP18] Emanuele Crosato, Ramil Nigmatullin, and Mikhail Prokopenko. On critical dynamics and thermodynamic efficiency of urban transformations. *Royal Society Open Science*, 5(10), 2018.
- [CRSW13] Simon L. Cotter, Gareth O. Roberts, Andrew M. Stuart, and David White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, pages 424–446, 2013.
- [CS07] Daniela Calvetti and Erkki Somersalo. *An introduction to Bayesian scientific computing: ten lectures on subjective computing*, volume 2. Springer Science and Business Media, 2007.
- [CVGO11] Luis A. Clementi, Jorge R. Vega, Luis M. Gugliotta, and Helcio R.B. Orlande. A Bayesian inversion method for estimating the particle size distribution of latexes from multiangle dynamic light scattering measurements. *Chemometrics and Intelligent Laboratory Systems*, 107(1):165–173, 2011.
- [Dau92] Ingrid Daubechies. *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, 1992.

- [DEH08] Paul Dostert, Yalchin Efendiev, and Thomas Y. Hou. Multiscale finite element methods for stochastic porous media flow equations and application to uncertainty quantification. *Computer Methods in Applied Mechanics and Engineering*, 197(43-44):3445–3455, 2008.
- [Dem97] James W. Demmel. *Applied numerical linear algebra*, volume 56. Society for Industrial and Applied Mathematics, 1997.
- [DEM09] Paul Dostert, Yalchin Efendiev, and Binayak Mohanty. Efficient uncertainty quantification techniques in inverse problems for Richards’ equation using coarse-scale simulation models. *Advances in water resources*, 32(3):329–339, 2009.
- [DEN⁺17] Kun Dong, David Eriksson, Hannes Nickisch, David Bindel, and Andrew G. Wilson. Scalable log determinants for Gaussian process kernel learning. In *Advances in Neural Information Processing Systems*, pages 6327–6337, 2017.
- [DGA00] Arnaud Doucet, Simon Godsill, and Christophe Andrieu. On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and computing*, 10(3):197–208, 2000.
- [DHS12] Masoumeh Dashti, Stephen J. Harris, and Andrew M. Stuart. Besov priors for Bayesian inverse problems. *Inverse Problems and Imaging*, 6(2):183–200, 2012.
- [DKPR87] Simon Duane, Anthony D. Kennedy, Brian J. Pendleton, and Duncan Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987.
- [DMDJ06] Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.
- [DMRT13] Peter J. Diggle, Paula Moraga, Barry Rowlingson, and Benjamin M. Taylor. Spatial and Spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563, 2013.
- [DS16] Matthew M. Dunlop and Andrew M. Stuart. The Bayesian formulation of EIT: Analysis and algorithms. *Inverse Problems & Imaging*, 10(4):1007–1036, 2016.

- [DS17] Masoumeh Dashti and Andrew M. Stuart. The Bayesian approach to inverse problems. *Handbook of Uncertainty Quantification*, pages 311–428, 2017.
- [DW11] Joel Dearden and Alan G. Wilson. A framework for exploring urban retail discontinuities. *Geographical Analysis*, 43(2):172–187, 2011.
- [DW15] Joel Dearden and Alan G. Wilson. *Explorations in urban and regional dynamics: a case study in complexity science*, volume 7. Routledge, 2015.
- [EDGG⁺05] Yalchin Efendiev, Akhil Datta-Gupta, Victor Ginting, Xiang Ma, and Bani Mallick. An efficient two-stage Markov chain Monte Carlo method for dynamic data integration. *Water Resources Research*, 41(12), 2005.
- [EGPW18] Louis Ellam, Mark Girolami, Grigorios A. Pavliotis, and Alan G. Wilson. Stochastic modelling of urban structure. *Proceedings of the Royal Society A*, 474(2213):20170700, 2018.
- [EH07] Yalchin Efendiev and Thomas Y. Hou. Multiscale finite element methods for porous media flows and their applications. *Applied Numerical Mathematics*, 57(5-7):577–596, 2007.
- [EH09] Yalchin Efendiev and Thomas Y. Hou. *Multiscale finite element methods: theory and applications*, volume 4. Springer Science & Business Media, 2009.
- [EHL06] Yalchin Efendiev, Thomas Y. Hou, and Wuan Luo. Preconditioning Markov chain Monte Carlo simulations using coarse-scale models. *SIAM Journal on Scientific Computing*, 28(2):776–803, 2006.
- [EHW14] Ahmed H. Elsheikh, Ibrahim Hoteit, and Mary F. Wheeler. Efficient Bayesian inference of subsurface flow models using nested sampling and sparse polynomial chaos surrogates. *Computer Methods in Applied Mechanics and Engineering*, 269:515–537, 2014.
- [ENS17] Simon Eberle, Barbara Niethammer, and André Schlichting. Gradient flow formulation and longtime behaviour of a constrained Fokker–Planck equation. *Nonlinear Analysis*, 158:142–167, 2017.

- [ESGM17] Louis Ellam, Heiko Strathmann, Mark Girolami, and Iain Murray. A determinant-free method to simulate the parameters of large Gaussian fields. *Stat*, 6(1):271–281, 2017.
- [Eva98] Lawrence C. Evans. *Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, 1998.
- [Eve03] Geir Evensen. The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.
- [EZG16] Louis Ellam, Nicholas Zabaras, and Mark Girolami. A Bayesian approach to multiscale inverse problems with on-the-fly scale determination. *Journal of Computational Physics*, 326:115–140, 2016.
- [FG14] Maurizio Filippone and Mark Girolami. Pseudo-marginal Bayesian inference for Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226, 2014.
- [Fie06] Stephen E. Fienberg. When did Bayesian inference become "Bayesian"? *Bayesian analysis*, 1(1):1–40, 2006.
- [FL88] Robin Flowerdew and Andrew Lovett. Fitting constrained Poisson regression models to interurban migration flows. *Geographical Analysis*, 20(4):297–307, 1988.
- [FL07] Marco A. R. Ferreira and Herbert K. H. Lee. *Multiscale Modeling : A Bayesian Perspective*. Springer Series in Statistics. Springer Science and Business Media, 2007.
- [FWN⁺15] Seth Flaxman, Andrew G. Wilson, Daniel Neill, Hannes Nickisch, and Alex Smola. Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In *International Conference on Machine Learning*, pages 607–616, 2015.
- [Gar09] Crispin Gardiner. *Stochastic methods*, volume 4. Springer, 2009.
- [GC11] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

- [GCS⁺14] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian data analysis*, volume 2. CRC Press, 2014.
- [GDGF10] Alan E. Gelfand, Peter J. Diggle, Peter Guttorp, andMontserrat Fuentes. *Handbook of Spatial Statistics*. CRC Press, 2010.
- [Gel06] Andrew Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1(3):515–534, 2006.
- [Gey91] Charles J. Geyer. Markov chain Monte Carlo maximum likelihood. In *Proceedings of the 23rd Symposium Interface*, 1991.
- [GG84] Stuart Geman and Donald Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.
- [GLPR15] Peter J. Green, Krzysztof Łatuszyński, Marcelo Pereyra, and Christian P. Robert. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Statistics and Computing*, 25(4):835–862, 2015.
- [GM94] Ulf Grenander and Michael I. Miller. Representations of knowledge in complex systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 549–603, 1994.
- [Gre95] Peter J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
- [GS97] David Gottlieb and Chi-Wang Shu. On the Gibbs phenomenon and its resolution. *SIAM review*, 39(4):644–668, 1997.
- [Has70] W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
- [HF84] Kingsley E. Haynes and A. Stewart Fotheringham. *Gravity and spatial interaction models*, volume 2. Sage, 1984.

- [HFL13] James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian processes for big data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 282–290. AUAI Press, 2013.
- [HG14] Matthew D. Hoffman and Andrew Gelman. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.
- [HHT08] Nicholas Hale, Nicholas J. Higham, and Lloyd N. Trefethen. Computing A^α , $\log(A)$ and related matrix functions by contour integrals. *Society for Industrial and Applied Mathematics Journal on Numerical Analysis*, 46(5):2505–2523, 2008.
- [HJK12] Martin Hutzenthaler, Arnulf Jentzen, and Peter E. Kloeden. Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients. *The Annals of Applied Probability*, 22(4):1611–1641, 2012.
- [HL13] Nicholas J. Higham and Lijing Lin. Matrix functions: A short course. *Matrix Functions and Matrix Equations*, 19:1–27, 2013.
- [HMS15] Insu Han, Dmitry Malioutov, and Jinwoo Shin. Large-scale log-determinant computation through stochastic Chebyshev expansions. In *International Conference on Machine Learning*, pages 908–917, 2015.
- [HN96] Koji Hukushima and Koji Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *Journal of the Physical Society of Japan*, 65(6):1604–1608, 1996.
- [Hut90] Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 19(2):433–450, 1990.
- [HW78] Britton Harris and Alan G. Wilson. Equilibrium values and dynamics of attractiveness terms in production-constrained spatial-interaction models. *Environment and planning A*, 10(4):371–388, 1978.
- [ILS15] Marco A. Iglesias, Yulong Lu, and Andrew M. Stuart. A Bayesian level set method for geometric inverse problems. *arXiv preprint arXiv:1504.00313*, 2015.

- [Jar97a] Christopher Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Physical Review E*, 56(5):5018, 1997.
- [Jar97b] Christopher Jarzynski. Nonequilibrium equality for free energy differences. *Physical Review Letters*, 78(14):2690, 1997.
- [Jay57] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.
- [Jeg96] Beat Jegerlehner. Krylov space solvers for shifted linear systems. *arXiv preprint arXiv:hep-lat/9612014*, 1996.
- [Jin08] Bangti Jin. Fast Bayesian approach for parameter estimation. *International Journal for Numerical Methods in Engineering*, 76(2):230–252, 2008.
- [JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.
- [KBJ14] Nikolas Kantas, Alexandros Beskos, and Ajay Jasra. Sequential Monte Carlo methods for high-dimensional inverse problems: A case study for the Navier–Stokes equations. *Society for Industrial and Applied Mathematics/American Statistical Association Journal on Uncertainty Quantification*, 2(1):464–489, 2014.
- [Ken04] Anthony D. Kennedy. Approximation theory for matrices. *Nuclear Physics B-Proceedings Supplements*, 128:107–116, 2004.
- [KKS00] Jari P. Kaipio, Ville Kolehmainen, Erkki Somersalo, and Marko Vauhkonen. Statistical inversion and Monte Carlo sampling methods in electrical impedance tomography. *Inverse problems*, 16(5):1487, 2000.
- [KO01] Marc C. Kennedy and Anthony O’Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.
- [Kou09] Phaedon-Stelios Koutsourelakis. A multi-resolution, non-parametric, Bayesian framework for identification of spatially-varying model parameters. *Journal of computational physics*, 228(17):6184–6211, 2009.

- [KP13] Peter E. Kloeden and Eckhard Platen. *Numerical solution of stochastic differential equations*, volume 23. Springer Science & Business Media, 2013.
- [KR10] Ali Kaveh and Hossein Rahami. Eigenvalues of the adjacency and Laplacian matrices for modified regular structural models. *International Journal for Numerical Methods in Biomedical Engineering*, 26(12):1836–1855, 2010.
- [KS06] Jari P. Kaipio and Erkki Somersalo. *Statistical and computational inverse problems*, volume 160. Springer Science and Business Media, 2006.
- [KV86] Claude Kipnis and S. R. Srinivasa Varadhan. Central limit theorem for additive functionals of reversible Markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19, 1986.
- [LGA⁺15] Anne-Marie Lyne, Mark Girolami, Yves Atchadé, Heiko Strathmann, and Daniel P. Simpson. On Russian roulette estimates for Bayesian inference with doubly-intractable likelihoods. *Statistical Science*, 30(4):443–467, 2015.
- [LHB⁺02a] Herbert K. H. Lee, Dave M. Higdon, Zhuoxin Bi, Marco A. R. Ferreira, and Mike West. Markov random field models for high-dimensional parameters in simulations of fluid flow in porous media. *Technometrics*, 44(3):230–241, 2002.
- [LHB⁺02b] Herbert K. H. Lee, Dave M. Higdon, Zhuoxin Bi, Marco A. R. Ferreira, and Mike West. Markov random field models for high-dimensional parameters in simulations of fluid flow in porous media. *Technometrics*, 44(3):230–241, 2002.
- [Lin72] Dennis V. Lindley. *Bayesian statistics, a review*, volume 2. SIAM, 1972.
- [Lin00] Dennis V. Lindley. The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):293–337, 2000.
- [Liu08] Jun S. Liu. *Monte Carlo strategies in scientific computing*. Springer Science and Business Media, 2008.
- [LM13] Andrzej Lasota and Michael C. Mackey. *Chaos, fractals, and noise: stochastic aspects of dynamics*, volume 97. Springer Science and Business Media, 2013.

- [LM15] Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods (Interdisciplinary Applied Mathematics)*. Springer, 2015.
- [LMW18] Ben Leimkuhler, Charles Matthews, and Jonathan Weare. Ensemble preconditioning for Markov chain Monte Carlo simulation. *Statistics and Computing*, 28(2):277–290, 2018.
- [LPS14] Gabriel J. Lord, Catherine E. Powell, and Tony Shardlow. *An introduction to computational stochastic PDEs*. Number 50. Cambridge University Press, 2014.
- [LRL11] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.
- [LS84] Pierre-Louis Lions and Alain-Sol Sznitman. Stochastic differential equations with reflecting boundary conditions. *Communications on Pure and Applied Mathematics*, 37(4):511–537, 1984.
- [LSZ15] Kody Law, Andrew M. Stuart, and Kostas Zygalakis. *Data Assimilation*. Springer, 2015.
- [LTT10] Guang Lin, Alexandre M. Tartakovsky, and Daniel M. Tartakovsky. Uncertainty quantification via random domain decomposition and probabilistic collocation on sparse grids. *Journal of Computational Physics*, 229(19):6995–7012, 2010.
- [MA10] Iain Murray and Ryan P. Adams. Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems*, volume 23, pages 1723–1731, 2010.
- [Mac03a] David J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.
- [Mac03b] Michael C. Mackey. *Time’s arrow: the origins of thermodynamic behavior*. Courier Corporation, 2003.

- [Mal89] Stéphane G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):674–693, 1989.
- [Mal08] Stéphane G. Mallat. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 3rd edition, 2008.
- [Mar20] Alfred Marshall. *Principles of economics: an introductory volume*. Royal Economic Society (Great Britain), 1920.
- [MB15] Masoumeh Mohammadi and Alfio Borzi. A Hermite spectral method for a Fokker-Planck optimal control problem in an unbounded domain. *International Journal for Uncertainty Quantification*, 5(3), 2015.
- [McF73] Daniel McFadden. Conditional logit analysis of qualitative choice behaviour. In *Frontiers in Econometrics*, pages 105–142. Academic Press New York, 1973.
- [MEMDG10] Anindan Mondal, Yalchin Efendiev, Bidyut Mallick, and Akhil Datta-Gupta. Bayesian uncertainty quantification for flows in heterogeneous porous media using reversible jump Markov chain Monte Carlo methods. *Advances in Water Resources*, 33(3):241–256, 2010.
- [MG05] Iain Murray and Zoubin Ghahramani. A note on the evidence and Bayesian Occam’s razor, 2005.
- [MGM06] Iain Murray, Zoubin Ghahramani, and David J. C. MacKay. MCMC for doubly-intractable distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2006.
- [MHCC18] Roman Marchant, Sebastian Haan, Garner Clancey, and Sally Cripps. Applying machine learning to criminology: semi-parametric spatial-demographic Bayesian regression. *Security Informatics*, 7(1):1, 2018.
- [MN09] Youssef M. Marzouk and Habib N. Najm. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. *Journal of Computational Physics*, 228(6):1862–1902, 2009.

- [MP92] Enzo Marinari and Giorgio Parisi. Simulated tempering: a new Monte Carlo scheme. *EPL (Europhysics Letters)*, 19(6):451, 1992.
- [MPRB06] Jesper Møller, Anthony N. Pettitt, Robert Reeves, and Kasper K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika*, 93(2):451–458, 2006.
- [MRR⁺53] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.
- [MS14] Georg Menz and André Schlichting. Poincaré and logarithmic Sobolev inequalities by decomposition of the energy landscape. *The Annals of Probability*, 42(5):1809–1884, 2014.
- [MT96] Kerrie L. Mengersen and Richard L. Tweedie. Rates of convergence of the Hastings and Metropolis algorithms. *The annals of Statistics*, 24(1):101–121, 1996.
- [MV99] Peter A. Markowich and Cedric Villani. On the trend to equilibrium for the Fokker-Planck equation: An interplay between physics and functional analysis. In *Physics and Functional Analysis, Matematica Contemporanea (SBM) 19*. Citeseer, 1999.
- [MZ09] Xiang Ma and Nicholas Zabaras. An efficient Bayesian inference approach to inverse problems based on an adaptive sparse grid collocation method. *Inverse Problems*, 25(3):035013, 2009.
- [Nea97] Radford M. Neal. Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. *arXiv preprint physics/9701026*, 1997.
- [Nea01] Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.
- [Nea11] Radford M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

- [NS16] Frank Nielsen and Ke Sun. Guaranteed Bounds on Information-Theoretic Measures of Univariate Mixtures Using Piecewise Log-Sum-Exp Inequalities. *Entropy*, 18(12), 2016.
- [NW06] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science and Business Media, 2006.
- [OCR97] Dean S. Oliver, Luciane B. Cunha, and Albert C. Reynolds. Markov chain Monte Carlo methods for conditioning a permeability field to pressure data. *Mathematical Geology*, 29(1):61–91, 1997.
- [Pav16] Grigorios A. Pavliotis. *Stochastic processes and applications*. Springer, 2016.
- [QCR05] Joaquin Quiñonero-Candela and Carl E. Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6:1939–1959, 2005.
- [QCRW07] Joaquin Quinonero-Candela, Carl Edward Rasmussen, and Christopher K. I. Williams. Approximation methods for Gaussian process regression. *Large-scale kernel machines*, pages 203–224, 2007.
- [RGG97] Gareth O. Roberts, Andrew Gelman, and Walter R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.
- [RH05] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC Press, 2005.
- [RHL14] Lassi Roininen, Janne M. J. Huttunen, and Sari Lasanen. Whittle-matérn priors for Bayesian statistical inversion with applications in electrical impedance tomography. *Inverse Problems and Imaging*, 8(2):561–586, 2014.
- [Rie97] Andreas Rieder. A wavelet multilevel method for ill-posed problems stabilized by Tikhonov regularization. *Numerische Mathematik*, 75(4):501–522, 1997.
- [Rob04] Christian P Robert. *Monte carlo methods*. Wiley Online Library, 2004.

- [Ros11] Jeffrey S. Rosenthal. Optimal proposal distributions and adaptive MCMC. *Handbook of Markov Chain Monte Carlo*, pages 93–112, 2011.
- [RR98] Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):255–268, 1998.
- [RR04] Gareth O. Roberts and Jeffrey S. Rosenthal. General state space Markov chains and MCMC algorithms. *Probability Surveys*, 1:20–71, 2004.
- [RS94] Gareth O. Roberts and Adrian F. M. Smith. Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216, 1994.
- [RS97] Gareth O. Roberts and Sujit K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317, 1997.
- [RT77] P. A. Raviart and J. M. Thomas. *Mathematical Aspects of Finite Element Methods: Proceedings of the Conference Held in Rome, December 10–12, 1975*, pages 292–315. Springer, 1977.
- [RT96a] Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.
- [RT96b] Gareth O. Roberts and Richard L. Tweedie. Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.
- [RV83] Francis J. A. Rijk and Antonius C. F. Vorst. Equilibrium points in an urban retail model and their connection with dynamical systems. *Regional Science and Urban Economics*, 13(3):383–399, 1983.
- [RW06] Carl E. Rasmussen and Chris Williams. *Gaussian processes for machine learning*. MIT Press, 2006.
- [Saa03] Yousef Saad. *Iterative methods for sparse linear systems*, volume 82. siam, 2003.

- [Saa12] Yunus Saatçi. *Scalable inference for structured Gaussian process models*. PhD thesis, Citeseer, 2012.
- [SAI17] Arvind K. Saibaba, Alen Alexanderian, and Ilse C. F. Ipsen. Randomized matrix-free trace and log-determinant estimators. *Numerische Mathematik*, 137(2):353–395, 2017.
- [SDP⁺08] Martin B. Short, Maria R. D’orsogna, Virginia B. Pasour, George E. Tita, Paul J. Brantingham, Andrea L. Bertozzi, and Lincoln B. Chayes. A statistical model of criminal behavior. *Mathematical Models and Methods in Applied Sciences*, 18(supp01):1249–1267, 2008.
- [Ser16] Metropolitan Police Service. Recorded crime: Geographic breakdown. Technical report, 2016.
- [Sim08] Daniel P. Simpson. *Krylov subspace methods for approximating functions of symmetric positive definite matrices with applications to applied statistics and anomalous diffusion*. PhD thesis, Queensland University of Technology, 2008.
- [Ski06] John Skilling. Nested sampling for general Bayesian computation. *Bayesian analysis*, 1(4):833–859, 2006.
- [Smi13] Ralph C. Smith. *Uncertainty quantification: theory, implementation, and applications*, volume 12. Society for Industrial and Applied Mathematics, 2013.
- [SS09] Matti L. E. Saksman and Samuli Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. Technical report, 2009.
- [SS12] Christopher Schwab and Andrew M. Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Inverse Problems*, 28(4):045003, 2012.
- [Sul15] Timothy J. Sullivan. *Introduction to uncertainty quantification*, volume 63. Springer, 2015.
- [SW86] Robert H. Swendsen and Jian-Sheng Wang. Replica Monte Carlo simulation of spin-glasses. *Physical review letters*, 57(21):2607, 1986.

- [Tar05] Albert Tarantola. *Inverse problem theory and methods for model parameter estimation*, volume 89. siam, 2005.
- [TB18] Rohit K. Tripathy and Ilias Bilonis. Deep uq: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *Journal of Computational Physics*, 375:565–588, 2018.
- [TBKZ16] Panagiotis Tsilifis, Ilias Bilonis, Ioannis Katsounaros, and Nicholas Zabaras. Computationally efficient variational approximations for Bayesian inverse problems. *ASME Journal of Verification, Validation and Uncertainty Quantification*, 2016.
- [Tik63] Andrei N. Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk*, volume 151, pages 501–504. Russian Academy of Sciences, 1963.
- [TLL19] Zofia Trstanova, Ben Leimkuhler, and Tony Lelièvre. Local and global perspectives on diffusion maps in the analysis of molecular systems. *arXiv preprint arXiv:1901.06936*, 2019.
- [TMM⁺15] Gašper Tkačik, Thierry Mora, Olivier Marre, Dario Amodei, Stephanie E. Palmer, Michael J. Berry, and William Bialek. Thermodynamics and signatures of criticality in a network of neurons. *Proceedings of the National Academy of Sciences*, 112(37):11508–11513, 2015.
- [Tob70] Waldo R. Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1):234–240, 1970.
- [TP18] Michalis K. Titsias and Omiros Papaspiliopoulos. Auxiliary gradient-based sampling algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):749–767, 2018.
- [Tra09] Kenneth E. Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.
- [Vor85] Antonius C. F. Vorst. A stochastic version of the urban retail model. *Environment and Planning A*, 17(12):1569–1580, 1985.

- [VV08] Jarno Vanhatalo and Aki Vehtari. Modelling local and global phenomena with sparse Gaussian processes. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pages 571–578, 2008.
- [Wei97] Iain S. Weir. Fully Bayesian reconstructions from single-photon emission computed tomography data. *Journal of the American Statistical Association*, 92(437):49–60, 1997.
- [Wen95] Holger Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in computational Mathematics*, 4(1):389–396, 1995.
- [WG13] Xiaojie Wang and Siqing Gan. The tamed Milstein method for commutative stochastic differential equations with non-globally Lipschitz continuous coefficients. *Journal of Difference Equations and Applications*, 19(3):466–490, 2013.
- [Wil67] Alan G. Wilson. A statistical theory of spatial distribution models. *Transportation research*, 1(3):253–269, 1967.
- [Wil71] Alan G. Wilson. A family of spatial interaction models, and associated developments. *Environment and Planning A*, 3(1):1–32, 1971.
- [Wil77] Huw CWL Williams. On the formation of travel demand models and economic evaluation measures of user benefit. *Environment and planning A*, 9(3):285–344, 1977.
- [Wil00] Alan G. Wilson. *Complex spatial systems: the modelling foundations of urban and regional analysis*. Pearson Education, 2000.
- [Wil08] Alan Wilson. Boltzmann, Lotka and Volterra and spatial structural evolution: an integrated methodology for some dynamical systems. *Journal of The Royal Society Interface*, 5(25):865–871, 2008.
- [Wil09] Alan Wilson. The “thermodynamics” of the city. In *Complexity and Spatial Networks*, pages 11–31. Springer, 2009.
- [Wil10] Alan Wilson. Entropy in urban and regional modelling: Retrospect and prospect. *Geographical Analysis*, 42(4):364–394, 2010.

- [Wil11] Alan G. Wilson. *Entropy in urban and regional modelling*. Routledge, 2011.
- [Wil12] Alan G. Wilson. *The science of cities and regions: lectures on mathematical model design*. Springer Science and Business Media, 2012.
- [WM17] Colin Wei and Iain Murray. Markov chain truncation for doubly-intractable inference. 54:776–784, 20–22 Apr 2017.
- [WT02] C Larrabee Winter and Daniel M. Tartakovsky. Groundwater flow in heterogeneous composite aquifers. *Water Resources Research*, 38(8), 2002.
- [WZ65] Eugene Wong and Moshe Zakai. On the convergence of ordinary integrals to stochastic integrals. *The Annals of Mathematical Statistics*, 36(5):1560–1564, 1965.
- [WZ04] Jingbo Wang and Nicholas Zabaras. A Bayesian inference approach to the inverse heat conduction problem. *International Journal of Heat and Mass Transfer*, 47(17-18):3927–3941, 2004.
- [WZ11] Jiang Wan and Nicholas Zabaras. A Bayesian approach to multiscale inverse problems using the sequential Monte Carlo method. *Inverse Problems*, 27(10):105004, 2011.
- [XT04] Dongbin Xiu and Daniel M. Tartakovsky. A two-scale nonperturbative approach to uncertainty analysis of diffusion in random composites. *Multiscale Modeling & Simulation*, 2(4):662–674, 2004.
- [ZJA16] Yan Zhou, Adam M. Johansen, and John A. D. Aston. Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.
- [ZLL05] Yunong Zhang, William E. Leithead, and Douglas J. Leith. Time-series Gaussian process regression based on Toeplitz computation of $O(n^2)$ operations and $O(n)$ -level storage. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 3711–3716. IEEE, 2005.

Appendix A

Diffusion processes and the Fokker-Planck equation

This appendix contains additional background for the diffusion processes described in Chapter 5, and is a concise summary of the relevant definitions and results that can be found in a textbook on stochastic differential equations (SDEs) [KP13, Gar09, LSZ15, Pav16]. A diffusion process is of the form

$$\frac{dX(t)}{dt} = b(X(t)) + \sigma(X(t))\xi(t), \quad X(0) = x, \quad (\text{A.1})$$

where $X(t) \in \mathbb{R}^d$, the drift function $b : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a smooth vector-valued function and the diffusion function $\sigma : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times m}$ is a smooth matrix-valued function. Here, $\xi(t) \in \mathbb{R}^m$ denotes a white noise process: it is a mean-zero, vector-valued Gaussian process with the autocorrelation function

$$E[\xi_i(t)\xi_j(t)] = \delta_{ij}\delta(t-s), \quad i, j = 1, \dots, m. \quad (\text{A.2})$$

A heuristic, but not always accurate¹, interpretation of (A.1) is that over a short time-step Δt , the stochastic process $X(t)$ changes by a Gaussian random variable with mean $b(X, t)\Delta t$ and covariance $\Sigma(X, t)\Delta t$, where the diffusivity matrix

$$\Sigma(X, t) = \sigma(X, t)\sigma(X, t)^T, \quad (\text{A.3})$$

is defined for notational convenience. The class of diffusion processes is sufficiently broad to provide a stochastic generalisation of the Harris and Wilson model. Under certain conditions it is also possible to understand some relevant properties.

¹This depends on the precise interpretation of the noise term and the properties of the drift and diffusion functions.

A.1 Formal definition

More formally, the white noise process is related to a standard m -dimensional Brownian motion $B(t)$ by $B(t) = \int_0^t d\xi$. This means that $B(t) = (B_1(t), \dots, B_m(t))^T$ is a vector of independent Brownian motions. That is, each $B_i(t)$ is a real-valued stochastic process with continuous sample paths that pass through $B_i(0) = 0$; and for every $0 \leq s < t$ the increment $B_i(t) - B_i(s)$ is an independent Gaussian random variable with mean zero and variance $t - s$. As a Brownian motion is nowhere differentiable, many authors prefer to express (A.1) in differential form by

$$dX(t) = b(X(t), t)dt + \sigma(X(t), t)dB(t). \quad (\text{A.4})$$

The precise interpretation of (A.4) is given by the integral equation

$$X(t) = x + \int_0^t b(X(s), s)ds + \int_0^t \sigma(X(s), s)dB(s). \quad (\text{A.5})$$

The second integral requires special treatment as a Brownian motion does not have bounded variation.

A.2 Stochastic integration

It is necessary to be explicit about the interpretation of stochastic integrals of the form

$$I(t) = \int_0^t G(s) dB(s), \quad (\text{A.6})$$

where $G : \mathbb{R}_{>0} \rightarrow \mathbb{R}^{d \times m}$ is a suitably defined matrix-valued random process. More specifically, let the Brownian motion B be defined on the probability space $(\Omega, \mathcal{F}, P)$ and let $\mathcal{F}_t \subset \mathcal{F}$ denote the σ -algebra defined by information only on $[0, t)$. Then assume that G is $\mathcal{F}_t$ -measurable, and therefore only depends on the Brownian motion on $[0, t)$, and is square integrable in that

$$E \left[\int_0^t \|G(s)\|_F^2 ds \right] < \infty, \quad (\text{A.7})$$

where $\|\cdot\|_F$ denotes the Frobenius norm that is defined as $\|A\|_F = \sqrt{\text{Tr}(AA^T)}$. Under these assumptions, the Itô integral of (A.6) may be defined to satisfy the following properties:

1. $I(t)$ is $\mathcal{F}_t$ -measurable and is linear in G ;

2. $E[I(t)] = 0$; and
3. $E[I(t)]^2 = E\left[\int_0^t \|G(s)\|_F^2 ds\right]$.

The Itô integral is not the only way of defining $I(t)$, however, the Itô integral is often studied because the properties listed above are mathematically convenient or appropriate modelling assumptions. It can be more appropriate to interpret the noise in a different way. An alternative is the Stratonovich integral.

A.3 Existence and uniqueness

The existence of a unique solution to the SDE in (A.4) can be established under certain assumptions. For Chapter 5, it suffices to assume that, for some $T > 0$:

- The drift function $b(\cdot, t)$ is measurable, locally Lipschitz continuous, in that for every $N > 0$ there exists $K_N > 0$ such that

$$|b(x, t) - b(y, t)| \leq K_N |x - y|, \quad \forall x, y \in \mathbb{R}^d \text{ such that } |x|, |y| \leq N \text{ and } t \in [0, T], \quad (\text{A.8})$$

and satisfies a one-sided growth condition, in that there exists $\alpha, \beta > 0$ such that

$$\langle b(x, t), x \rangle \leq \alpha + \beta |x|^2, \quad \forall x \in \mathbb{R}^d, \text{ and } t \in [0, T]. \quad (\text{A.9})$$

- The diffusion function $\sigma(\cdot, t)$ is measurable and globally Lipschitz continuous, in that there exists $K > 0$ such that

$$|\sigma(x, t) - \sigma(y, t)| \leq K |x - y|, \quad \forall x, y \in \mathbb{R}^d, \text{ and } t \in [0, T]. \quad (\text{A.10})$$

- The initial condition $X(0) = x$ is a random variable independent of the Brownian motion $B(t)$ with a finite second moment:

$$E|x|^2 < \infty. \quad (\text{A.11})$$

Under these conditions, the SDE in (A.4) has a unique solution $X \in C([0, T]; \mathbb{R}^d)$ with

$$E \left[\int_0^T |X(t)|^2 dt \right] < \infty, \quad (\text{A.12})$$

and is non-exploding on $[0, T]$, almost surely.

A.4 The Itô formula and the Fokker-Planck equation

One of the most attractive properties of SDEs is that they are Markov processes [Pav16, LSZ15]. The appeal is that a function of a Markov process can be studied via the Itô formula. It is then possible to derive the Fokker-Planck equation (FPE), which is a partial differential equation that describes the evolution of the probability distribution of the process. Whilst the FPE is notoriously difficult to solve, especially in high-dimensions, for some diffusion processes its stationary distribution is available in closed form.

The Itô formula describes the process $\varphi(X(t), t)$, where $X(t)$ solves the SDE (A.4) and $\varphi \in C^{2,1}(\mathbb{R}^d \times [0, T], \mathbb{R})$. Provided that the conditions in Section A.3 hold, it can be shown that the process $\varphi(X(t), t)$ satisfies:

$$\begin{aligned} \varphi(X(t), t) = \varphi(x, 0) &+ \int_0^t \left[\frac{\partial \varphi}{\partial t}(X(s), s) + \mathcal{L}\varphi(X(s), s) \right] ds \\ &+ \int_0^t \left\langle \nabla \varphi(X(s), s), \sigma(X(s)) dB(s) \right\rangle. \end{aligned} \quad (\text{A.13})$$

for $t \in [0, T]$, where the operator $\mathcal{L}$ is defined as

$$\mathcal{L}\varphi = f \cdot \nabla \varphi + \frac{1}{2} \Sigma : \nabla \nabla \varphi. \quad (\text{A.14})$$

A proof of this result can be found in [Pav16].

The backward Kolmogorov equation and FPE can be derived by an application of the Itô formula for the function $u(x, t) = E[f(X(t)) | X(0) = x]$, with $f \in C^2(\mathbb{R}^d)$. In doing so, one requires the martingale property $E[I(t)] = 0$. The backward Kolmogorov equation is given by

$$\frac{\partial u}{\partial t} = \mathcal{L}u, \quad (\text{A.15})$$

which is studied with the initial condition $u(x, 0) = f(x)$. The FPE is given by

$$\frac{\partial \rho}{\partial t} = \mathcal{L}^* \rho, \quad (\text{A.16})$$

which is studied with the initial condition $\rho(x, 0) = \rho_0(x)$. Here, $\rho(x, t)$ is the marginal probability density function of $X(t)$, and $\mathcal{L}^*$ is the L^2 adjoint of $\mathcal{L}$, which is given by

$$\mathcal{L}^* \varphi = -\nabla \cdot (b\varphi) + \frac{1}{2} \nabla \cdot \nabla (\Sigma \varphi). \quad (\text{A.17})$$

It is useful to make reference to the concept of ergodicity. The time average of an ergodic Markov process $X(t)$ is equal to its stationary average for any suitably defined test function $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$ in that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \varphi(X(t)) dt = \int_{\mathbb{R}^d} \varphi(x) \rho_\infty(x) dx. \quad (\text{A.18})$$

In this case, the probability density function of the stationary distribution can be studied via the FPE, since

$$-b\rho_* + \frac{1}{2}(\Sigma\rho_*) = 0. \quad (\text{A.19})$$

A.5 The Smoluchowski equation

In the case that $\Sigma = 2\gamma^{-1}I$ and $b(x) = -\nabla V(x)$, (A.19) has a closed form solution given by

$$\rho_*(x) = \frac{1}{Z} \exp(-\beta V(x)). \quad (\text{A.20})$$

Under appropriate conditions, the Markov process with generator $\mathcal{L}$ is ergodic and

$$Z = \int_{\mathbb{R}^d} \exp(-\gamma V(x)) dx < \infty. \quad (\text{A.21})$$

Here, it is assumed that the following Lyapunov condition is satisfied: there exists $b \in (0, 1)$ and $C > 0$ such that

$$\liminf_{|x| \rightarrow \infty} (b|\nabla V|^2 - \gamma^{-1} \Delta V) \geq -C. \quad (\text{A.22})$$

The condition is satisfied for any potential functions that grow at least linearly but no faster than exponentially at infinity. The Lyapunov condition is strong enough to ensure that (A.21) holds, for example, by Lemma 3.14 in [MS14]. Moreover, the Lyapunov condition implies that a

Poincaré holds: there exists $C_p > 0$, such that

$$\text{Var}_{\rho_*}(f) \leq C_p \int_{\mathcal{X}} |\nabla f|^2 \rho_* dx, \quad (\text{A.23})$$

for all $f \in C^1(\mathcal{X}) \cap L^2(\mathcal{X}; \rho_*)$, as proved by Corollary 4.1 in [BCG08].

It is more convenient to study the probability density function $\rho(x, t)$ via a change of variables that gives a backward Kolmogorov equation. The quantity $h(x, t)$ is introduced by

$$\rho(x, t) = h(x, t) \rho_*(x). \quad (\text{A.24})$$

By substituting (A.24) into (A.16), one obtains the backward Kolmogorov equation for $h(x, t)$:

$$\frac{\partial h}{\partial t} = \mathcal{L}h, \quad (\text{A.25})$$

with the initial condition $h(x, 0) = \rho_0(x) \rho_*(x)^{-1}$. There are two reasons for studying the FPE via the backward Kolmogorov equation. First, the backward operator $\mathcal{L}$ satisfies a symmetry property in $L^2(\rho_*)$:

$$\langle \mathcal{L}f, h \rangle_{\rho_*} = -\gamma^{-1} \int_{\mathbb{R}^d} (\nabla f \cdot \nabla h) \rho_* dx. \quad (\text{A.26})$$

for all $f, g \in C^1(\mathcal{X}) \cap L^2(\mathcal{X}; \rho_*)$. Moreover,

$$\langle \mathcal{L}f, f \rangle_{\rho_*} = -\gamma \|\nabla f\|_{\rho_*}^2, \quad (\text{A.27})$$

which means that $\mathcal{L}$ is a self-adjoint non-positive operator, and the nullspace of $\mathcal{L}$ comprises of the constant functions. Second, the Poincaré inequality in (A.23) can be used to describe the trend to equilibrium.

A.5.1 Convergence to equilibrium

It is now possible to show the process will converge to the unique stationary distribution exponentially fast [Pav16]. This can be shown from the backward equation for the centred function $h - 1$:

$$\frac{\partial(h - 1)}{\partial t} = \mathcal{L}(h - 1). \quad (\text{A.28})$$

The argument follows by taking the inner product of (A.28) against the test function $h - 1$, and differentiating with respect to time:

$$\begin{aligned}
\frac{d}{dt} \|h - 1\|_{\rho_*}^2 &= 2 \left\langle \frac{\partial}{\partial t} (h - 1), h - 1 \right\rangle_{\rho_*}, \\
&= 2 \langle \mathcal{L}(h - 1), h - 1 \rangle_{\rho_*}, \\
&= -2\gamma^{-1} \langle \nabla(h - 1), \nabla(h - 1) \rangle_{\rho_*}, \\
&\leq -2\gamma^{-1} C_p^{-1} \text{Var}_{\rho_*}(h - 1) \\
&= -2\gamma^{-1} C_p^{-1} \|h - 1\|_{\rho_*}^2.
\end{aligned} \tag{A.29}$$

The third line follows from the symmetry property and the fourth line is an application of the Poincaré inequality. By separating the variables and integrating from time 0 to an arbitrary time t , one obtains

$$\|h(\cdot, t) - 1\|_{\rho_*} \leq e^{-C_p^{-1}\gamma^{-1}t} \|h(\cdot, 0) - 1\|_{\rho_*}. \tag{A.30}$$

Lastly, by substituting $h(\cdot, t) = \rho(\cdot, t)\rho_*^{-1}(\cdot)$ it can be seen that

$$\|\rho(\cdot, t) - \rho_*(\cdot)\|_{\rho_*^{-1}} \leq e^{-C_p^{-1}\gamma^{-1}t} \|\rho(\cdot, 0) - \rho_*(\cdot)\|_{\rho_*^{-1}}, \tag{A.31}$$

Therefore, $\rho(\cdot, t)$ converges to the unique stationary distribution ρ_* at a rate that is exponentially fast. This result holds for any initial condition $\rho_0 \in L^2(\mathbb{R}^d, \rho_*^{-1})$.

Two remarks are in order. First, the requirement that $\rho_0 \in L^2(\mathbb{R}^d, \rho_*^{-1})$ is satisfied by a smooth but compactly supported initial condition. Dirac masses can also be considered, albeit with an appropriate mollification that make little difference to the numerical approximation of the solution. Second, owing to the exponential decay in (A.31), the system will rapidly return to equilibrium after a small perturbation away from it.

A.5.2 Connection with entropy maximisation

Lastly, it can be shown that the conditional entropy of $\rho(\cdot, t)$ with respect to ρ_* can only increase over time [Pav16]. The relevant conditional entropy is given by

$$H_c(\rho|\rho_*) = - \int_{\mathbb{R}^M} \rho \ln \left[\frac{\rho}{\rho_*} \right] dx. \tag{A.32}$$

This can be seen by defining $h(\cdot, t)$ as before, then

$$\begin{aligned}
\frac{dH_c}{dt} &= -\frac{d}{dt} \int_{\mathbb{R}^M} (h \ln h - h + 1) \rho_* dx. \\
&= -\int_{\mathbb{R}^M} \ln h \frac{dh}{dt} \rho_* dx, \\
&= \int_{\mathbb{R}^M} (\nabla \ln h \cdot \nabla h) \rho_* dx, \\
&= \int_{\mathbb{R}^M} h^{-1} |\nabla h|^2 \rho_* dx, \\
&\geq 0.
\end{aligned} \tag{A.33}$$

Moreover, the conditional entropy stops increasing when it attains its maximum value of zero when $\rho = \rho_*$. It has already been shown that process is ergodic with respect to ρ_* , and therefore

$$\lim_{t \rightarrow \infty} H_c(\rho(\cdot, t) | \rho_*) = 0. \tag{A.34}$$

The result is analogous to the second law of thermodynamics.

A.6 Other interpretations of the stochastic integral

An alternative to the Itô integral is the Stratonovich integral. It is standard notation to write “ dW ” to indicate the Itô interpretation and “ $\circ dW$ ” to indicate the Stratonovich interpretation. The Ito integral satisfies the martingale property, whereas the Stratonovich integral satisfies Newton-Leibniz calculus and is stable with respect to changes in the random terms. The latter means that Stratonovich noise can be viewed as a smooth approximation of white noise [WZ65, Pav16].

It can be useful to switch between Ito and Stratonovich SDEs. For example, the Itô SDE

$$dX = b(X)dt + \sigma(X)dW, \tag{A.35}$$

can be written as the Stratonovich SDE

$$dX = \left(b(X) - h(X) \right) dt + \sigma(X) \circ dW, \tag{A.36}$$

where $h : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is known as the Itô-to-Stratonovich correction, whose i th component is given

by

$$h_i(x) = \frac{1}{2} \sum_{j=1}^d \sum_{k=1}^m \sigma_{jk}(x) \frac{\partial \sigma_{ik}}{\partial x_j}(x). \quad (\text{A.37})$$

The two interpretations coincide when the noise is additive - this is the case where σ is a constant-valued function.