IMPERIAL COLLEGE LONDON

Interpretable Models for Spatially Dependent and Heterogeneous Phenomena

by

Jan Povala

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the
Faculty of Natural Sciences
Department of Mathematics

June 22, 2022

IMPERIAL COLLEGE LONDON

Abstract

Faculty of Natural Sciences

Department of Mathematics

Doctor of Philosophy

by Jan Povala

Over the past decades, we have seen an increase in the availability of data that includes spatial information. Incorporating spatial information in models may result in performance improvements, which may then be used to better inform decision-making processes. When modelling spatial data, typical assumptions such as independence of observations across locations, no longer hold. As a consequence, careful methodology is required. This thesis addresses the modelling of two common types of data encountered in spatial modelling: measurements of a quantity at pre-specified locations (*e.g.*, sensor measurements), and events for which geographical location and time are recorded. We develop effective approaches for modelling spatial data in an interpretable manner, thus making it suitable for application domains where the transparency of a model is a desired property. We demonstrate the developed approaches with empirical simulation studies.

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged in accordance with the standard referencing practices of the discipline.

Jan Povala
May, 2022

Copyright

Acknowledgements

I would like to express gratitude to the following people and institutions:

1. My supervisors Professor Niall Adams and Professor Mark Girolami for guidance and support.

2. My collaborators on projects included in this thesis – Dr Ieva Kazlauskaite, Dr Eky Febrianto, Prof. Fehmi Cirak, Dr Seppo Virtanen – for all the help and advice.

3. Dr Lara Vomfell for the collaboration on a crime reporting project which is not part of this thesis but has contributed to my development as a researcher.

4. Engineering and Physical Sciences Research Council for funding my MRes and PhD studies.

5. Imperial College Mathematics Department for its exceptional faculty, students, and a productive environment.

6. The Alan Turing Institute for the people I have met there, research seminars, and exceptional facilities.

7. My brother Pavol for encouraging me throughout my doctorate and for his feedback on parts of this thesis.

8. My family and friends – thanks for always being there.

9. My thesis examiners – Dr Andrew Duncan, Professor Finn Lindgren – for their thorough feedback and discussion.

Contents

List of Figures

List of Tables

Abbreviations

ELBO	**E**vidence **L**ower **B**ound
FCVB	**F**ull-**C**ovariance parametrisation for **VB**
FEM	**F**inite **E**lement **M**ethod
GLM	**G**eneralised **L**inear **M**odel
GP	**G**aussian **P**rocess
HMC	**H**amiltonian **M**onte **C**arlo
i.i.d.	**i**ndependetly and **i**dentically **d**istributed
IMP	Covariate **IMP**ortance measure
KL	**K**ullback-**L**eibler
LGCP	**L**og-**G**aussian **C**ox **P**rocess
LSOA	**L**ower **S**uper **OA**
MCMC	**M**arkov **C**hain **M**onte **C**arlo
MFVB	**M**ean-**F**ield paramtetrisation for **VB**
MSOA	**M**iddle **S**uper **OA**
MH	**M**etropolis **H**astings
OA	**O**utput **A**rea
PAI	**P**redictive **A**ccuracy **I**ndex
pCN	**p**reconditioned **C**rank-**N**icholson
PDE	**P**artial **D**ifferential **E**quation
PI	**P**redictive **E**fficiency **I**ndex
PMVB	**P**recision **M**atrix paramtetrisation for **VB**
RMSE	**R**oot **M**ean **S**quare **E**rror
SAM-GLM	**S**patially **A**ware **M**ixture of **GLM**s
SVI	**S**tochastic **VI**
VB	**V**aritational **B**ayes
VI	**V**aritational **I**nference

Symbols

D	domain
$\bar{D}$	closure of domain D
$C^r(D, Y)$	Space of functions whose derivatives up to and including order r are continuous
$C^r(D)$	$C^r(D, \mathbb{R})$
$C_c^\infty(D)$	compact, bounded continuous function from D to $\mathbb{R}$
$L^p(D, Y)$	Lebesgue space of the p-th order of functions from D to Y
$L^p(D)$	$L^p(D, \mathbb{R})$

$\mathcal{D}^\alpha$	Weak derivative of order α

κ	log-diffusivity parameter
λ	intensity

$\boldsymbol{\theta}$	vector of model parameters

$\mathcal{L}$	operator from a vector space to another vector space
$\mathcal{N}$	Gaussian distribution (multi-variate, possibly infinite-dimensional)
a.e.	almost everywhere (in the measure-theoretic sense)
a.s.	almost surely (in the probability measure sense)

Dedicated to my family. Mami a Oci, veľká vďaka za všetko!

Chapter 1

Introduction

Data collection for many phenomena studied in science and engineering involves recording spatial information for the purposes of providing more context for the rest of the collected data. Examples of such processes include recording census data, mapping crime locations in a city, measuring properties of a material at specified locations, and collecting sensor measurements for air equality monitoring. Data with spatial information is not a new concept (Krige 1951, Ripley 1977, Cressie 1993), but the increasing availability of this kind of data calls for effective techniques to leverage its full potential (Gelfand 2010, Girolami 2020). Despite advances in making these measurements more available, the collected spatial data is still 'sparse': census takes place only once every 10 years; data with sensitive information must be aggregated to protect anonymity; for many social phenomena there are no repeated measurements; deploying numerous sensors can be costly; in domains where the object of interest is continuous such as physical materials, we can take measurements only in a finite number of locations. For these reasons, we require careful methodology to build effective spatial models. This thesis aims to contribute to the methodology of modelling spatial data with strong emphasis on interpretability of the resulting models, motivated by application domains such as criminology and materials science, where interpretability is desired.

In Section 1.1, we introduce the main tenets of spatial models, different forms of spatial data, and the approaches to model them in an interpretable manner. Section 1.2 summarises the contributions made in my doctoral research. The rest of the thesis is structured as follows. In Chapter 2, we give an overview of statistical inference and give details of the methods used in this thesis. Chapters 4 and 6 include two original contributions, while Chapters 3 and 5 provide prerequisite background for Chapters 4 and 6, respectively. Chapter 7 concludes the thesis.

1.1 Spatial Models

Spatial information provides context for collected data. Using this information effectively can improve modelling of phenomena occurring over a spatial domain. Below, we discuss two important aspects of the spatial context – spatial dependence and spatial heterogeneity. Subsequently, we discuss relevant modelling approaches.

1.1.1 Spatial Dependence and Spatial Heterogeneity

The overarching principle for building spatial models is the first law of geography: "everything is related to everything else, but near things are more related than distant things" (Tobler 1970). This principle of *spatial dependence* manifests itself in different ways. The most common way is through clustering as a result of most phenomena not occurring with complete spatial randomness (Gelfand 2010). An example of this is increased intensity of crime at a neighbourhood of a city due to socio-economic factors. Spatial dependence can also be a result of interaction between events – an event triggering subsequent events. Similarly, in physical systems, the physical properties of systems often behave continuously – they do not change rapidly from one location to the next. For example, atomic structure of a material imposes spatial dependence of the properties of the material.

Another important aspect of spatial data that needs to be carefully considered is *heterogeneity*. This encompasses accounting for systematic differences across space without relying on spatial dependence. Generally, we expect to see different regimes in different parts of the modelling domain, especially if the domain is large, but there are also phenomena where the regime change is sudden and non-smooth. For example, natural boundaries such as rivers may produce two neighbourhoods that have distinctly different social composition (Piquero & Weisburd 2010) or a crack in a material can create discontinuities in the properties of the material.

As part of the modelling approach, *interpretability* and *uncertainty quantification* are key tenets of the methodology we develop in this thesis. We emphasise the importance of positing models with interpretable data-generating processes which are driven by the domain knowledge, and of quantifying uncertainty of the inferences being made.

1.1.2 Modelling

In many contexts, the typical modelling assumptions such as independence no longer hold, and the methodology needs to account for that. The measurements of spatial phenomena we would like to model, or use as part of the model, come in different forms and this determines the choice of methodology. We consider three types of spatial data: *point patterns*, *sensor measurements*, and aggregated *areal data*.

1. *Point patterns.* This form of data records the location, and potentially the time, of an event that has occurred, for example, a crime occurrence for which the location and the time is recorded. These measurements are a realisation of a spatial point process (Daley & Vere-Jones 2003). A common objective of modelling point processes is inferring the intensity of the events across the space or space-time. Additionally, one may also be interested in understanding what factors, such as exogenous environmental indicators, are driving the variation in the intensity across the space. One of the drivers could be interactions between the events themselves – this class of models is called self-exciting point processes (Hawkes 1971).

2. *Sensor measurements.* The quantity of interest is measured by a set of sensors at fixed locations. The measurements are often repeated. An example of sensor measurements in a spatial context is monitoring air quality in a city through sensor networks. In structural mechanics experiments, material properties are measured at a pre-specified set of locations. Depending on the context, the modelling objectives are either predicting a quantity of interest at locations where measurements are not available (interpolation or extrapolation), or inferring a quantity of interest of which we can take only indirect measurements. The latter class of problems is called *inverse problems*.

3. *Aggregated areal data*: as is very common in social sciences, individual measurements of the quantity of interest are aggregated into areal units. This is often done

to ensure privacy and security of individual citizens. The field of spatial econometrics and social sciences relies heavily on this form of data (Anselin 2010). The data measured at areal level are either the quantity of interest that we would like to model, or they are an input to another spatial model. For example, one may want to regress the unemployment rate on other socio-economic indicators. A popular class of models of this kind is *spatial linear regression* analysis.

In this thesis, our primary focus will be modelling phenomena for which the data are collected as the first two categories: point process realisations, and sensor measurements of a physical system. If relevant, we will also leverage areal data as input to the models. Next, we summarise relevant modelling approaches for both point patterns and sensor measurements.

1.1.2.1 Point Patterns

A point process is a stochastic process for which an observation is a finite or countably infinite set of points in the domain, D, on which the process is defined (Daley & Vere-Jones 2003). We assume that observed data consists of spatial locations of events that occur over a fixed period of time (spatial point pattern). In general, data may consist of a set of events where both the location and the timestamp are recorded for each event (spatio-temporal point pattern). An intuitive definition of a point process that extends to spaces of higher dimensions is using a *counting measure*. For any given measurable subset A of D, the counting measure counts the number of events (points) in the set A. It is denoted by $N(A)$ and it is a non-negative integer- (possibly ∞-) valued quantity. Then, a point process is defined if the joint probability distributions are known for $N(A)$ for all finite disjoint subsets A (Daley & Vere-Jones 2003, ch. 1, ch. 5). In the spatial-only case where $D = \mathbb{R}^2$, subsets A represent surface areas; in the spatio-temporal case, subsets A refer to the space-time-discretised volumes.

By assuming independence of the number of events in the disjoint subsets of D and the assumption that an event may occur at any point in (space-)time, $N(A)$ for any subset A follows Poisson distribution (Cox 1955, Ripley 1977, Diggle 1985, Gelfand 2010). The resulting point process is then referred to as the *Poisson process*, where the intensity function $\lambda(\cdot)$ is the object of interest. The intensity is either constant or varies across the domain – it is assumed that conditional on the intensity function, event locations / times within a given subset A are i.i.d. with density proportional to the intensity function. In real world, there are phenomena where this assumption is invalid – an earthquake event causing aftershocks, a criminal event may spark a new vengeful crime event, or an increase in activity on financial markets can be a catalyst for further events (Hawkes 1971, Ogata 1988, Laub et al. 2015). For these phenomena, the intensity also depends on the past events, and as a result, the process is not Poisson process any more. In this thesis, we do not consider such triggered events.

We may be interested in the intensity itself, or we may want to find out what factors or drivers contribute to the intensity. In the former case, having a model of the intensity can help us predict the intensity in parts of the domain with non-existent or limited observations. In the latter case, determining factors that contribute to the variations in the intensity are of interest for modellers and decision makers. From the macroscopic point of view, they may be interested in the effect of an explanatory variable on the intensity of specific event types, such as crime. On the microscopic level, they may be interested in the interactions between the individual events themselves and how that contributes to the overall intensity.

The choice of the form of the intensity function determines how spatial dependence and spatial heterogeneity are accounted for in the model. In line with interpretability as one of the tenets of this work, the specification for the intensity function will be motivated by the application domain. Throughout Chapter 4, we will work with point patterns of criminal offences. The objective will be to model the intensity of burglary occurrences and quantifying the effect that different socio-economic indicators have on the intensity. Even though this kind of models cannot prevent individual occurrences of crime, we can learn what potential drivers of certain behaviours are and use it for operational insights and policy changes (Felson & Clarke 1998, Taddy 2010, Mohler et al. 2011, Aldor-Noiman et al. 2016, Flaxman et al. 2019, PredPol 2019).

1.1.2.2 Sensor Measurements of Physical Systems

Another approach to modelling spatial dependence and heterogeneity is through partial differential equations (PDEs) which provide a mechanistic and interpretable way of specifying the relationship between different quantities. The partial derivatives encode how a quantity changes as we move through the domain D, which can be specified as spatial-only or spatio-temporal.

Let $\boldsymbol{x} \in D$, $f(\cdot)$ and $\kappa(\cdot)$ be suitable functions (full details provided in Chapter 5) and let $\mathcal{L}$ be an operator involving partial derivatives, parametrised by $\kappa(\cdot)$, and acting on a function $u(\cdot)$. Then a general form for the modelled systems we consider here is

$$\mathcal{L}(\kappa(\cdot))u(\boldsymbol{x}) = f(\boldsymbol{x}), \tag{1.1}$$

with given boundary and/or initial conditions. Solving (1.1) means finding $u(\cdot)$. The operator $\mathcal{L}$, characterised by $\kappa(\cdot)$, encodes how the solution $u(\cdot)$ and $f(\cdot)$ are related.

In a spatial-only context, where $D \subset \mathbb{R}^2$, one could for example model $u(\boldsymbol{x})$ as the solution of the following elliptic PDE:

$$\nabla \cdot (\exp(\kappa(\boldsymbol{x}))\nabla u(\boldsymbol{x})) = f(\boldsymbol{x}), \tag{1.2}$$

where $\nabla = [\partial/\partial x_1, \partial/\partial x_2]^\top$, and the divergence operator $\nabla \cdot ([v_1, v_2]^\top) = \partial v_1/\partial x_1 + \partial v_2/\partial x_2$. Informally, this particular form of $\mathcal{L}$ measures how much the average value of u over the neighbourhood of $\boldsymbol{x}$ deviates from the value of u at $\boldsymbol{x}$. If we assume that $\exp(\kappa(\boldsymbol{x})) = 1$, the operator $\mathcal{L}$ is the Laplacian operator. By choosing $f(\boldsymbol{x})$ and allowing $\exp(\kappa(\boldsymbol{x}))$ to change over the domain D, we can control the spatial dependence of $u(\boldsymbol{x})$ as well as model different regimes in different parts of the domain, thus enabling incorporation of both spatial dependence and spatial heterogeneity. Concrete applications of systems modelled by PDEs of this form include electrostatics (Jackson 1999), steady-state flow of groundwater through an aquifer (Wang & Anderson 2014), and elasticity equations (Bauchau & Craig 2009). More recently, PDEs of this form have proven to be an effective way of scaling models in spatial statistics (Lindgren et al. 2011, Lindgren & Rue 2015).

We focus on problems where we observe a noisy version of the solution $u(\boldsymbol{x})$ at a specified number of locations, and we assume that for a given $\kappa(\boldsymbol{x})$, the dynamics of $u(\boldsymbol{x})$ are controlled by an elliptic PDE of the form given in (1.1). For the observations of u, our objective will be to infer the properties of the unknown parameter κ. This problem is termed the inverse problem (Tarantola 2005, Kaipio & Somersalo 2005, Stuart 2010). Once the unknown parameter is inferred, the model can be used to perform simulations of the studied system under different scenarios, such as how different choices of f affect the solution u. The ability to infer system properties and then use it for simulations under different scenarios has been an indispensable tool in building the so-called *digital twins* (Grieves 2015, Jones et al. 2020).

1.2 Contributions

The thesis is concerned with advancing methodology for modelling spatially correlated phenomena in an interpretable manner and with focus on uncertainty quantification. We make two contributions to this area.

1.2.1 Spatial Poisson Mixture for Modelling Point Patterns

Analysis of the intensity of point patterns is a central task in spatial statistics. Building a model of the intensity as well as understanding what possible factors contribute to its variation are the two main objectives. To make models practically useful, especially on large spatial domains, both spatial dependence and spatial heterogeneity need to be accounted for. This poses several challenges to estimation and simulation of these models. Most notably, computational scalability. We propose a spatial extension to mixtures of generalised linear models to model crime events. The mixture formulation allows for incorporating spatial heterogeneity, while spatial dependence is imposed through the mixture allocation component. The main contributions include:

- We discretise the domain of interest into a grid of cells and develop a Bayesian model for the counts of points in cells using a mixture of Poisson regressions where mixture membership allocation is modelled in a probabilistic manner.
- We leverage findings from criminology literature to shortlist socio-economic variables that are relevant to criminal activity and estimate their effect on the intensity of the observed point pattern.
- Compared to the go-to model of spatial point patterns (log-Gaussian Cox process), our proposed model achieves superior predictive performance, is more computationally scalable, and the mixture components can be interpreted to provide criminological insights.

The core of this contribution is presented in Chapter 4, with necessary technical background discussed in Chapters 2 and 3. This work has been published as Povala et al. (2020) and the code that replicates the analysis is available for public use.

1.2.2 Assimilation of Sensor Measurements into PDEs Using Variational Bayes

Partial differential equations provide a mechanistic way for incorporating spatial dependence and spatial heterogeneity into models. The inverse problems involving PDEs are of great importance in science and engineering. Although such problems are generally ill-posed, regularisation is used to ameliorate this problem. One of the viewpoints in which to view regularised solutions is the Bayesian formulation, where a prior probability measure is placed on the quantity of interest. The resulting posterior probability measure is usually analytically intractable. The Markov Chain Monte Carlo (MCMC) method has been the go-to method for sampling from those posterior measures. MCMC is computationally infeasible for large-scale problems that arise in engineering practice. Lately, variational Bayes (VB) has been recognised as a more computationally tractable method for Bayesian inference, approximating a Bayesian posterior distribution with a simpler trial distribution by solving an optimisation problem. We argue, through an extensive empirical assessment, that Variational Bayes methods, when appropriately parameterised, are a preferable alternative to MCMC methods. The main contributions include:

- We propose a variational Bayes parametrisation that leverages sparsity arising from the discretisation of PDE models using finite elements.
- We assess the expected error in the mean, as well as the ability to quantify the uncertainty of the estimate.
- Our results on examples of elliptic PDEs show that variational Bayes methods provide a good estimate for the mean and variance of the posterior distribution in a time that is an order of magnitude faster than MCMC methods.

The main part of this contribution is in Chapter 6. Prerequisite technical material can be found in Chapters 2 and 5. This work has been published as Povala et al. (2022).

Chapter 2

Statistical Inference

This chapter provides an overview of the statistical inference methods we use in subsequent chapters. After general introduction of statistical inference, we shift focus to Bayesian inference methods and give detailed explanations of the schemes we employ in the thesis.

2.1 Introduction

This introductory section closely follows Young & Smith (2005). Statistical inference provides a framework where observational or experimental data are modelled as observed values of random variables to allow for inductive conclusions to be drawn about the mechanism giving rise to the data.

Given a vector of n possibly vector-valued observations $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, we regard $\boldsymbol{y}$ as the observed value of a random variable $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ with (unknown) probability measure, often specified by a probability density, or probability mass function, $f(\boldsymbol{y})$. In *parametric statistical inference*, $f(\boldsymbol{y})$ is of known analytic form, but involves a finite number of real unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. We specify the region $\Theta \subseteq \mathbb{R}^p$ of possible values of $\boldsymbol{\theta}$, the *parameter space*. For a parametric model, we write $f(\boldsymbol{y}; \boldsymbol{\theta})$. Alternatively, the observed data, $\boldsymbol{y}$, could be modelled non-parametrically.

The objective of statistical inference is to assess some aspect of $\boldsymbol{\theta}$, having observed $\boldsymbol{y}$, by considering a suitable family of distributions, $\mathcal{F} = \{f(\boldsymbol{y}; \boldsymbol{\theta}) \colon \boldsymbol{\theta} \in \Theta\}$. Common types of inference include: *point estimation*, *hypothesis testing*, *confidence set estimation*, *prediction of a yet unobserved random variable*, and *examination of model specification* by $\mathcal{F}, \Theta$.

There are two broad approaches to statistical inference: *Bayesian* and *frequentist* (Cox 2006). The differentiating property is the interpretation of probability, whereas common to both approaches is the concept of *likelihood*. Likelihood measures the probability that different values of the parameter $\boldsymbol{\theta}$ assign to the actual observed data $\boldsymbol{y}$. After observing $\boldsymbol{y}$, the likelihood function is given by

$$L(\boldsymbol{\theta}) \equiv L(\boldsymbol{\theta}; \boldsymbol{y}) = f(\boldsymbol{y}; \boldsymbol{\theta}), \tag{2.1}$$

where f is a probability density function, or probability mass function if $\mathbf{Y}$ is a discrete random variable.

In the frequentist approach, no further probabilistic assumptions are made. The parameter $\boldsymbol{\theta}$ is treated as an unknown constant, and statistical inferences about $\boldsymbol{\theta}$ must be based on probabilities with direct experimental interpretation. Central to this approach is the *repeated sampling principle*: the inference drawn about $\boldsymbol{\theta}$ from observing $\boldsymbol{y}$ should be based on an analysis of how the conclusions change with variations in the data samples which would be obtained through hypothetical repetitions, under the same conditions of the experiment which generated data $\boldsymbol{y}$ (Young & Smith 2005). Through the repeated sampling principle, we derive long-run behaviour to allow for sound conclusions to be made from the particular instance of data under analysis, $\boldsymbol{y}$. The challenge here is ensuring the relevance of the derived long-run behaviour to the observed particular instance (Cox 2006). This approach was spearheaded by Ronald A. Fisher who used the likelihood to develop the maximum likelihood estimate (MLE) methodology. He provided a description of the optimum that is achievable in a given estimation problem. He derived the asymptotic standard error of the MLE estimate and has shown that no other consistent and sufficiently regular estimator can do better (Fisher 1922, Efron 1998, Young & Smith 2005). A further fundamental element of Fisher's viewpoint is that inference about $\boldsymbol{\theta}$, to be as relevant as possible to the data $\boldsymbol{y}$, must be carried out conditional on everything that is known and uninformative about $\boldsymbol{\theta}$ (Fisher 1934, Young & Smith 2005). For example, conditioning on an ancillary statistic for $\boldsymbol{\theta}$.

The appeal of the frequentist methodology lies in no a priori assumptions about the parameter $\boldsymbol{\theta}$. Any inferences that we draw about $\boldsymbol{\theta}$ are based on the observed data $\boldsymbol{y}$. This is in contrast with the Bayesian paradigm, where we treat the parameter $\boldsymbol{\theta}$ as a random variable itself and we place a prior probability distribution over $\boldsymbol{\theta}$ *before* any data analysis. For reasons which will become clear later on, we will adopt the Bayesian approach to inference. We give a detailed review of the main ideas in the next sections.

2.2 Bayesian Methods

The Bayesian paradigm of statistical inference goes back to the ideas developed by Reverend Thomas Bayes and P. S. Laplace (Bayes 1763, Laplace 1812). The fundamental concept in this approach is that the unknown parameter $\boldsymbol{\theta}$ should itself be treated as a random variable, as opposed to a fixed value. In the frequentist approach, the model expresses the natural randomness or uncertainty by treating data $\boldsymbol{y}$ as a sample from random variables $\mathbf{Y}$, whose law is parametrised by fixed parameter $\boldsymbol{\theta}$. This type of uncertainty is referred to as *aleatory* uncertainty. Bayesian paradigm, in addition, advocates for using random variables for uncertainty due to our lack of knowledge about a past or present fact or number, which could be in principle known, but we do not have access to it (van der Bles et al. 2019). This type of uncertainty is referred to as *epistemic*. We specify the epistemic uncertainty about the unknown parameter $\boldsymbol{\theta}$ before the data analysis through a prior probability distribution which reflects our current knowledge (or lack thereof) and it may be subjective. The subjectivity of the prior is the main source of disagreement between the Bayesian approach and the frequentist one (Young & Smith 2005). The likelihood of the data is conditional on the random variable $\boldsymbol{\theta}$, and we denote it as $p(\boldsymbol{y} \mod \boldsymbol{\theta}$.

To simplify notation in the subsequent sections, we often do not distinguish between $\boldsymbol{\theta}$ as a random variable, and $\boldsymbol{\theta}$ as a particular realisation of that random variable. This will be clear from the context.

Bayesian inference is the formalisation of how the prior probability distribution over an unknown quantity changes into the posterior probability density in the light of evidence from data $\boldsymbol{y}$. This update of the probability distribution is expressed through Bayes' formula:

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{2.2}$$

where $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the conditional density of $\mathbf{Y}$ given $\boldsymbol{\theta}$ (often referred to as the likelihood function), and $p(\boldsymbol{y}) = \int_\Theta p(\boldsymbol{\theta}')p(\boldsymbol{y} \mid \boldsymbol{\theta}')\mathrm{d}\boldsymbol{\theta}'$ is the marginal likelihood. The specification of $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is often a straightforward task and agrees with the frequentist approach. It expresses any discrepancy between the model and the data. Specification of $p(\boldsymbol{\theta})$ is more controversial and different approaches have been proposed in the literature, as we discuss below. Given that we employ the Bayesian framework in the subsequent chapters of this thesis, we give a detailed exposition of the Bayesian methods and the associated issues.

2.2.1 Prior Distribution

Firstly, we give a summary of the main approaches for specifying the prior distribution $p(\boldsymbol{\theta})$ as discussed in Young & Smith (2005):

(a) Physical reasoning priors – advocated for and used by Bayes, but too restrictive for many practical situations.

(b) "Non-informative priors" – the view adopted by Jeffreys and Laplace (Laplace 1812, Jeffreys 1998) and a widely used method in practice. Following this procedure may lead into improper priors, which cannot be normalised to form a proper density. However, the resulting posterior is often a proper probability density. It is important to note that all prior are informative.

(c) Subjective priors – an approach promoted by B. de Finetti, L.J. Savage, and D.V. Lindley. According to their theory of subjective probability, each individual behaves in such a way as to maximise his/her expected utility according to his/her own judgement of probabilities of different outcomes. It is used in applications where subjective prior information, *e.g.*, from an expert, may be important.

(d) Convenient priors – often chosen to simplify the calculations. For example, choosing priors that form a conjugate pair with $f(\boldsymbol{y};\boldsymbol{\theta})$. In Example 2.1 below, we give an illustrative example of the prior that forms a conjugate pair with the binomial density. This choice makes $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ being available in closed form.

Example 2.1 (Analytic posterior distribution)**.** *Let y be the number of times a coin lands heads in a series of n independent coin tosses. We model the number of head outcomes by the binomial probability distribution, with the probability of an individual toss landing heads given by θ. We have $Y \sim Bin(n, \theta)$ with known n and unknown θ. Suppose that the prior density is $Beta(a, b)$ on $(0, 1)$,*

$$p(\theta) = \frac{\Gamma(a+b)\theta^{a-1}(1-\theta)^{b-1}}{\Gamma(a)\Gamma(b)}, \quad 0 < \theta < 1,$$

where $a > 0$, $b > 0$, and $\Gamma(\cdot)$ is the gamma function, $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\mathrm{d}x$. The density for y is given by

$$f(y; \theta) = \binom{n}{x}\theta^y(1-\theta)^{n-y}. \tag{2.3}$$

Applying (2.2), we obtain

$$p(\theta \mid y) \propto \theta^{a+x-1}(1-\theta)^{n-x+b-1},$$

which is proportional to the Beta density with parameters $a + y$ and $n - x + b$, which gives the posterior distribution.

The analytic tractability of the posterior in the example above is enabled by the conjugacy of the Beta density with the binomial density. Although there are several useful examples of conjugate pairs (see Gelman et al. (2013) for more examples), most of the time the posterior distribution is not analytically available, and we must resort to approximation methods which we discuss in Section 2.2.2.

2.2.1.1 Hierarchical Models

We can use the specification of the prior distribution to incorporate dependence structure into the model by making the components of $\boldsymbol{\theta}$ related to one another, as opposed to imposing that structure in the specification of $f(\boldsymbol{y}; \boldsymbol{\theta})$. We add the dependence by assuming that different components of $\boldsymbol{\theta}$ are sampled from a common probability distribution with its own parameters of which we place another prior distribution, hence 'hierarchical'. Apart from inducing a prior dependence structure in the components of $\boldsymbol{\theta}$, hierarchical specification allows for deferring the specification of prior parameters to another stage: the prior parameter values are themselves given a prior distribution, often an uninformative one.

2.2.2 Approximation Techniques

As mentioned above, most of the choices of $f(\boldsymbol{y};\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ will not lead to a closed-form expression for the posterior distribution. As a consequence, approximation methods need to be used when making inferences about the posterior. We categorise approximation techniques into two categories: 1) *simulation-based* and 2) those comibining *direct numerical integration* with *optimisation*. In the following sections, we will give details of only methods we use in subsequent chapters: Markov Chain Monte Carlo method as an example of a simulation-based approximation scheme and variational Bayes, also known as variational inference, as an example of an optimisation-based scheme.

2.2.3 Markov Chain Monte Carlo

Markov Chain Monte Carlo method overcomes the analytical intractability of $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ by drawing samples from approximate distributions and then correcting those draws to better approximate the posterior. The output is a set of sequential draws (not i.i.d.), $\{\boldsymbol{\theta}_t\}_{t=1}^{T}$, such that the sequence forms a discrete-time Markov Chain whose stationary distribution is $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. For the purposes of this section, we denote $\pi(\boldsymbol{\theta}) \equiv p(\boldsymbol{\theta} \mid \boldsymbol{y})$, and $\pi_u(\boldsymbol{\theta})$ its unnormalised version, *i.e.*, $\pi(\boldsymbol{\theta}) \propto \pi_u(\boldsymbol{\theta})$. The exposition below largely follows Roberts & Rosenthal (2004). A discrete-time Markov Chain is a stochastic process $\{\boldsymbol{\theta}_t \in \Theta \colon t \in I\}$, with $I \subset \mathbb{N}$ and $\Theta \subset \mathbb{R}^d$ that satisfies the Markov property,

$$\mathbb{P}(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_{n-1}, \dots, \boldsymbol{\theta}_1) = \mathbb{P}(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_{n-1}),$$

where $A \in \sigma(\Theta)$ and $\sigma(\Theta)$ is the Borel σ-algebra on Θ. In other words, the only relevant distribution for the next state is the one at the current state. We emphasise again that it is clear from the context whether $\boldsymbol{\theta}_n$ refers to a random variable or its particular instantiation.

To ensure that $\pi(\cdot)$ is the stationary distribution of the chain on Θ, the evolution of the chain must satisfy transition probabilities $P(\boldsymbol{\theta}_i, \mathrm{d}\boldsymbol{\theta}_j)$ for $\boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \Theta$, such that

$$\int_{\boldsymbol{\theta}_i \in \Theta} \pi(\boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i, \mathrm{d}\boldsymbol{\theta}_j) = \pi(\mathrm{d}\boldsymbol{\theta}_j),$$

i.e., the transition kernel $P(\boldsymbol{\theta}_i, \mathrm{d}\boldsymbol{\theta}_j)$ leaves $\pi(\cdot)$ invariant. A sufficient, but not necessary, condition for this is the *reversibility* condition.

Definition 2.2 (Reversibility)**.** A Markov chain on a state space Θ is *reversible* with respect to a probability distribution $\pi(\cdot)$ on Θ, if

$$\pi(\mathrm{d}\boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i, \mathrm{d}\boldsymbol{\theta}_j) = \pi(\mathrm{d}\boldsymbol{\theta}_j) P(\boldsymbol{\theta}_j, \mathrm{d}\boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i, \boldsymbol{\theta}_j \in \Theta. \tag{2.4}$$

To ascertain that a constructed Markov chain indeed converges to the required posterior $\pi(\cdot)$, we define n-step transition law of the Markov chain,

$$P^n(\boldsymbol{\theta}_0, A) = \mathbb{P}(\boldsymbol{\theta}_n \in A \mid \boldsymbol{\theta}_0) \tag{2.5}$$

and the total variation distance between two probability measures $\nu_1(\cdot)$ and $\nu(\cdot)$:

$$\|\nu_1(\cdot) - \nu_2(\cdot)\|_{\mathrm{TV}} = \sup_A |\nu_1(A) - \nu_2(A)|. \tag{2.6}$$

Using these definitions, we formulate the asymptotic convergence in total variation-distance as

$$\lim_{n \to \infty} \|P^n(\boldsymbol{\theta}, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} = 0, \quad \boldsymbol{\theta} \in \Theta. \tag{2.7}$$

For this to hold, the following two conditions are sufficient:

1. *φ-irreducibility*: a chain is φ-irreducible if there exists a non-zero σ-finite measure φ on Θ such that for all $A \subset \Theta$ with $\varphi(A) > 0$, and for all $\boldsymbol{\theta} \in \Theta$, there exists a positive integer $n = n(\boldsymbol{\theta}, A)$ such that $P^n(\boldsymbol{\theta}, A) > 0$.

2. *Aperiodicity*: A Markov chain with stationary distribution $\pi(\cdot)$ is aperiodic if there do not exist $d \geq 2$ and disjoint subsets $\Theta_1, \Theta_2, \ldots, \Theta_d \subset \Theta$ with $P(\boldsymbol{\theta}, \Theta_{i+1}) = 1$ for all $\boldsymbol{\theta} \in \Theta_i (1 \leq i \leq d-1)$, and $P(\boldsymbol{\theta}, \Theta_1) = 1$ for all $\boldsymbol{\theta} \in \Theta_d$, such that $\pi(\Theta_1) > 0$ (and hence $\pi(\Theta_i) > 0$ for all i).

These two properties lead to the following theorem.

Theorem 2.3. *If a Markov chain on a state space with countably generated σ-algebra is φ-irreducible and aperiodic, and has a stationary distribution $\pi(\cdot)$, then for π-a.e., and any $\boldsymbol{\theta} \in \Theta$,*

$$\lim_{n\to\infty} \|P^n(\boldsymbol{\theta}, \cdot) - \pi(\cdot)\|_{TV} = 0 \tag{2.8}$$

In particular, $\lim_{n\to\infty} P^n(\boldsymbol{\theta}, A) = \pi(A)$ *for all measurable* $A \subset \Theta$.

To assess the convergence rate of (2.8), *geometric ergodicity*, which is defined as follows, gives the convergence rates of the Markov chain to the true distribution $\pi(\cdot)$.

Definition 2.4 (Geometric ergodicity)**.**

$$\|P^n(\boldsymbol{\theta}, \cdot) - \pi(\cdot)\|_{\mathrm{TV}} \leq M(\boldsymbol{\theta})\rho^n, \quad n = 1, 2, 3, \ldots \tag{2.9}$$

for some $\rho < 1$, where $M(\boldsymbol{\theta}) < \infty$ for π-a.e., and $\boldsymbol{\theta} \in \Theta$.

Note that this rate depends on the initial position x, where the stronger version, referred to as *uniform ergodicity*, is independent of the initial state, but its assumptions are rarely found to hold in practice (Roberts & Rosenthal 2004).

2.2.3.1 Metropolis-Hastings Algorithm

One of the most popular algorithms for constructing reversible Markov Chains which have the required stationary probability distribution is the Metropolis-Hastings algorithm (Metropolis et al. 1953, Hastings 1970). The algorithm defines the transition probabilities $P(\boldsymbol{\theta}_i, \mathrm{d}\boldsymbol{\theta}_j)$ using a proposal distribution $Q(\boldsymbol{\theta}_i, \cdot)$ with its own (possibly unnormalised) density, *i.e.*, $Q(\boldsymbol{\theta}_i, \mathrm{d}\boldsymbol{\theta}_j) \propto q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta}_j$ and the acceptance ratio $\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ which is defined as

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = \min\left[1, \frac{\pi(\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i)}{\pi(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)}\right]. \tag{2.10}$$

Given the current state $\boldsymbol{\theta}_n$, we generate a proposal $\boldsymbol{\theta}^*_{n+1}$ from $Q(\boldsymbol{\theta}_n, \cdot)$. We flip a coin whose probability of heads is $\alpha(\boldsymbol{\theta}_n, \boldsymbol{\theta}^*_{n+1})$. If the coin lands heads, we accept the proposal and set $\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}^*_{n+1}$. The transition kernel is then given as

$$P(\boldsymbol{\theta}_i, \mathrm{d}\boldsymbol{\theta}_j) = q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta}_j. \tag{2.11}$$

To prove correctness, we need to show reversibility

$$\pi(\mathrm{d}\boldsymbol{\theta}_i)P(\boldsymbol{\theta}_i,\mathrm{d}\boldsymbol{\theta}_j)=\pi(\mathrm{d}\boldsymbol{\theta}_j)P(\boldsymbol{\theta}_j,\mathrm{d}\boldsymbol{\theta}_i). \tag{2.12}$$

For $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_j$ (if $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ then the equation is trivial), we have

$$\begin{aligned}
\pi(\mathrm{d}\boldsymbol{\theta}_i)P(\boldsymbol{\theta}_i,\mathrm{d}\boldsymbol{\theta}_j) &= [c^{-1}\pi_u(\boldsymbol{\theta}_i)\mathrm{d}\boldsymbol{\theta}_i][q(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)\alpha(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta}_j] && (2.13)\\
&= c^{-1}\pi_u(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)\min\left[1,\frac{\pi_u(\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_j,\boldsymbol{\theta}_i)}{\pi_u(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)}\mathrm{d}\boldsymbol{\theta}_i\mathrm{d}\boldsymbol{\theta}_j\right] && (2.14)\\
&= c^{-1}\min\left[\pi_u(\boldsymbol{\theta}_i)q(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j),\pi_u(\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_j,\boldsymbol{\theta}_i)\right]\mathrm{d}\boldsymbol{\theta}_i\mathrm{d}\boldsymbol{\theta}_j, && (2.15)
\end{aligned}$$

which is symmetric in $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, showing that (2.12) holds.

We conclude the discussion with the following remarks.

- Only the unnormalised posterior density, $\pi_u(\cdot)$, is necessary as the normalising constants cancel out in the acceptance ratio α.
- There is a trade-off between the acceptance rate and the distance between successive samples. Larger moves get rejected more frequently, but smaller moves lead to higher autocorrelation of the Markov Chain.
- A number of Metropolis-Hastings steps 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 chain is ergodic with respect to the target distribution (Roberts & Rosenthal 2004)

2.2.3.2 Gibbs Sampler

For some problems, especially in higher dimensions, constructing an MH updating scheme which *jointly* updates all components of $\boldsymbol{\theta}$ is challenging as the proportion of accepted samples would be low. It may be possible to split $\boldsymbol{\theta}$ into *non-overlapping* groups of random variables, each of which is updated separately. This is the main idea behind the Gibbs sampler (Geman & Geman 1984).

Gibbs sampler can be viewed as a special case of the MH transition kernel, in which only a subset I of $\boldsymbol{\theta}$ is updated at a time, whilst the remaining variables, $\boldsymbol{\theta}_{i\notin I}$, are held fixed:

$$q(\boldsymbol{\theta},\boldsymbol{\theta}') = q(\boldsymbol{\theta}_{i\in I},\boldsymbol{\theta}'_{i\in I})\delta_{\boldsymbol{\theta}_{i\notin I}}(\boldsymbol{\theta}'_{i\notin I}), \tag{2.16}$$

where $\delta_{\boldsymbol{\theta}_0}(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta}-\boldsymbol{\theta}_0)$ is the Dirac delta function.

If one is able to draw from full conditionals, *i.e.*, $q(\boldsymbol{\theta}_{i\in I},\boldsymbol{\theta}'_{i\in I}) = \pi(\boldsymbol{\theta}'_{i\in I} \mid \boldsymbol{\theta}_{i\notin I})$, the acceptance ratio in (2.10) becomes

$$\begin{aligned}\alpha(\boldsymbol{\theta},\boldsymbol{\theta}') &= \frac{\pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}',\boldsymbol{\theta})}{\pi(\boldsymbol{\theta})q(\boldsymbol{\theta},\boldsymbol{\theta}')}\\ &= \frac{\pi(\boldsymbol{\theta}')}{\pi(\boldsymbol{\theta})}\frac{q(\boldsymbol{\theta}'_{i\in I},\boldsymbol{\theta}_{i\in I})\delta_{\boldsymbol{\theta}_{i\notin I}}(\boldsymbol{\theta}'_{i\notin I})}{q(\boldsymbol{\theta}_{i\in I},\boldsymbol{\theta}'_{i\in I})\delta_{\boldsymbol{\theta}'_{i\notin I}}(\boldsymbol{\theta}_{i\notin I})}\\ &= \frac{\pi(\boldsymbol{\theta}'_{i\in I} \mid \boldsymbol{\theta}'_{i\notin I})\pi(\boldsymbol{\theta}'_{i\notin I})}{\pi(\boldsymbol{\theta}_{i\in I} \mid \boldsymbol{\theta}_{i\notin I})\pi(\boldsymbol{\theta}_{i\notin I})}\frac{q(\boldsymbol{\theta}'_{i\in I},\boldsymbol{\theta}_{i\in I})}{q(\boldsymbol{\theta}_{i\in I},\boldsymbol{\theta}'_{i\in I})}\\ &= \frac{\pi(\boldsymbol{\theta}'_{i\in I} \mid \boldsymbol{\theta}'_{i\notin I})\pi(\boldsymbol{\theta}'_{i\notin I})}{\pi(\boldsymbol{\theta}_{i\in I} \mid \boldsymbol{\theta}_{i\notin I})\pi(\boldsymbol{\theta}_{i\notin I})}\frac{\pi(\boldsymbol{\theta}_{i\in I} \mid \boldsymbol{\theta}'_{i\notin I})}{\pi(\boldsymbol{\theta}'_{i\in I} \mid \boldsymbol{\theta}_{i\notin I})} = 1,\end{aligned} \tag{2.17}$$

implying that updates to variables with indices in I will always get accepted.

We conclude the discussion of Gibbs sampling with the following remarks.

- Although the proposals in Gibbs sampling are always accepted, the exploration of the posterior distribution $\pi(\boldsymbol{\theta})$ may be slow due to strong correlations between the groups of variables. Techniques such as rescaling or transformation of $\boldsymbol{\theta}$ have been shown to alleviate issues related to strong correlations (Gelman et al. 2013).

- Selecting the group of non-overlapping variables (each of which is a subset of $\boldsymbol{\theta}$) is done either deterministically in a sequential manner (also known as *deterministic-scan Gibbs sampler*) or one group is chosen at random for each iteration (often referred to as *random-scan Gibbs sampler*) (Roberts & Rosenthal 2004).

- For scenarios where drawing directly from full conditionals is not available for a group of variables, a Metropolis-Hastings update step can be used to propose samples which are accepted as per the acceptance ratio in (2.10). This is known as *Metropolis-within-Gibbs scheme.*

2.2.3.3 Hamiltonian Monte Carlo

One of the most effective ways of exploring high-dimensional posterior distribution is to leverage gradient information of the posterior into the proposal mechanism of an MCMC scheme. Hamiltonian Monte Carlo (HMC), introduced into the statistics community by Duane et al. (1987), is a MH variant where the proposals are made by computing a trajectory according to Hamiltonian dynamics, implemented with the leapfrog method (Neal 2011). A new state proposed in this manner can be distant from the current state of $\boldsymbol{\theta}$ but nevertheless have a high probability of acceptance.

The core of the method lies in defining a Hamiltonian function H in terms of the probability distribution we wish to sample from, $\pi(\boldsymbol{\theta})$, and artificially introduced auxiliary 'momentum' variables, $\boldsymbol{\phi}$, which have density $\pi(\boldsymbol{\phi})$, and are often defined as independent Gaussians. One momentum variable is added for every component of $\boldsymbol{\theta}$. The MCMC updates are performed on the joint distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, and the extra variables are subsequently discarded. The joint distribution is given by

$$\pi(\boldsymbol{\theta}, \boldsymbol{\phi}) \propto \exp\big(-H(\boldsymbol{\theta}, \boldsymbol{\phi})\big), \tag{2.18}$$

where

$$H(\boldsymbol{\theta}, \boldsymbol{\phi}) = U(\boldsymbol{\theta}) + K(\boldsymbol{\phi}), \tag{2.19}$$

where $U(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta})$ is called the *potential energy*, and $K(\boldsymbol{\phi}) = -\log \pi(\boldsymbol{\phi})$ is called the *kinetic energy*, which for the independent Gaussians reads as

$$K(\boldsymbol{\phi}) \overset{\text{const}}{=} \boldsymbol{\phi}^\top \boldsymbol{M}^{-1} \boldsymbol{\phi}/2, \tag{2.20}$$

with $\boldsymbol{M}$, referred to as 'mass matrix', being diagonal.

The Hamiltonian dynamics describe how $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ change over time, t:

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \frac{\partial H}{\partial \boldsymbol{\phi}} \tag{2.21}$$

$$\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = -\frac{\partial H}{\partial \boldsymbol{\theta}}, \tag{2.22}$$

which for the case of $\pi(\boldsymbol{\phi})$ being Gaussian density reads as

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \boldsymbol{M}^{-1}\boldsymbol{\phi} \tag{2.23}$$

$$\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = \frac{\partial \log \pi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \tag{2.24}$$

For any time interval of size s, these equations define a mapping T_s, from the state at any time t to the state at time $t+s$ (note that H, and hence T_s are assumed to not depend on t). Several properties of Hamiltonian dynamics are crucial to its use as an update mechanism for MCMC which we discuss next.

Reversibility: The Hamiltonian dynamics can be reversed as the mapping T_s from the state at time t, $(\boldsymbol{\theta}(t), \boldsymbol{\phi}(t))$ to the state at time $t+s$, $(\boldsymbol{\theta}(t+s), \boldsymbol{\phi}(t+s))$, is one-to-one, and hence has an inverse T_{-s}. If $K(p) = K(-p)$, which is the case for the independent Gaussians case, the inverse mapping can be obtained by negating $\boldsymbol{\phi}$, applying T_s, and then negating $\boldsymbol{\phi}$ again (Neal 2011). The reversibility of Hamiltonian dynamics implies the reversibility of the corresponding Markov chain, which is used to show that MCMC updates based on Hamiltonian dynamics leave the target distribution invariant.

Conservation of Hamiltonian: The update dynamics conserve the Hamiltonian as

$$\frac{\mathrm{d}H}{\mathrm{d}t} = \frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}\frac{\partial H}{\partial \boldsymbol{\theta}} + \frac{\partial H}{\partial \boldsymbol{\phi}}\frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} = \frac{\partial H}{\partial \boldsymbol{\phi}}\frac{\partial H}{\partial \boldsymbol{\theta}} - \frac{\partial H}{\partial \boldsymbol{\phi}}\frac{\partial H}{\partial \boldsymbol{\theta}} = 0. \tag{2.25}$$

If the Hamiltonian, H, is kept invariant, the acceptance probability for updates based on Hamiltonian dynamics is one. In practice, H can only be approximately invariant. We discuss this in the discretisation section below.

Volume preservation: The Hamiltonian dynamics preserve the volume in the $(\boldsymbol{\theta}, \boldsymbol{\phi})$ space. As a consequence of Liouville's theorem, applying the mapping T_s to the points in some region R of the $(\boldsymbol{\theta}, \boldsymbol{\phi})$ space with boundary S, the image of R under T_s will have the same volume:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_R \mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\phi} = \int_S \left(\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}, \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t}\right) \cdot \hat{\boldsymbol{n}}\mathrm{d}S = \int_R \nabla \cdot \left(\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}, \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t}\right)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\phi} = 0. \tag{2.26}$$

We used the fact that the flow in the $(\boldsymbol{\theta}, \boldsymbol{\phi})$ space is divergence free:

$$\nabla \cdot \left(\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t}, \frac{\mathrm{d}\boldsymbol{\phi}}{\mathrm{d}t} \right) = \frac{\partial^2 H}{\partial \boldsymbol{\theta} \partial \boldsymbol{\phi}} - \frac{\partial^2 H}{\partial \boldsymbol{\phi} \partial \boldsymbol{\theta}} = 0. \tag{2.27}$$

As a consequence of volume preservation, which means that the determinant of the Jacobian matrix of the mapping T. has absolute value one, we do not need to account for any change in volume in the acceptance probability for the updates.

Discretising Hamiltonian dynamics: In practice, Hamiltonian dynamics are discretised using a leapfrog integrator scheme for a length of time ϵL, where ϵ is the step size and L is the number of leapfrog step. A single leapfrog step is given as:

$$\boldsymbol{\phi}(t + \epsilon/2) = \boldsymbol{\phi}(t) - (\epsilon/2) \frac{\partial}{\partial \boldsymbol{\theta}} U(\boldsymbol{\theta}(t)) \tag{2.28}$$

$$\boldsymbol{\theta}(t + \epsilon) = \boldsymbol{\theta}(t) + \epsilon \frac{\partial}{\partial \boldsymbol{\phi}} K(\boldsymbol{\phi}(t + \epsilon/2)) \tag{2.29}$$

$$\boldsymbol{\phi}(t + \epsilon) = \boldsymbol{\phi}(t + \epsilon/2) - (\epsilon/2) \frac{\partial}{\partial \boldsymbol{\theta}} U(\boldsymbol{\theta}(t + \epsilon)) \tag{2.30}$$

The leapfrog integrator exactly preserves the volume as the equations above correspond to shear transformations with the determinant of the Jacobian equal to one, and it is also reversible. After running these updates for L steps and changing the sign of the momentum variables, $\boldsymbol{\phi}$, the proposal, $(\boldsymbol{\theta}', \boldsymbol{\phi}')$, is accepted with the Metropolis-Hastings acceptance probability given by

$$\alpha((\boldsymbol{\theta}, \boldsymbol{\phi}), (\boldsymbol{\theta}', \boldsymbol{\phi}')) = \min \left\{ 1, \exp\left(H(\boldsymbol{\theta}, \boldsymbol{\phi}) - H(\boldsymbol{\theta}', \boldsymbol{\phi}') \right) \right\}. \tag{2.31}$$

The full steps of the algorithm are given in Algorithm 1.

The performance of the algorithm can be tuned in three ways: (i) choice of the momentum distribution $p(\boldsymbol{\phi})$, which in the version with independent Gaussians requires specifying the mass matrix, $\boldsymbol{M}$, (ii) adjusting the scaling factor of the leapfrog step, ϵ, and (iii) the number of leapfrog steps, L. Gelman et al. (2013) suggest setting ϵ and L so that $\epsilon L = 1$. They suggest tuning these so that the acceptance rate is about 65%. As for the mass matrix, the authors suggest that it should approximately scale with the inverse covariance matrix of the posterior distribution, $(\mathrm{Cov}(\boldsymbol{\theta} \mid \boldsymbol{y}))^{-1}$. This can be achieved by a pre-run from which the empirical covariance matrix can be computed.

Algorithm 1: HAMILTONIAN MONTE CARLO as presented in Gelman et al. (2013)

```
Input: $\pi_u(\boldsymbol{\theta})$: unnormalised target density and its gradients , $\pi(\boldsymbol{\phi})$: momentum
       density, $L$: leapfrog steps, $\epsilon$: scaling factor
Output: A list of samples from $\pi(\boldsymbol{\theta})$.
for $t \leftarrow 1, 2, \dots$ do
    Sample $\boldsymbol{\phi}$ from $p(\boldsymbol{\phi})$
    $\boldsymbol{\theta}_* \leftarrow \boldsymbol{\theta}_{t-1}$
    for $i \leftarrow 1$ to $L$ do
        $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{1}{2}\epsilon\frac{\partial \log \pi_u(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}}$
        $\boldsymbol{\theta}_* \leftarrow \boldsymbol{\theta}_* + \epsilon\frac{\partial}{\partial \boldsymbol{\phi}} K(\boldsymbol{\phi})$
        $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{1}{2}\epsilon\frac{\partial \log \pi_u(\boldsymbol{\theta}_*)}{\partial \boldsymbol{\theta}}$
    $r \leftarrow \frac{\pi_u(\boldsymbol{\theta}_*)\pi(\boldsymbol{\phi}_*)}{\pi_u(\boldsymbol{\theta}_{t-1})\pi(\boldsymbol{\phi}_{t-1})}$
    $\boldsymbol{\theta}_t \leftarrow \begin{cases} \boldsymbol{\theta}_* & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}_{t-1} & \text{otherwise} \end{cases}$
return $[\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \dots]$
```

Recently, the no-U-turn sampler has been proposed as a way of determining L adaptively (Hoffman & Gelman 2014). Further efficiency gains can be achieved by incorporating the geometry of the posterior into the proposal (Girolami & Calderhead 2011).

2.2.3.4 Example

Example 2.5 (highly-correlated bivariate posterior)**.** *We show simulations for obtaining samples from a target probability distribution that exhibits high correlations. We compare three methods: Metropolis-Hastings MCMC, Gibbs sampler, and Hamiltonian Monte Carlo. The distribution we wish to sample from is given by*

$$\begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right) \tag{2.32}$$

where we set $\rho = 0.9$. We show the sampling process for all three methods in Figure 2.1. It is clear that HMC can explore the posterior most effectively, while the samples made using MH and Gibbs show high autocorrelation. The MH proposals are based on a scaled identity matrix, where the scaling parameter is tuned to achieve acceptance rate of 0.23.

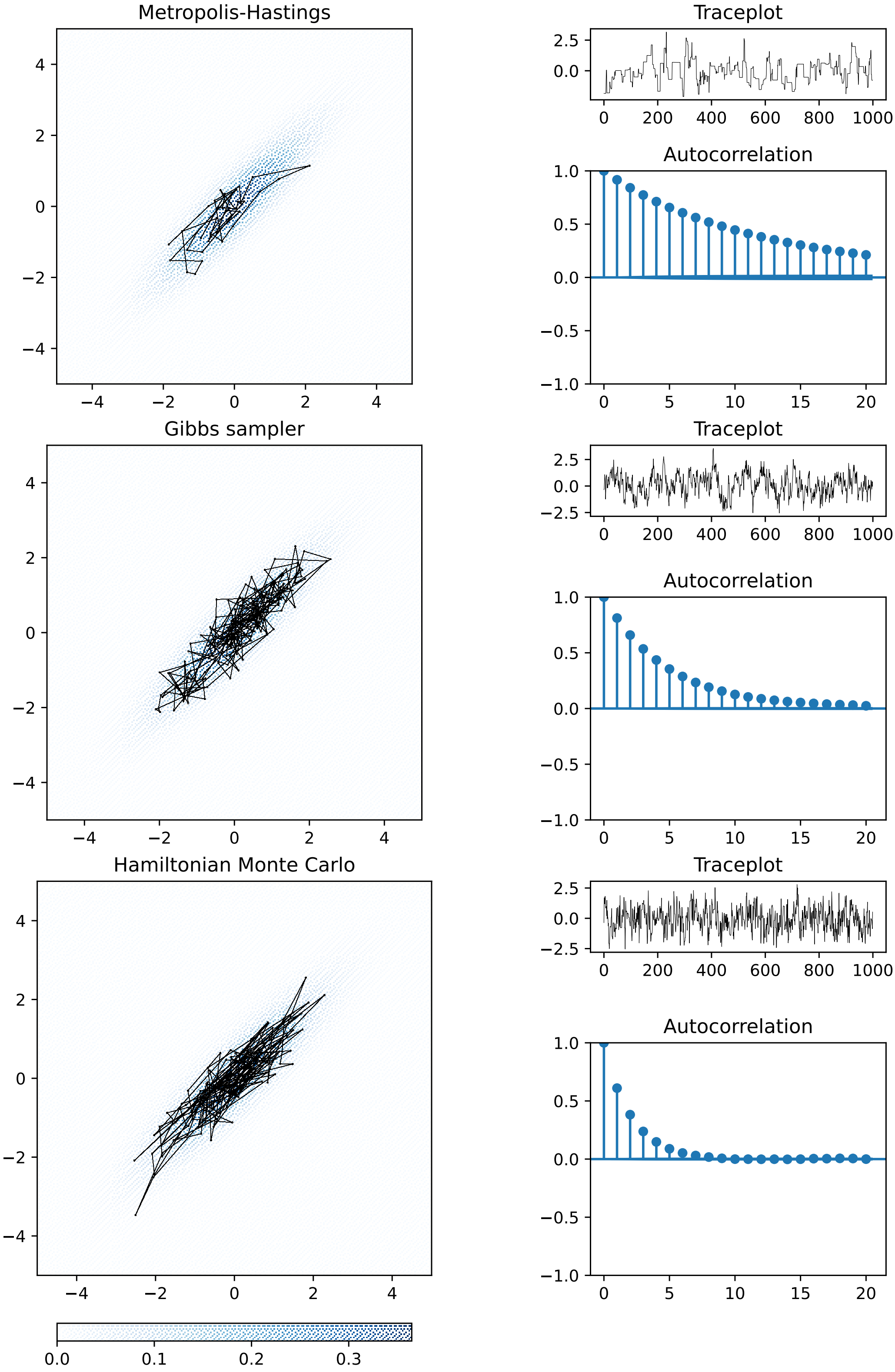


FIGURE 2.1: Comparison of the performance of MCMC algorithms for a strongly correlated bivariate example (see Example 2.2.3.4). The true distribution is shown in light blue in the plots on the left.

2.2.3.5 MCMC Methods for Functions

The MCMC algorithms presented so far are defined for finite-dimensional $\boldsymbol{\theta}$. If one is interested in inferring probability distribution of infinite-dimensional objects such as functions, care needs to be taken as such distributions do not have probability density functions. Although one could discretise the problem at hand and proceed with the derivation of an inference scheme on a finite-dimensional approximation, infinite-dimensional MCMC schemes derive the inference procedure on infinite-dimensional objects directly on probability measures, as opposed to probability density functions. Such methods are robust and scalable as the dimensions increases to infinity (Hairer et al. 2014).

For the purposes of Chapter 6, we consider the pre-conditioned Crank-Nicholson MCMC scheme proposed by Cotter et al. (2013). The objective is to sample from the posterior measure $\mu^{\boldsymbol{y}}(\theta)$ given a finite-dimensional observation $\boldsymbol{y}$ and a prior measure μ_0. Note that θ denotes a function in this section. We assume that a Gaussian prior measure is placed on θ. The posterior measure is related to the prior measure through Radon-Nikodym derivative.

The Markov Chain proposals are a combination of the current state and a sample from the prior, weighted by the parameter β. The uniform value of β may lead to slow convergence if the prior $\mu_0(\theta)$ is different from the posterior $\mu^y(\boldsymbol{\theta})$. An example of this is when for two different points in the domain, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, we have that $\frac{\mathrm{Var}_{\mu_0}\theta(\boldsymbol{x}_i)}{\mathrm{Var}_{\mu_0}\theta(\boldsymbol{x}_j)} \neq \frac{\mathrm{Var}_{\mu^y}\theta(\boldsymbol{x}_i)}{\mathrm{Var}_{\mu^y}\theta(\boldsymbol{x}_j)}$. To alleviate such problems, Rudolf & Sprungk (2018) developed an extension which accounts for the anisotropy in the covariance of the posterior or the local curvature of $\Phi(\theta, \boldsymbol{y})$ when MCMC proposals are made.

We summarise the procedure in Algorithm 2. For further details, we refer the reader to Cotter et al. (2013).

2.2.4 Variational Bayes

Variational Bayes is an optimisation-based technique for inferring posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. The true posterior is approximated by a variational distribution $q(\boldsymbol{\theta})$ which is the minimiser of the discrepancy between a chosen variational family $\mathcal{D}_q$ and the true

Algorithm 2: Pre-conditioned Crank-Nicholson MCMC (Cotter et al. 2013)

```
Input: $\Phi(\theta, \boldsymbol{y}) = -\log p(\boldsymbol{y} \mid \theta)$: negative likelihood of the data, $\mu_0(\theta)$: prior
       measure, $\beta$: corresponds to the amount of innovation in the proposal. If the
       value is small, there is little innovation and the proposed sample will be
       close to the previous sample.
Output: A list of samples from $\mu^y(\theta)$.
for $t \leftarrow 1, 2, \dots$ do
    Sample $\xi_{(t)} \sim \mu_0(\theta)$
    $v_{(t)} \leftarrow \sqrt{(1-\beta^2)}\theta_{(t)} + \beta\xi_{(t)}$
    $\theta_{(t+1)} \leftarrow \begin{cases} v_{(t)} & \text{with probability} \min\Big(1, \exp\big(\Phi(\theta_{(t)}; \boldsymbol{y}) - \Phi(v_{(t)}; \boldsymbol{y})\big)\Big) \\ \theta_{(t)} & \text{otherwise} \end{cases}$
return $[\theta_{(1)}, \theta_{(2)}, \dots]$
```

posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ (Jordan et al. 1999, Jordan & Wainwright 2007). A typical choice for the measure of discrepancy between distributions is the Kullback-Leibler (KL) divergence (which due to the lack of symmetry is not a metric). To find the approximate posterior distribution we have:

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta}) \in \mathcal{D}_q}{\arg\min} \; \mathrm{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \boldsymbol{y})). \tag{2.33}$$

Expanding the KL divergence term we obtain

$$\begin{aligned} \mathrm{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta} \mid \boldsymbol{y})) &= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})} \mathrm{d}(\boldsymbol{\theta}) \\ &= \mathbb{E}_q\big[\log q(\boldsymbol{\theta})\big] - \mathbb{E}_q\big[\log p(\boldsymbol{\theta} \mid \boldsymbol{y})\big] \\ &= \mathbb{E}_q\big[\log q(\boldsymbol{\theta})\big] - \mathbb{E}_q\Big[\log \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{y})}\Big] \\ &= \mathbb{E}_q\big[\log q(\boldsymbol{\theta})\big] - \mathbb{E}_q\big[\log p(\boldsymbol{y}, \boldsymbol{\theta})\big] + \log p(\boldsymbol{y}) \end{aligned} \tag{2.34}$$

The last term of the KL divergence, the log-marginal likelihood $\log p(\boldsymbol{y})$, is usually not known. However, we use the fact that the KL divergence is non-negative to obtain the bound

$$\log p(\boldsymbol{y}) \geq \mathbb{E}_q\big[\log p(\boldsymbol{y}, \boldsymbol{\theta})\big] - \mathbb{E}_q\big[\log q(\boldsymbol{\theta})\big]. \tag{2.35}$$

This inequality becomes an equality when the trial density $q(\boldsymbol{\theta})$ and the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ are equal. To minimise the KL divergence, it is sufficient to maximise $\mathbb{E}_q\big[\log p(\boldsymbol{y}, \boldsymbol{\theta})\big] - \mathbb{E}_q\big[\log q(\boldsymbol{\theta})\big]$, which is commonly referred to as the evidence lower bound (ELBO). The

ELBO term can be rewritten as

$$\begin{aligned}\mathrm{ELBO}(q) &= \mathbb{E}_q\big[\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})\big] - \mathbb{E}_q\big[\log q(\boldsymbol{\theta})\big] \\ &= \mathbb{E}_q\big[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})\big] - \mathrm{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})).\end{aligned} \tag{2.36}$$

To summarise, the task now becomes:

$$q^*(\boldsymbol{\theta}) = \underset{q(\boldsymbol{\theta})\in\mathcal{D}_q}{\arg\max}\ \mathbb{E}_q\big[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})\big] - \mathrm{KL}(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})). \tag{2.37}$$

To maximise the ELBO with a gradient-based optimiser, we need to evaluate the ELBO and its gradients with respect to the parameters of $q(\boldsymbol{\theta})$. Although the KL divergence term of the ELBO is often available in closed form, $\mathbb{E}_q\big[\log p(y \mid \boldsymbol{\theta})\big]$ involving the likelihood is generally not available. It can be approximated using a Monte Carlo approximation with N_{SVI} samples from the trial density $q(\boldsymbol{\theta})$ as follows:

$$\mathbb{E}_q\big[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})\big] \approx \frac{1}{N_{\mathrm{SVI}}} \sum_{i=1}^{N_{\mathrm{SVI}}} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(i)}), \tag{2.38}$$

where $\boldsymbol{\theta}^{(i)}$ is the i-th sample from $q(\boldsymbol{\theta})$. The computations of gradients with respect to parameters of $q(\boldsymbol{\theta})$ is problem-dependent. A common technique is the reparametrisation trick, as described in Section 2.2.4.1. The choice of N_{SVI} which leads to fast convergence of the optimisation has been shown in the literature to be in the range of 2–5 (Kingma & Welling 2014). This approach is often referred to as stochastic variational inference (SVI).

2.2.4.1 Reparametrisation Trick

The reparametrisation trick allows computing the gradients of quantities derived from samples from a probability distribution with respect to the parameters $\boldsymbol{\phi}$ of that probability distribution. This holds for probability distributions where samples can be obtained by a deterministic mapping, parametrised by $\boldsymbol{\phi}$, of other random variables.

Let $\boldsymbol{\epsilon}$ be a set of random variables. We assume that samples of $\boldsymbol{\theta} \sim q(\boldsymbol{\theta}; \boldsymbol{\phi})$ are given by a deterministic mapping

$$\boldsymbol{\theta} = t(\boldsymbol{\phi}, \boldsymbol{\epsilon}). \tag{2.39}$$

The KL divergence between approximating distribution $q(\boldsymbol{\theta})$ and the prior $p(\boldsymbol{\theta})$ is often available in closed form and so are its gradients with respect to $\boldsymbol{\phi}$. To estimate the gradients of the Monte Carlo estimate of the log-likelihood of the data,

$$\mathbb{E}_q\big[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})\big] \approx N_{\text{SVI}}^{-1} \sum_{i=1}^{N_{\text{SVI}}} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(i)}), \tag{2.40}$$

we can use the chain rule of differentiation to obtain

$$\nabla_{\boldsymbol{\phi}} \left[N_{\text{SVI}}^{-1} \sum_{i=1}^{N_{\text{SVI}}} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(i)}) \right] = N_{\text{SVI}}^{-1} \sum_{i=1}^{N_{\text{SVI}}} \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(i)}) \cdot \nabla_{\boldsymbol{\phi}} t(\boldsymbol{\phi}, \boldsymbol{\epsilon}^{(i)}). \tag{2.41}$$

2.2.4.2 Variational Bayes Remarks

- The choice of the approximating distribution is task-dependent, but a multivariate Gaussian distribution has been a popular choice due to analytical tractability (Blei et al. 2017). Different parametrisations of the Gaussian distribution and other non-Gaussian options such as mixture distributions are discussed in Section 6.3.

- To maximise the ELBO in (2.37), the ADAM algorithm (Kingma & Ba 2015) has been widely used, and it is what we use in this thesis. One may wish to employ a different optimisation algorithm, and this choice should be made depending on the problem-specific requirements. This is discussed in more detail in Section 6.3.4.

Chapter 3

Spatial Point Processes

3.1 Definition

Point processes are mathematical models for random point patterns over a domain D which we assume to be a complete separable metric space (Daley & Vere-Jones 2003). We consider planar point patterns, *i.e.*, $D \subset \mathbb{R}^2$. Let X denote a point process. For a given Borel set $A \subset D$, the counting measure $N(A)$ is the random number of points of X contained in A (Stoyan & Stoyan 1994). We only consider simple processes, *i.e.*, no multiple points. Thus, the set of *distinct* points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ is a random countable set

$$\mathcal{X} = \{\boldsymbol{x}_i\}_i, \tag{3.1}$$

where $\mathcal{X}$ is referred to as a configuration. Each $\boldsymbol{x}_i$ lies in D, and a set of n points lies in D^n.

If the number of points is not finite, we require that configurations place at most a finite number of points in any *bounded* Borel set $A \subset D$:

$$N_{\mathcal{X}}(A) = \sum_{i=1}^{\infty} \mathbf{1}(\boldsymbol{x}_i \in A) < \infty. \tag{3.2}$$

We denote the family of locally finite configurations as N^{lf} and we equip it with the smallest σ-algebra $\mathcal{N}^{\text{lf}}$ such that the mapping $\mathcal{X} \mapsto N_{\mathcal{X}}(A)$ is measurable. The formal definition of a point process then follows.

Definition 3.1 (Point process (van Lieshout 2010))**.** Let (D, d) be a complete, separable metric space equipped with its Borel σ-algebra $\mathcal{B}$. A point process on D is a mapping X from an abstract probability space $(\Omega, \mathcal{A}, \mathbb{P})$ into N^{lf} such that for all bounded Borel sets $A \subset D$, the number $N(A) = N_X(A)$ of points falling in A is a finite random variable. In other words, a point process X is a random variable with values in the measurable space $(N^{\text{lf}}, \mathcal{N}^{\text{lf}})$.

The induced probability measure $\mathcal{P}$ on $\mathcal{N}^{\text{lf}}$ is the *distribution* of the point process. For $F \in \mathcal{N}^{\text{lf}}$, we have

$$\mathcal{P}(F) = \mathbb{P}(\{\omega \in \Omega \colon X(\omega) \in F\}). \tag{3.3}$$

If $\mathcal{P}$ is translation invariant, *i.e.*, the distribution does not change if all points $\boldsymbol{x}_i \in X$ are translated over $\boldsymbol{y}$ into $(\boldsymbol{x}_i + \boldsymbol{y})$, then X is *stationary*. Similarly, if $\mathcal{P}$ is rotation invariant, X is *isotropic*.

The probability measure $\mathcal{P}$ is induced by integer-valued random variables $N(A)$ counting the number of points in a bounded Borel set A. Thus, it is natural to characterise $\mathcal{P}$ by specifying the properties of $N(A)$. One can do so through the family of *finite-dimensional distributions* which are a collection of joint distributions $(N(A_1), N(A_2), \ldots, N(A_m))$, where $(A_1, \ldots, A_m)$ ranges over the bounded Borel sets $A_i \subset D$, $i = 1, \ldots, m$, and $m \in \mathbb{N}$. It has been shown that if two point processes have identical finite-dimensional distributions, they also share the same distribution (van Lieshout 2010, Theorem 16.1). In other words, the distribution of a point process X is completely specified by its finite-dimensional distributions. If the point process is simple (no multiple points), as we assume throughout the thesis, void probabilities of bounded Borel sets $A \subset D$ also completely determine the distribution of the process:

$$\mathbb{P}(N(A) = 0). \tag{3.4}$$

Before we proceed with further summary of point process methodology, we show an example of a point pattern observed in real life.

3.1.1 Point Process Example

We present an example of an observed point pattern of the locations of trees in a tropical rain forest (Baddeley & Turner 2005).

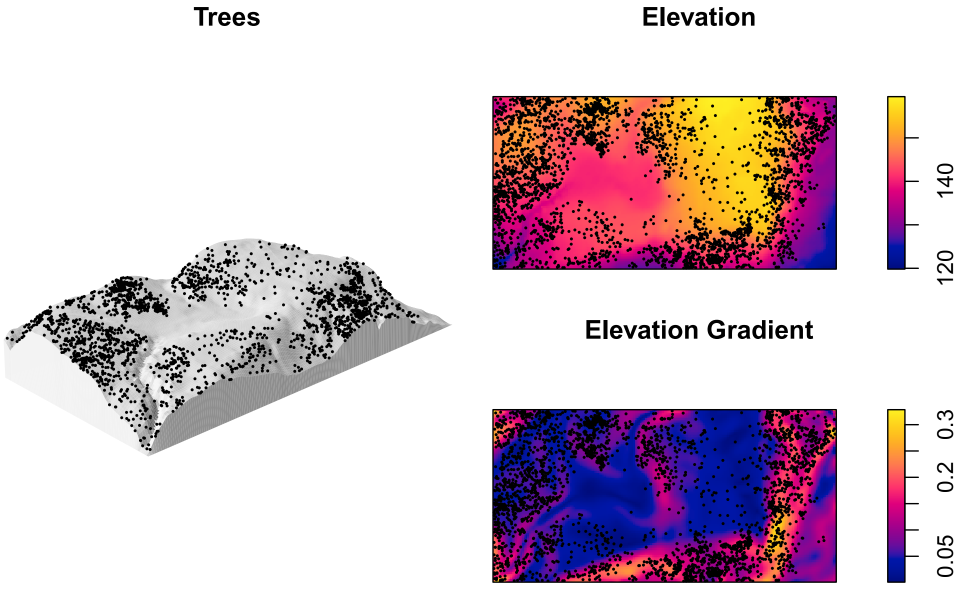


FIGURE 3.1: A point pattern of 3605 trees on Barro Colorado Island together with the corresponding elevation and the norm of the elevation gradient. This figure was produced using an *R* package provided by Baddeley & Turner (2005).

Example 3.2. *Figure 3.1 shows the locations of 3605 trees of the species Beilschmiedia pendula in a 1000 by 500 metre region in the tropical rainforest of Barro Colorado Island. The figure also shows the elevation (in metres) as well as the norm of the elevation gradient. One may for example be interested in analysing the relationship between the intensity of the point pattern and the elevation, or its gradient.*

3.2 Summary Statistics

Summary statistics provide a concise characterisation of an observed point pattern, which may then be used for further analysis. Firstly, we consider moment measures of the finite-dimensional distributions used to specify the distribution of point process X. The first-order measure, often called intensity measure, is a set function M such that

$$M(A) = \mathbb{E}\, N(A) < \infty \tag{3.5}$$

for all bounded Borel sets $A \subset D$. If M is absolutely continuous with respect to the Lebesgue measure ν on D, then

$$M(A) = \int_A \lambda(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x}), \tag{3.6}$$

for Borel sets A, and $\lambda(\cdot)$ is referred to as the *intensity function* of the point process X. As a consequence of this, we can write

$$\int_A \mathrm{d}M(\boldsymbol{a}) = \mathbb{E}\left[\sum_{\boldsymbol{x} \in X} \mathbf{1}\{\boldsymbol{x} \in A\}\right] \tag{3.7}$$

for all $A \in \mathcal{B}(D)$. Thanks to linearity and monotonicity we obtain the Campbell theorem:

$$\mathbb{E}\left[\sum_{\boldsymbol{x} \in X} g(\boldsymbol{x})\right] = \int_D g(\boldsymbol{x}) dM(\boldsymbol{x}) \tag{3.8}$$

for any measurable function $g \colon D \to \mathbb{R}$. This allows computing expectations of random sums using integrals involving the mean measure M.

The moment measures may be refined to restrict attention to configurations with specific properties, for example the Campbell measure of the first order is defined as follows.

Definition 3.3 (Campbell measure)**.** Let X be a point process on D. The first-order Campbell measure is defined as

$$C(A \times F) = \mathbb{E}\left[N(A)\mathbf{1}\{X \in F\}\right] \tag{3.9}$$

for all bounded Borel sets $A \subset D$ and all $F \in \mathcal{N}^{\mathrm{lf}}$.

Higher-order moment measures are defined by considering tuples of $(N(A_1), N(A_2), \dots)$ jointly. The n-th order moment measure is defined as

$$M_n(A_1 \times \cdots \times A_n) = \mathbb{E}\left[N(A_1), \dots, N(A_n)\right]. \tag{3.10}$$

The Campbell theorem involving nth order measure in the integral representation reads as

$$\mathbb{E}\left[\sum_{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n} g(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)\right] = \int_D \cdots \int_D g(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \mathrm{d}M_n(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) \tag{3.11}$$

For example, the covariance of the random variables $N(A)$ and $N(B)$ can be written as

$$\mathrm{Cov}(N(A), N(B)) = M_2(A \times B) - M(A)M(B). \tag{3.12}$$

In (3.11), we considered tuples which include both identical and distinct points. By considering only distinct points, we obtain the n-th order *factorial moment measure* μ_n, defined by the integral representation:

$$\mathbb{E}\left[\sum_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}^{\neq} g(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\right] = \int_D \cdots \int_D g(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \mathrm{d}\mu_n(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n), \tag{3.13}$$

for all measurable functions $g\colon D^n \to \mathbb{R}$. The n-th order factorial moment measure exists if $\mu_n(A)$ is finite for all bounded Borel sets $A \subset D$ (van Lieshout 2010). If μ_n is absolutely continuous with respect to some n-fold product measure ν^n on D^n, the we can rewrite (3.13) as

$$\mathbb{E}\left[\sum_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n}^{\neq} g(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\right] = \int_D \cdots \int_D g(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\rho_n(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n) \mathrm{d}\nu(\boldsymbol{x}_1)\ldots\mathrm{d}\nu(\boldsymbol{x}_n), \tag{3.14}$$

where the Radon-Nikodym derivative $\rho_n(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$ is referred to as *product density*. The expression $\rho_n(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)\mathrm{d}\nu(\boldsymbol{x}_1)\ldots\mathrm{d}\nu(\boldsymbol{x}_n)$ may be interpreted as the join probability of a point falling in each of the infinitesimal regions centred at $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n$. The product density is used for deriving likelihood for point processes, as we shall see below in (3.23).

Further summary statistics such as Campbell reduced measures or Palm conditioning are often used for analysis of point processes. In this thesis, we will not make use of such techniques and therefore omit them from this review. For the interested reader, we recommend Daley & Vere-Jones (2003), van Lieshout (2010).

3.3 Finite Point Processes

Most of the point patterns in practice are finite in the number of points and are observed on a bounded region, which is either dictated by the application or a smaller observation 'window' is chosen for convenience. The distribution of a point process which is finite (with probability 1) can be conveniently described as follows. Suppose that D is equipped with the Lebesgue measure ν and $0 < \nu(D) < \infty$, then it is sufficient to specify:

1. A discrete probability distribution, $(p_n)_{n\in\mathbb{N}_0}$ for the total number N of points.
2. A family of symmetric probability densities $\pi_n(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)_{n\in\mathbb{N}_0}$ with respect to the n-fold product of ν for the point locations.

The symmetry requirement for π_n is necessary to ensure that the patterns generated by π_n are permutation invariant, *i.e.*, not dependent on the order in which they are listed. This formulation is the basis for the models we consider in this thesis.

3.4 Models and Estimation

We limit this exposition to finite point processes which are observed on a bounded observation window D. After observing a sample of a point pattern, the goal is to describe the distribution of points and build a statistical model explaining the variation in the observed pattern. We assume that the data consists of exactly one observation of the point pattern. We use the inference methods described in Chapter 4, where we focus on inferring the first moment of the point process, and specifically, the *intensity function* (which we assume exists, see (3.6) above).

3.4.1 Non-parametric Models

The first class of models which is mostly used for exploratory analysis is *non-parametric models*. One of the objectives of exploratory analysis is to ascertain whether the observed point pattern is spatially random, commonly referred to as *complete spatial randomness* (CSR). A point process with CSR property is a stationary Poisson process which we discuss in one of the subsequent sections. From a finite sample of n points, $\{\boldsymbol{x}_i\}_{i=1}^{n}$, it is not trivial to understand whether the underlying process which gave rise to the observed point pattern is CSR or not. Additional context is often required to make this assumption. A common approach to test this assumption is to perform a statistical hypothesis test using statistics based on various distances such as K-function or L-function (Stoyan & Stoyan 1994, Ripley 1976).

For the estimation of the intensity function in a non-parametric way, kernel estimators are a powerful tool (Diggle 1985). An elementary estimator is given as

$$\hat{\lambda}_h(\boldsymbol{x}) = \frac{N(b(\boldsymbol{x}, h)}{\nu(b(\boldsymbol{x}, h))}, \tag{3.15}$$

where $b(\boldsymbol{x}, h)$ is the ball of radius h around $\boldsymbol{x}$, and ν is the Lebesgue measure on D. This histogram-like estimator can be smoothed out by applying a smoothing kernel k_h:

$$\hat{\lambda}_h(\boldsymbol{x}) = \sum_{i=1}^{n} k_h(\boldsymbol{x} - \boldsymbol{x}_i). \tag{3.16}$$

Popular choices of k_h include Epanechnikov kernel and Gaussian kernel (Stoyan & Stoyan 1994). After a family of kernel functions has been chosen, the quality of the estimator is mostly determined by the choice of h, which is often referred to as bandwidth.

For stationary processes, the intensity can be simply estimated by

$$\hat{\lambda} = n/|D|. \tag{3.17}$$

This estimator is always unbiased, and if the underlying point process is a stationary Poisson process, it is also the maximum likelihood estimator with variance equal to $\lambda/|D|$ (Gelfand 2010).

3.4.2 Parametric Models

The second class of models assumes a parametric form for the distribution of the point process. Once a class of models has been chosen, the objective is to infer the parameter(s) and validate the model specification. The most fundamental model in spatial statistics is the stationary Poisson process, often called *homogeneous*. It is used for modelling point patterns where locations of points are completely random and independent of each other. If such independence conditions do not hold, then a Poisson process is often used as either a rough approximation, or as a null hypothesis model. Below, we give a formal definition, discuss the non-homogeneous case, as well as the case where the intensity function is stochastic.

3.4.2.1 Homogeneous Poisson Process

Definition 3.4 (Homogeneous Poisson process (Stoyan & Stoyan 1994))**.** The homogeneous Poisson process is defined by two properties:

1. If k is any integer and $A_1, \dots, A_k$ are any disjoint Borel sets on D then the random variables $N(A_1), \dots, N(A_k)$ are stochastically independent.
2. The number of points $N(A)$ in any bounded Borel set A has a Poisson distribution with the intensity parameter equal to $M(A) = \mathbb{E}\,N(A)$. For a finite Poisson process, we have $M(A) = \lambda\nu(A)$, where $\nu(A)$ denotes the Lebesgue measure on D, which throughout this thesis represents an area in $\mathbb{R}^2$ space.

This definition implies that the field is stationary and isotropic. More specifically, if a bounded set $A \subset D$ has exactly n points in it, then these points are distributed uniformly and independently in A. This property is useful for the simulation of homogeneous Poisson point processes.

The probability of observing k points in a bounded Borel set A is then given by the probability mass function of Poisson distribution:

$$\mathbb{P}(N(A) = k) = \frac{\left[\lambda\nu(A)\right]^k}{k!} e^{-\lambda\nu(A)}. \tag{3.18}$$

The variance of $N(A)$ follows from the properties of the Poisson distribution:

$$\mathrm{Var} N(A) = M(A) = \lambda\nu(A). \tag{3.19}$$

The estimate of the intensity for an observed sample $\{\boldsymbol{x}_i\}_{i=1}^n$ on a bounded domain D simply reads as

$$\hat{\lambda} = \frac{N(D)}{\nu(D)}. \tag{3.20}$$

3.4.2.2 Non-homogeneous Poisson Process

In situations where the intensity of points is not constant throughout the domain, non-homogeneous Poisson process may be a suitable alternative. The definition follows similarly to the homogeneous case, except that the intensity measure for a bounded Borel set

A, $M(A)$, is no longer dependent only on the Lebesgue measure $\nu(A)$, but also considers the spatially varying intensity function as defined below.

Definition 3.5 (Non-homogeneous Poisson Process (Stoyan & Stoyan 1994))**.** Non-homogeneous Poisson process can be defined by changing the second property of Definition 3.4 to:

2. The number of points $N(A)$ in a bounded Borel set A has a Poisson distribution with parameter $M(A) = \mathbb{E}\,N(A)$. Provided the intensity function $\lambda(\cdot)$ exists (see (3.6)), we have

$$M(A) = \int_A \lambda(\boldsymbol{x})\mathrm{d}\nu(\boldsymbol{x}). \tag{3.21}$$

Similarly to the homogeneous case we have

$$\mathrm{Var} N(A) = M(A). \tag{3.22}$$

The product density (see (3.14)) takes a simple form for an inhomogeneous Poisson process (due to the independence property in Definition 3.5 (Møller & Waagepetersen 2007)):

$$\rho_n(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \prod_{i=1}^{n} \lambda(\boldsymbol{x}_i), \tag{3.23}$$

which implies that there are no interactions between the points themselves.

While the intensity can be estimated in a non-parametric manner as described in Section 3.4.1, we can perform likelihood-based inference of $\lambda(\cdot)$. The likelihood is given as the probability of obtaining a given number of points n on a bounded domain D times the joint conditional density for the locations of these points, given the number of observed points n. While the probability of obtaining n points follows from the probability mass function of Poisson distribution, $\frac{(\int_D \lambda(\boldsymbol{x})\mathrm{d}\nu(\boldsymbol{x}))^n}{n!\exp(-\int_D \lambda(\boldsymbol{x})\mathrm{d}\nu(\boldsymbol{x}))}$, each of the n points at locations $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ is distributed across D with density

$$\frac{\lambda(\boldsymbol{x})}{\int_D \lambda(\boldsymbol{x}')\mathrm{d}\nu(\boldsymbol{x}')}. \tag{3.24}$$

This is a consequence of the independence property of the Poisson processes which allows for the factorisation of the product density as in (3.23). Taken together, the likelihood follows as

$$\begin{aligned}L(\lambda; \boldsymbol{x}_1, \dots, \boldsymbol{x}_n) &= \frac{\left(\int_D \lambda(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x})\right)^n}{n! \exp(-\int_D \lambda(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x}))} \prod_{i=1}^{n} \frac{\lambda(\boldsymbol{x}_i)}{\int_D \lambda(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x})} \\ &\propto \exp\left(-\int_D \lambda(\boldsymbol{x}) \mathrm{d}\nu(\boldsymbol{x})\right) \prod_{i=1}^{n} \lambda(\boldsymbol{x}_i).\end{aligned} \tag{3.25}$$

Finding the $\lambda(\cdot)$ which maximises the likelihood is the central task of likelihood-based inference for inhomogeneous Poisson processes (Møller & Waagepetersen 2004, 2007, Gelfand 2010).

3.4.2.3 Cox Process

The Cox process is an extension of the inhomogeneous Poisson process. This model is used when the aggregation of points in a point pattern is due to stochastic environmental heterogeneity (Diggle et al. 2013, Stoyan & Stoyan 1994). The intensity of a Cox process X is driven by a non-negative process $\Lambda = (\Lambda(\boldsymbol{x}))_{\boldsymbol{x} \in D}$, such that conditional on $\Lambda(\boldsymbol{x}) = \lambda(\boldsymbol{x})$, X is a non-homogeneous Poisson process with intensity function $\lambda(\boldsymbol{x})$. The properties of Λ determine the properties of X. Most importantly, if Λ is a stationary process, then so is X. One cannot distinguish the Cox process X from its corresponding Poisson process $X \mid \Lambda$ when only one realisation of X is available. The likelihood is, in general, unknown while the product densities may be tractable (Gelfand 2010, Møller & Waagepetersen 2007).

Of particular interest are Cox processes, where $\log \Lambda$ is defined to be a Gaussian process (Møller et al. 1998). If we include the fixed-effects term with spatial covariates $\boldsymbol{z}$, the log-intensity is then given as

$$\log \Lambda(\boldsymbol{x}) = \boldsymbol{z}(\boldsymbol{x})^\top \boldsymbol{\beta} + f(\boldsymbol{x}), \tag{3.26}$$

where $\boldsymbol{\beta}$ are coefficients, f is a zero-mean Gaussian process (see a summary of Gaussian processes in Section 6.2.4 of Chapter 6, where we use them extensively) with the covariance function $c(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{Cov}[f(\boldsymbol{x}), f(\boldsymbol{y})]$.

Estimation of this process is non-trivial due to the doubly-stochastic nature of the process. For example, each likelihood computation requires integration over the process Λ, *i.e.*, evaluating $\mathbb{E}_\Lambda L(\Lambda; \boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$. In Chapter 4, we use a log-Gaussian Cox process as the baseline model. Our estimation procedure follows the Bayesian framework where we sample the posterior distribution of f and $\boldsymbol{\beta}$ using MCMC methods. For details, see Section A.2.

Chapter 4

Burglary in London: Insights from Statistical Heterogeneous Spatial Point Processes

4.1 Introduction

Use of statistical models for understanding and predicting criminal behaviour has become increasingly relevant for police forces, and policymakers (Felson & Clarke 1998, Bowers & Hirschfield 1999, PredPol 2019). While short-term forecasting of criminal activity has been used to better allocate policing resources (Taddy 2010, Mohler et al. 2011, Aldor-Noiman et al. 2016, Flaxman et al. 2019, PredPol 2019), understanding the criminal behaviour and target selection process through statistical models has a potential to be used for designing policy changes and development programs (Felson & Clarke 1998). In this work, we consider the problem of burglary crime in London. In the UK, burglary is a well-reported crime, but the detection rate remains at the 10-15% level (Smith et al. 2013). Rather than being concerned with short-term forecasting, we focus on understanding the effects of spatially varying explanatory variables on the target selection through descriptive regression models. Inferences made using these models help us understand the underlying mechanisms of burglary. The main contribution of this work is the integration of statistical methods in spatial modelling with the findings from the criminological literature.

Instances of burglary can be represented as a *spatial point pattern* – a finite or countably infinite set of points in the study region. Understanding the intensity of the occurrences through spatially varying covariates is the main objective of this work. The task of estimating the effects of the covariates on the intensity can be classified as a multivariate regression modelling, in which systematic effects of the explanatory variables are of interest while taking into account other random effects such as measurement errors and spatial correlation (McCullagh & Nelder 1998). In the context of spatial data, it has been widely recognised that multivariate regression modelling techniques which do not account for *spatial dependence* and *spatial heterogeneity* can lead to biased results and faulty inferences (Anselin et al. 2000). Spatial dependence refers to the Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things"(Tobler 1970). Spatial dependence manifests itself mostly in the spatial correlation of the residuals of a model. In non-spatial settings, the residuals are often assumed to be independent and identically distributed (McCullagh & Nelder 1998). Spatial heterogeneity is exhibited when the object of interest, in our case, the intensity of a point pattern, shows location-specific behaviour. For example, properties of the burglary point pattern in a city centre are going to be different from the properties in a residential area. Formalising these two concepts and incorporating them into modelling methodology results in more accurate spatial models (Anselin et al. 2000).

Log-Gaussian Cox process (Møller et al. 1998, Møller & Waagepetersen 2007) has been a common approach for modelling intensity of spatial point patterns (Diggle et al. 2013, Serra et al. 2014). The flexibility of the model is due to the Gaussian process part through which complex covariance structures, including spatial dependence and heterogeneity, can be accounted for. In practice, stationary covariance functions are used for computational and identifiability reasons (Diggle et al. 2013). Identifiability of non-stationary covariance models is challenging without the knowledge of problem-specific structure. As a result, log-Gaussian Cox process models with stationary covariance functions handle spatial dependence but do not account for spatial heterogeneity.

Mixture based approaches have been adopted as a way of enriching the collection of probability distributions to account for spatial heterogeneity often observed in practice (Green 2010, Fernández & Green 2002). Notably, Knorr-Held & Raßer (2000), Fernández & Green (2002), Green & Richardson (2002) used mixtures for modelling the elevations of disease prevalence. While these methods improve the model fit by accounting for spatial heterogeneity as wells as spatial dependence, they provide little interpretation as to why the level is elevated in certain areas. Also, these three methods have been tested only at a modest scale. Following this line of work, Hildeman et al. (2018) proposed a method in which each mixture component can take a rich representation that may include covariates. Although this model is very rich in representation, the empirical study in the paper was limited to the case of two mixtures, with one of the components being held constant. Their study of a tree point pattern and its dependence on soil type was carried out on a region discretised into a grid with 2461 cells.

A very different approach to controlling for spatial heterogeneity has been taken by Gelfand et al. (2003) who allow regression coefficients to vary across the spatial region. The method treats the coefficients of the covariates as a multivariate spatial process. The process is, however, very challenging to fit and is often limited to 2 or 3 covariates (Banerjee et al. 2015, p.288). A simpler version of the same idea is geographically-weighted regression (Brunsdon et al. 1996), where the regression coefficients are weighted by a latent component whose properties have to be specified a priori or learned through cross-validation.

Motivated by the computational challenges and limited interpretability of the aforementioned approaches, we propose a mixture based method that takes into account spatial dependence and is able to discover latent groups of locations and characterise each group by group-specific effects of spatially varying covariates. To estimate the model parameters from the limited data and to quantify the uncertainty of the estimates, we follow the Bayesian framework.

More specifically, our approach builds upon the mixtures of generalised linear models (Grün & Leisch 2008), in which observations are modelled as a mixture of different models. We cater for spatial dependence using an approach inspired by Fernández & Green (2002) and Knorr-Held & Raßer (2000). Our model probabilistically assigns each location to a particular mixture component, while imposing spatial dependence through prior information. The prior information will suggest that locations that are close to each other are likely to belong to the same component. We define a pair of locations to be close if both of them are in the same block. We use the blocking structure predefined by the census tracts, but our method allows defining custom ones. We further model spatial dependence of the blocks using latent Gaussian processes, following Fernández & Green

(2002). The posterior inferences for the individual components consisting of regressions coefficients and the assignments of locations to clusters are used to draw conclusions and provide insights about the heterogeneity of the spatial point pattern across the study region.

In contrast to Fernández & Green (2002) and Green & Richardson (2002), this work considers including the covariates into each mixture component, rather than having intercept-only components. Compared to the approach of Hildeman et al. (2018) who model the log-intensity of a point pattern as a mixture of Gaussian random fields, our model is more constrained but provides better scalability.

We show that the proposed methodology effectively models burglary crime in London. By comparing our approach to log-Gaussian Cox process (LGCP), a standard model for spatial point patterns (Diggle et al. 2013), we show that our method outperforms LGCP and is more computationally tractable. Lastly, the interpretation of inferred quantities provides useful criminological insights.

The rest of the paper is structured as follows. Section 4.2 defines the model and details the inference method, Section 4.3 elaborates on our application and gives the discussion of model choices specific to our application. The obtained results are discussed in Section 4.4. Section 4.5 concludes the chapter.

4.2 Modelling methodology

It is widely recognised that burglary crime is spatially concentrated (Brantingham & Brantingham 1981, Clare et al. 2009, Johnson & Bowers 2010). It is also apparent that some areas in the study region will exhibit extreme behaviour. For example, areas with no buildings such as parks will have no burglary for structural reasons. To effectively model burglary, these phenomena need to be accounted for using *spatial effects*. The two important spatial effects are *spatial dependence* and *spatial heterogeneity* (Anselin et al. 2000).

For our modelling framework, we choose the Bayesian paradigm because it allows us to formalise prior knowledge, and to quantify uncertainty in the unknown quantities of our model. In our application, burglary data are given as a point pattern over a fixed period of time. Chapter 3 provides the prerequisite background on point patterns and discusses several modelling approaches and related issues. In this chapter, we discretise the point pattern onto a grid of N cells by counting the points in each cell. Although any form of discretisation is allowed, throughout this paper, we work with a regular grid.

We model the count of points in a cell n, y_n, conditioned on the mixture component k as a Poisson-distributed random variable, with the logarithm of the intensity driven by a linear term, which is specific for each mixture component, indexed by $k = 1, \dots, K$. The linear term is a linear combination of J covariates for cell n, $\boldsymbol{X}_n$, and the corresponding coefficients, $\boldsymbol{\beta}_k$. The covariates need to be specified for the application of interest and usually include the intercept. To specify the prior distribution for the regression coefficients, we use a prior that shrinks the estimate towards zero. For each coefficient, we set $\beta_{k,j} \sim \mathcal{N}(0, \sigma^2_{k,j})$, where $\sigma^2_{k,j} \sim \text{InvGamma}(1, 0.01)$. We put the uniform prior on the intercepts, if present.

Each cell n is probabilistically allocated to one of the K components through an allocation variable, z_n, which is a categorical random variable with event probabilities given by the mixture weights prior, $\boldsymbol{\pi}_{b[n]}$. The value of $\boldsymbol{\pi}_{b[n]}$ is shared for all locations within cell n's block, $b[n]$. The blocks for the study region are defined as non-overlapping spatial areas spanning the whole study region. In many practical applications, the block structure is already defined by administrative units or census tracts. Block $b[n]$ is the block that contains the centroid point of cell n. The block-specific event probabilities will express the belief that the effect of the covariates is the same within the block unless evidence from the observed data outweighs this information. Note that conditional predictors (conditional on the allocations $(z_n)_{n=1}^N$) only use the fixed covariate effects.

To model the mixture weights prior for block b, $\boldsymbol{\pi}_b = (\pi_{1,b}, \dots, \pi_{K,b})$, we allow for different choices provided that $\pi_{k,b} \geq 0$ and $\sum_{k=1}^K \pi_{k,b} = 1$, *i.e.*, it is a valid probability measure. One possible choice which also takes into account the spatial dependence between the blocks is to model the mixture weights prior for block b and mixture component k as

$$\pi_{k,b} = \frac{\exp(f_{k,b})}{\sum_{l=1}^K \exp(f_{l,b})},$$

where $f_{k,b}$ is the shorthand for the evaluation of $f_k(\cdot)$ at the centroid of the block b, and $f_k(\cdot)$ is an independent zero-mean Gaussian Process defined over $\boldsymbol{x} \in D \subset \mathbb{R}$ with hyper-parameters $\boldsymbol{\theta}_k$. The prior for $\boldsymbol{\theta}_k$ is specified depending on the kernel function used.

We will use the squared exponential kernel throughout this work (Rasmussen & Williams 2006).

We refer to the proposed model as a *spatially-aware mixture of Poisson generalised linear models* (SAM-GLM). The formulation is summarised in the equation and the graphical representation shown in Figure 4.1. In the proposed model, we handle *spatial heterogeneity* using the mixture components, each of which specifies a set of J regression coefficients, $\boldsymbol{\beta}_k$. *Spatial dependence* is considered first within each block and also through inter-block dependence imposed by K Gaussian processes. Modelling the spatial dependence using Gaussian processes with the training points defined by the coarse block centroids instead of finer cell centroids allows for more efficient estimation procedures, as we discuss later.

$$
\begin{aligned}
y_n | z_n = k, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{X}_n &\sim \text{Poisson}\left(\exp\left(\boldsymbol{X}_n^\top \boldsymbol{\beta}_k\right)\right) \\
z_n | \boldsymbol{\pi} &\sim \text{Cat}(\pi_{1,b[n]}, \ldots, \pi_{K,b[n]}) \\
\boldsymbol{\pi}_{k,b} | f_k(\cdot) &= \frac{\exp(f_{k,b[n]})}{\sum_{l=1}^{K} \exp(f_{l,b[n]})} \\
f_k(\cdot) | \boldsymbol{\theta}_k &\sim \mathcal{GP}(0, \kappa_{\boldsymbol{\theta}_k}(\cdot,\cdot)) \\
\boldsymbol{\theta}_k &\sim \text{kernel-dependent prior} \\
\beta_{k,j} | \sigma_{k,j}^2 &\sim \mathcal{N}(0, \sigma_{k,j}^2) \\
\sigma_{k,j}^2 &\sim \text{InvGamma}(1, 0.01).
\end{aligned}
$$

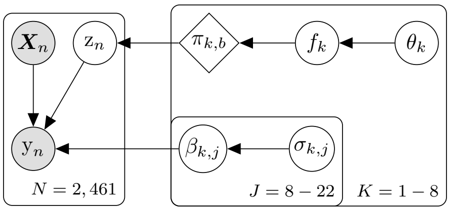


FIGURE 4.1: Summary of the SAM-GLM model and its graphical representation.

4.2.1 Excess of Zeros, Overdispersion

Two common challenges encountered when modelling count data using standard Poisson generalised linear models (GLM) are *excess of zeros* and *overdispersion* (McCullagh & Nelder 1998, Breslow 1984). The former refers to the presence of zeros that are structural, rather than due to chance. In the context of burglary, structural zeros occur in locations with no buildings, *e.g.*, parks. The latter issue refers to the situation when the variability of the observed data is higher than what would be expected based on a particular statistical model. The standard Poisson GLM for the burglary point pattern, a special case of our model ($K = 1$), suffers from overdispersion for different specifications of the covariates term – see Section A.1 in the appendix. The flexibility of our proposed model can account for the excess of zeros by identifying a low-count component to which areas of low intensity will be assigned. Similarly, introducing mixtures can reduce overdispersion. Two cells with similar values for the covariates, but with very different observed counts are likely to have the same expected count under the standard Poisson GLM. Under the mixture model, each cell would be allowed to follow a different model.

4.2.2 Inference

Statistical inference in the Bayesian setting involves inferring the posterior probability distribution for the quantities of interest. In this work, we choose the Markov Chain Monte Carlo (MCMC) method to sample from the posterior distributions (Gelman et al. 2013). We introduced the MCMC methods in Chapter 2.

Firstly, the scale parameter for the regression coefficients, σ^2_{kj}, is analytically integrated out to simplify the inference (see (A.14) in the supplementary material). The quantities of interest are the allocation vector $\mathbf{z}$, regression coefficient vector for each mixture component, $\boldsymbol{\beta}_k$, unnormalised mixture weights priors at the centroids of the blocks, $f_{k,b}$, and its hyper-parameters. For brevity, let $\boldsymbol{\beta}$ be a $K \times J$ matrix of regression coefficients for all mixture components and each covariate, $\boldsymbol{X}$ be an $N \times J$ matrix of all covariates for each location, $\boldsymbol{F}$ be a $B \times K$ matrix such that $\boldsymbol{F}_{b,k} = f_{k,b}$, and $\boldsymbol{\theta}$ the vector of kernel hyperparameters for all f_k's. The unnormalised joint posterior probability distribution is given as

$$p(\boldsymbol{\beta}, \mathbf{z}, \boldsymbol{F}, \boldsymbol{\theta} | \mathbf{y}, \boldsymbol{X}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{X}, \mathbf{z}) p(\mathbf{z}|\boldsymbol{F}) p(\boldsymbol{F}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\beta}) \tag{4.1}$$

We employ the Metropolis-within-Gibbs scheme (see Chapter 2 and Geman & Geman (1984), Metropolis et al. (1953)) and sample from the posterior in three steps:

1. We sample the regression coefficients $\beta_{k,j}$ jointly for all $k = 1, \ldots, K$ and $j = 1, \ldots, J$. The unnormalised density of the conditional distribution is given as

$$p(\boldsymbol{\beta}|\boldsymbol{X}, \mathbf{y}, \mathbf{z}) \propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{X}, \mathbf{z}) p(\boldsymbol{\beta}). \tag{4.2}$$

   Equation (4.2) is sampled using the Hamiltonian Monte Carlo method (see Section 2.2.3.3 and Duane et al. (1987)), for which efficient sampling schemes are available, *e.g.*, Girolami & Calderhead (2011).

2. Mixture allocation is sampled cell by cell directly using the following equation

$$p(z_n = k | \mathbf{z}^{\bar{n}}, \alpha, \boldsymbol{X}_n, \boldsymbol{\beta}, \mathbf{y}, \boldsymbol{F}) \propto p(y_n | z_n = k, \boldsymbol{X}_n, \boldsymbol{\beta}_k) \frac{\exp(f_{k,b[n]})}{\sum_{l=1}^{K} \exp(f_{l,b[n]})}, \tag{4.3}$$

   where $\mathbf{z}^{\bar{n}}$ are the components of $\mathbf{z}$ without z_n.

3. We sample all K functions with the GP prior and their hyperparameters jointly using the Hamiltonian Monte Carlo. The joint posterior density is proportional to the expression below

$$p(\boldsymbol{F}, \boldsymbol{\theta} | \mathbf{y}, \mathbf{z}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\frac{\exp(f_{k,b[n]})}{\sum_{l=1}^{K} \exp(f_{l,b[n]})} \right)^{I(z_n = k)} \prod_{k=1}^{K} p(f_k | \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k), \tag{4.4}$$

   where $I(\cdot)$ is the indicator function.

For the full expansion of the conditional distributions in equations (4.2), (4.3), and (4.4), see Section A.3 in the online supplementary material.

In terms of computational tractability, (4.2) takes $\mathcal{O}(N + J)$ steps, (4.3) requires $\mathcal{O}(N \times K)$ steps, and (4.4) requires $\mathcal{O}\big(B^3 \times K\big)$ steps due to matrix inversions of size $B \times B$ for each of the K components. To contrast it with a standard model for spatial point patterns, one sample from a log-Gaussian Cox process involves matrix inversions that require $\mathcal{O}\big(N^3\big)$ steps (Diggle et al. 2013). Thanks to our blocking, the inference requires inversions of smaller matrices. Note that other approximation methods of Gaussian Processes such as Gaussian Markov Random Fields (Lindgren et al. 2011), or low-rank GPs (Rasmussen & Williams 2006, Chapter 8) may be utilised to reduce the standard computational cost of $\mathcal{O}\big(N^3\big)$ mentioned above.

4.2.3 Special Case: Independent Blocks

The model and the associated inference introduced in this section provide a very flexible framework for modelling the spatial dependence of cells via blocks that are also spatially dependent. However, this comes at a high cost – inferring posterior distribution over K Gaussian processes that are combined using the logistic function is challenging at scale as each sample requires $\mathcal{O}\big(B^3 \times K\big)$ operations.

If we assume that the mixture weights priors ($\boldsymbol{\pi}_b$) for all blocks are independent and, conditioned on α, distributed as

$$\boldsymbol{\pi}_b|\alpha \sim \text{Dirichlet}(\alpha, \dots, \alpha), \tag{4.5}$$

the inference becomes more tractable. Specifically, (4.4) is not needed anymore, (4.2) stays the same, and (4.3) is replaced by

$$p(z_n = k|\mathbf{z}^{\bar{n}}, \alpha, \boldsymbol{X}_n\boldsymbol{\beta}, \mathbf{y}) \propto p(y_n|z_n = k, \boldsymbol{X}_n\boldsymbol{\beta}_k)\frac{c^{\bar{n}}_{b[n]k} + \alpha}{K\alpha + \sum_{i=1}^{K} c^{\bar{n}}_{b[n]k}}. \tag{4.6}$$

As a result, the time complexity to take one sample from the unknown quantities is dominated by resampling z_n's in (4.6), which can be computed in $\mathcal{O}(N \times K)$ steps. For the full derivation of (4.6), see Section A.3.2.4 in the supplementary material.

In the literature, $\alpha = 1/K$ is a recommended choice, see, *e.g.*, Alvares et al. (2018). This prior formulation induces sparsity and is able to cancel out the components in an overfitted mixture (Rousseau & Mengersen 2011). In the experiments, we compare the trade-off between computational complexity and modelling flexibility.

4.2.4 Identifiability

Specifying a mixture model means that the model likelihood is invariant under the relabelling of the mixture components (Celeux et al. 2000). This issue is commonly referred to as lack of identifiability. In the context of SAM-GLM model, $p(\mathbf{y}|\mathbf{z}, \boldsymbol{X}, \boldsymbol{\beta})$ is invariant under the relabelling of $\boldsymbol{\beta}_k$ and f_k's, which are the component-specific model parameters.

Exploration in high dimensional spaces is, in general, hard for an MCMC sampler. As the dimension of the parameter space for the mixture model increases, the sampler is likely to explore only one of the $K!$ possible modes. For the sampler to switch to a different mode, it would have to get past the area of low probability mass surrounding the chosen mode. However, note that as the number of mixture components increases, the chance of the sampler switching to a different mode increases as the shortest distance between a pair of component-specific parameters is likely to decrease.

Since the identifiability issue poses a problem only for the interpretation of the parameters, we inspect the trace plot of the Markov chain for each identifiable parameter to assert that relabelling is not present when interpreting the mixtures.

4.3 Application: London Burglary Crime

The methodology above has been developed to enable the analysis of our application – burglary in London. We discuss the data, criminological background and how we use them to inform model selection.

4.3.1 Data Description

The data, published online by the UK police forces (police.uk 2019), are provided monthly as a spatial point pattern over the area of 1572 km^2 of both residential and non-residential burglary occurrences. The non-residential burglary refers to instances where the target is not a dwelling, *e.g.*, commercial or community properties. We discretise our study area into a regular grid by counting the number of burglary occurrences within each cell. We choose a grid for computational reasons when comparing to competing methods (see Section A.2 in the supplementary material). Given our focus on spatial modelling, we temporally aggregate the point pattern into two datasets: the one-year dataset, starting 01/2015 and ending 12/2015, with 70,234 burglaries, and the three-year dataset, starting 01/2013 and ending 12/2015, with 224,747 burglaries.

Our analysis uses land use data, socioeconomic census data from 2011, and points of interest data from 2018 to estimate their effect on the intensity of the burglary point pattern. Land use data are available as exact geometrical shapes. The census variables are measured with respect to census tracts, called output areas (OA). The OAs have been designed to have similar population sizes and be as socially homogeneous as possible, based on the tenure of households and dwelling types. Each of the 25,053 OAs in London has between 100 people or 40 households and 625 people or 250 households. The OAs are aggregated into 4,835 lower super output areas (LSOA), which in turn are aggregated into 983 middle super output areas (MSOA). An LSOA has at least 1,000 people or 400 households and at most 3,000 people or 1,200 households. For an MSOA, the minimum is 5,000 people or 2,000 households, and the maximum is 15,000 people or 6,000 households.

The points of interest data are given as a point pattern. To project the data measured at non-grid geometries (the census and land use data) onto the grid we use weighted interpolation. The method assumes that the data is uniformly distributed across the OA. For cells that have an overlap with more than one OA, we compute the value for each such cell by combining the overlapping OAs and adjusting for the size of the overlap.

4.3.2 Criminology Background

We use existing criminology studies to identify explanatory variables and formulate hypotheses about burglary target selection. The target choice is a decision-making process of maximising *reward* with minimum *effort*, and managing the *risk* of being caught (a process analogous to optimal foragers in wildlife (Johnson & Bowers 2004)). Therefore, we categorise the explanatory variables into these three categories: reward, effort, and risk.

4.3.2.1 Reward, Opportunities, Attractiveness

Theoretically supported by rational choice theory (Clarke & Cornish 1985), offenders seek to maximise their reward by choosing areas of many opportunities and attractive targets. Firstly, the *number of dwellings* is used in the literature as a measure of the abundance of residential targets (Bernasco & Nieuwbeerta 2005, Clare et al. 2009, Townsley et al. 2015, 2016). *Real estate prices* and *household income* have been used in previous works as a proxy for the attractiveness of targets. The significance of their positive effect on residential burglary victimisation rate has been mixed and varied depending on the study region and the statistical method used (Bernasco & Luykx 2003, Bernasco & Nieuwbeerta 2005, Clare et al. 2009, Townsley et al. 2015, 2016). The finding that the effect of affluence was weak in some studies can be explained by the fact that most burglars do not live in affluent areas and hence are not in their awareness spaces, i.e. operating in an affluent neighbourhood is for them an unfamiliar terrain and the risk of being caught is higher (Evans 1989, Rengert & Wasilchick 2010). Other measures of affluence that have been used include *house ownership rates* (Bernasco & Luykx 2003).

With regard to non-residential burglary, the literature is more sparse. An analysis of non-residential burglary in Merseyside county in the UK by Bowers & Hirschfield (1999) shows that non-residential facilities have a higher risk of both victimisation and repeat victimisation. In particular, sport and educational facilities have a disproportionately higher risk of being targeted compared to other types of facilities. In the crime survey of business owners in the UK, the retail sector is the most vulnerable to burglaries (gov.uk 2017). For our application, we will use points of interest database from Ordnance Survey which include retail outlets, eating and drinking venues, accommodation units, sport and entertainment facilities, and health and education institutions (Ordnance Survey 2018).

4.3.2.2 Effort, Convenience

Using the framework of crime pattern theory (Brantingham & Brantingham 1993) and routine activity theory (Cohen & Felson 1979), offenders will prefer locations that are part of their routine activities or are convenient to them, i.e. they are in their activity or awareness spaces. The studies performed using the data on detected residential burglaries unanimously agree that areas *close to the offender's home* are more likely to get targeted (Bernasco & Nieuwbeerta 2005, Townsley et al. 2015, Menting et al. 2019, Clare et al. 2009). In the study based on a survey of offenders, Menting et al. (2019) argue that other awareness spaces than their residence play a significant role in target selection. These include previous addresses, neighbourhoods of their family and friends, as well as places where they work and go about their recreation and leisure.

As confirmed by numerous studies, the spatial topology of the environment plays a significant role in the choice of a target. Brantingham & Brantingham (1975) have shown that houses in the interior of a block are less likely to get burgled. Similarly, Townsley et al. (2015), Bernasco & Nieuwbeerta (2005) showed that *single-family dwellings* are more vulnerable to burglaries than multi-family dwellings such as blocks of flats. Beavon et al. (1994) studied the effects of the street network and traffic flow on residential burglary and found that crime was higher in *more accessible* and *more frequented* areas. Similarly, Johnson & Bowers (2010) show that main street segments are more likely to become a burglary target. Clare et al. (2009), Bernasco et al. (2015) showed that the presence of connectors such as train stations increases the likelihood of being targeted, while the so-called barriers such as rivers or highways decrease it.

4.3.2.3 Risk, Likelihood of Completion

In the social disorganisation theory of crime (Shaw & McKay 1942, Sampson & Groves 1989), it is argued that social cohesion induces collective efficacy. The effect of collective efficacy on crime is twofold. First, strong social control deters those who are thinking of committing one. Second, it decreases the chance of a successful completion once an offender has chosen to do so. This theory focuses on the impact that social deprivation, economic deprivation, family disruption, ethnic heterogeneity, and residential turnover have on the crime rates within an area. Most offenders live in disadvantaged areas and often commit a crime in their awareness spaces (minimise effort). The attraction to 'prosperous targets' applies mostly to the local context (maximise gain). On the other hand, when a neighbourhood has high social cohesion (also known as 'collective efficacy'), there is mutual trust among neighbours and residents are more likely to intervene on behalf of the common good (Sampson et al. 1997).

In the context of residential burglary, *ethnic diversity* has been shown to be positively related to burglary rates (Sampson & Groves 1989, Bernasco & Nieuwbeerta 2005, Bernasco & Luykx 2003, Clare et al. 2009). *Residential turnover* is another measure of collective efficacy. Although Bernasco & Luykx (2003) document a positive relationship between residential turnover and the burglary rates, results in Bernasco & Nieuwbeerta (2005), Townsley et al. (2015) do not confirm that hypothesis. *Socio-economic variation* among residents has been shown to be positively related to general crime rates (Sampson et al. 1997, Johnson & Summers 2015), but it was either not considered or shown insignificant in the studies on burglary we have reviewed. Other indicators of social disorganisation and their effect on general crime rates (not only burglary) are the high rate of single-parent households, one-person households as well as younger households Bernasco (2014), Sampson et al. (1997), Andresen (2010).

4.3.3 Covariates Selection

Based on the criminological overview above and the availability of covariates, we form four model specifications, from very rich representations to sparse ones. Table 4.1 shows the covariates used in each of the specifications.

Variables that represent density, i.e. given by the count per cell, are log-transformed to improve the fit. For the same reason, mean household income and mean house price are in log form. Indicators of heterogeneity are computed using the index of variation introduced in Agresti & Agresti (1978). These include ethnic heterogeneity and occupation variation within an area. Both are indicators of the lack of social cohesion. Subsequently, all variables were standardised to have zero mean and standard deviation of one.

The first specification, *specification 1*, is the richest representation and includes variables that are a proxy for the same phenomenon. For example, both household income and house price are a measure of affluence. This choice is deliberate as we use a shrinkage prior for the regression coefficients to choose the most relevant variables.

The second specification, *specification 2*, removes covariates that are strongly correlated to others or lack strong evidence in the criminological literature. We remove *owner-occupied dwellings* for its strong correlations with the house dwellings and the fraction of houses that are detached or semi-detached. We remove *house dwellings* due to high correlation with (semi-)detached houses and stronger theoretical backing for the latter (Bernasco & Nieuwbeerta 2005). We remove the *urbanisation level* because of little empirical evidence found in the literature. Naturally, it acts as a proxy for where buildings are, which is accounted for to a large extent by households and points of interest variables. We remove *single-parent households* due to a high correlation with social housing and unemployment rate, and the latter two being preferable indicators of social disorganisation.

In the third specification, *specification 3*, we exclude the following variables on top of those excluded in specification 2. *Median age*, as a proxy for collective efficacy, is removed due to weak evidence in previous studies and other measures of collective efficacy already present: ethnic and socio-economic heterogeneity. *One-person households* and *accommodation POIs* are removed because of weak empirical evidence from previous studies. *Mean household income* is removed due to insufficient evidence from previous studies and an already present and more preferable measure of affluence – house price. *Social housing* variable is removed because of weak empirical evidence and a high correlation with unemployment.

In the last specification, *specification 4*, we additionally remove *unemployment rate* due to weak empirical support from previous studies. This specification aggregates all POIs

Table 4.1: Models specifications that are used throughout the evaluation of the proposed model.

	1	2	3	4
log households (count per cell)	●	●	●	●
log retail POIs (count per cell)	●	●	●	
log eating/drinking POIs (count per cell)	●	●	●	
log edu/health POIs (count per cell)	●	●	●	
log accommodation POIs (count per cell)	●	●		
log sport/entertainment POIs (count per cell)	●	●	●	
log POIs (all categories count per cell)				●
houses (fraction of dwellings)	●			
(semi-)detached houses (fraction of dwellings)	●	●	●	●
social housing (fraction of dwellings)	●	●		
owner-occupied dwelling (fraction of dwellings)	●			
single-parent households (fraction of households)	●			
one-person households (fraction of households)	●	●		
unemployment rate	●	●	●	
ethnic heterogeneity measure (index of variation)	●	●	●	●
occupation variation measure (index of variation)	●	●	●	●
accessibility (estimated by Transport for London)	●	●	●	●
residential turnover (ratio of residents who moved in/out)	●	●	●	●
median age	●	●		
log mean household income	●	●		
log mean house price	●	●	●	●
urbanisation index (proportion of urban area)	●			

into a single variable (including accommodation POIs). This is to remove the strong correlations between them. As a single variable, it signifies the level of social activity: retail, education, entertainment, etc.

4.4 Results

After discussing the modelling choices and experimental settings, we compare SAM-GLM model to the log-Gaussian Cox process (LGCP), based on the out-of-sample generalisation and crime hotspot prediction. For LGCP, we use the standard formulation with a Matèrn covariance function (see Section A.2 in the online supplementary material for full details). Lastly, we interpret the results obtained using the proposed method and show the relevance for obtaining criminological insights.

4.4.1 Evaluation and Interpretation

We give the details of the evaluation criteria for assessing the predictive power of the model and discuss how we interpret the model components.

4.4.1.1 Out-of-sample Performance

Firstly, we evaluate the performance of the proposed and competing models using the Poisson likelihood of one-period-ahead data given the model parameters obtained from training data. The likelihood denotes how likely the observed data are for given parameters. For a given sample from the posterior distribution of the model parameters, $\boldsymbol{\phi}^{(s)}$, the average pointwise *held-out log-likelihood* is defined as

$$\text{Held-out log likelihood}(s) = \frac{1}{N}\sum_{n=1}^{N} \log p(\tilde{y}_n|\boldsymbol{\phi}^{(s)}), \tag{4.7}$$

where $p(\cdot)$ is the Poisson density function with the intensity parameter being a function of $\boldsymbol{\phi}^{(s)}$ determined by the model we use (LGCP or a SAM-GLM formulation), $\tilde{y}_n$ is the realised next-period value for cell n. Log-likelihood is a relative measure used for model comparison and can only be used to compare models within the same family of models, in our case, Poisson-based models. A higher value indicates superior predictive power.

Next, we use the *root mean square error* (RMSE) metric. It is independent of the model and is measured at the same scale as the target variable. Given a sample from the posterior distribution of the model parameters, $\boldsymbol{\phi}^{(s)}$, we obtain the expected count of events for all N locations, $\bar{\mathbf{y}}^{(s)}$, which is simply the value of the cell-specific intensity parameter. Then, using the realised next-period value, $\tilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_N)$, the RMSE is defined as

$$\text{RMSE}(s) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\bar{y}_n^{(s)} - \tilde{y}_n)^2}. \tag{4.8}$$

A lower value of RMSE indicates a better predictive performance.

4.4.1.2 Hotspot Prediction

Given that burglary is our object of interest, we also evaluate models with respect to their ability to effectively model areas of high intensity, so-called *hotspots*. The predictive accuracy index (PAI) and predictive efficiency index (PEI) are two standard approaches in criminology for assessing the ability to predict crime hotspots. These measures are driven by the objective of identifying where to deploy limited resources (police officers) to capture as many crime events as possible.

For a given area size a that a model is allowed to flag as hotspots, PAI (Chainey et al. 2008) is defined as

$$\text{PAI} = \frac{c_a/C}{a/A},$$

where c_a is the number of crimes in areas (of total size a) where crimes are predicted to occur by the model, A is the total area of the study region, and C is the total number of crimes in the study region. The objective of this index is to assess the ability to capture as many crime instances as possible with the as little area as possible.

PEI measures how effective the given model's forecasts are compared to those of a perfect model, which exactly predicts all events and their locations (Hunt 2016). For a given size of the area to be marked as hotspots, a, it is defined as

$$\text{PEI} = \frac{c_a}{c_a^*},$$

where c_a is defined the same as above, and c_a^* is the number of crimes in areas (of total size a) where crimes are predicted to occur by the *perfect* model. The c_a^* term can be thought of as the maximum possible number of crimes that could have been captured using an area of size a. PEI measures performance relating to what would the perfect model predict, whereas PAI focuses on what proportion of the overall crime the model can 'cover' with a finite area a.

In our context of a regular grid, we use both measures to compare competing models when up to n cells are flagged as hotspots. More concretely, for n cells to be flagged, c_a corresponds to the number of crimes in n cells (with total area a) with the highest *predicted* count of events; c_n^* corresponds to the number of crimes in n cells (with total area a) with the highest *realised* count of events. For a given n, a higher value of PAI or PEI indicates superior hotspot prediction.

4.4.1.3 Interpretation of Results

Estimating the effects of different spatial covariates helps us understand the underlying mechanisms of the point pattern.

In the mixtures of regressions literature, the interpretation of the individual regression coefficients is of no interest, or the focus is on reporting the regression coefficients ($\boldsymbol{\beta}_k$) for each component and quantifying their uncertainty so that their significance can be assessed (Frühwirth-Schnatter et al. 2019, ch. 8). To further interpret the coefficients, one could look at each mixture component specifically and interpret the coefficients in a classical way, conditional on the partitioning of observations. For example, for a GLM with the exponential link function, increasing a covariate by 1 unit multiplies the mean value of the observed variable by the exponential of the regression coefficient for that covariate, provided other covariates are held constant. However, this approach only allows component-specific conclusions as it depends on the distribution of the covariate for the associated component. For example, one mixture component may be active in areas with very small values for a specific covariate, while some other component is active in areas with high values. Comparing regression coefficients for that covariate across different components would not be appropriate.

Instead, to be able to compare the covariates across mixture components, we derive a covariate importance measure (`IMP`) that is motivated by the coefficient of determination, R^2. The objective of this measure is to assess the magnitude and the sign (positive/negative) of the effect of a covariate for a specific mixture component on the data fit. We measure the magnitude of the effect for a covariate j of the mixture component k as the ratio of the sum of squared residuals for the full model and the sum of squared residuals for the same model without covariate j, which is then subtracted from one. For a component k and a covariate j,

$$\texttt{IMP}_{kj} = 1 - \frac{\sum_n I(z_n = k)(y_n - \hat{y}_{n\tilde{\boldsymbol{\beta}}})^2}{\sum_n I(z_n = k)(y_n - \hat{y}_{n\tilde{\boldsymbol{\beta}}^j})^2}, \tag{4.9}$$

where, $I(z_n = k)$ is the indicator function of whether cell n is allocated to component k, $\hat{y}_{n\tilde{\boldsymbol{\beta}}}$ is the predicted count using the full vector of inferred regression coefficients, and $\hat{y}_{n\tilde{\boldsymbol{\beta}}^j}$ is the predicted count using the regression coefficients with the jth coefficient set to zero. The magnitude of `IMP` is interpreted as a measure of the relative importance of the corresponding covariate for the model fit. A value of `IMP` closer to 1 represents that removing the corresponding covariate is more detrimental to model fit.

We determine the sign of `IMP` for a given covariate and a mixture component by inspecting the distribution of the covariate for the given component. We need to be careful with negative values as our covariates are centred around zero and standardised. To obtain the sign, we take the mean of the covariate across the cells that are allocated to the given component, and if that is positive, we take the sign of the corresponding β_{kj} estimate. Otherwise, we take the negative of the sign of the β_{kj} estimate.

4.4.2 Simulation Study Details

For the methodology developed in Section 4.2, we need to choose the grid size, blocking structure, number of mixture components (K) and model specification.

4.4.2.1 Model Choices

To choose grid size, we take into account the precision of the burglary point pattern. The published data have been anonymised by mapping exact locations to predefined (snap) points (police.uk 2018). We follow the recommendations in Tompson et al. (2015) who assess the accuracy of the anonymisation method by aggregating both the original and obfuscated data to areal counts at different resolutions and looking at the difference. They show that the aggregation at lower super output area (LSOA) level does not suffer from the bias introduced by the anonymisation process. Therefore, for our cell size, we approximately match an average-size LSOA to avoid the loss of precision caused by the anonymisation process. As a result, our grid has $N = 9824$ cells, each of which corresponds to an area of 400×400 metres.

For the blocking structure, we take advantage of the existing census output areas, that are designed to group homogeneous groups of households and people together (Office for National Statistics 2019). Given that our grid is approximately at the LSOA level, we choose MSOAs as the blocking structure. We assess the sensitivity of this choice in Section 4.4.4.

Table 4.2: Model performance comparison of two variants of the model – dependent blocks using the logistic transform of K Gaussian processes, and independent blocks with Dirichlet prior. Reported values are a mean and standard deviation obtained from MCMC samples. Blocking: MSOA, training data: burglary 2015, test data: 2016, model specification 4.

K	Held-out log-likelihood		RMSE	
	Independent	Dependent	Independent	Dependent
2	-2.607 ± 0.010	$\mathbf{-2.605 \pm 0.010^*}$	$\mathbf{4.999 \pm 0.028^*}$	5.010 ± 0.028
3	-2.598 ± 0.012	$\mathbf{-2.593 \pm 0.011^*}$	4.973 ± 0.036	$\mathbf{4.950 \pm 0.031^*}$
4	$\mathbf{-2.588 \pm 0.011^*}$	-2.606 ± 0.012	$\mathbf{4.964 \pm 0.034^*}$	4.988 ± 0.031

The number of components, K, is a crucial parameter of our model. We run our model for varying K and use the performance measures introduced above to decide on the optimal number of components. From our experience, after a certain number of components, interpretation becomes harder while performance does not significantly improve.

We choose model specification based on the four options mentioned in Section 4.3.3.

4.4.2.2 Dependence of Blocks

In Section 4.2 we have proposed two possible formulations for the prior on the mixture weights: the multinomial logit transformation of K Gaussian random fields and independent Dirichlet random variables. To assess whether assuming block dependence has a major effect on the quality of the model, we compare the out-of-sample performance for both variants of the model. For this comparison, we set the blocking scheme to MSOA, use model specification 4, and estimate the model on the burglary 2015 dataset. To fit the model with dependent blocks, we use the squared exponential kernel (Rasmussen & Williams 2006) where we choose the lengthscale parameter by optimising out-of-sample RMSE using grid search. Table 4.2 shows the mean and the standard deviation of the samples of held-out log-likelihood and RMSE for both variants of the model, and for different values of K. The bold typeface signifies which method performed better for the given K and for the given metric. The star indicates statistical significance with p-value $< 10^{-3}$ obtained from a two-sample t-test of samples of each metric for each variant of the model.

The results in Table 4.2 show that the model with dependent blocks does not consistently lead to improved performance. This indicates that block dependence structure in the burglary point pattern data that we consider is not a major effect. These findings highlight some aspects of the data structure in terms of capturing these effects and suggest that the point pattern data at a higher precision would be needed to uncover these effects, if they are present. For this reason, in the rest of the paper we only consider independent blocks with Dirichlet prior weights as described in Section 4.2.3.

4.4.2.3 Identifiability

As mentioned in Section 4.2, the traceplot of the log-likelihood can be inspected for label-switching. From our experience, the sampler would choose one of the $K!$ modes, that are a consequence of the likelihood invariance, and is unlikely to switch to another mode due to the high dimensionality of the parameter space.

4.4.3 SAM-GLM Performance

Figures 4.2 and 4.3 report performance for the 2015 and 2013-2015 datasets, respectively. On the left panels of the figures, we report the box-plot of the posterior distribution of the average held-out log-likelihood. We show the box-plot for different model specifications for both SAM-GLM with an increasing number of components (K) and LGCP models. On the right panels, we report analogous plots for the root mean square error metric (RMSE).

For the one-year dataset, SAM-GLM model matches the predictive performance of the LGCP model for $K = 2$ components on both metrics. For the three-year dataset, $K = 3$ components are enough to match the LGCP model using the held-out log-likelihood, but at least $K = 4$ components are required for RMSE. The extra components required to match the performance of LGCP could be explained by the fact that the three-year point pattern will naturally be smoother and thus easier to interpolate non-parametrically using the Gaussian random field part of LGCP. The probability distribution for both metrics and for all models are more concentrated for the three-year dataset. For the one-year dataset, it is clear that $K = 2$ or $K = 3$ is the optimal number of components. For the three-year counterpart, the range between 3 and 5 components would be an appropriate choice. For both datasets, the performance does not vary significantly for different

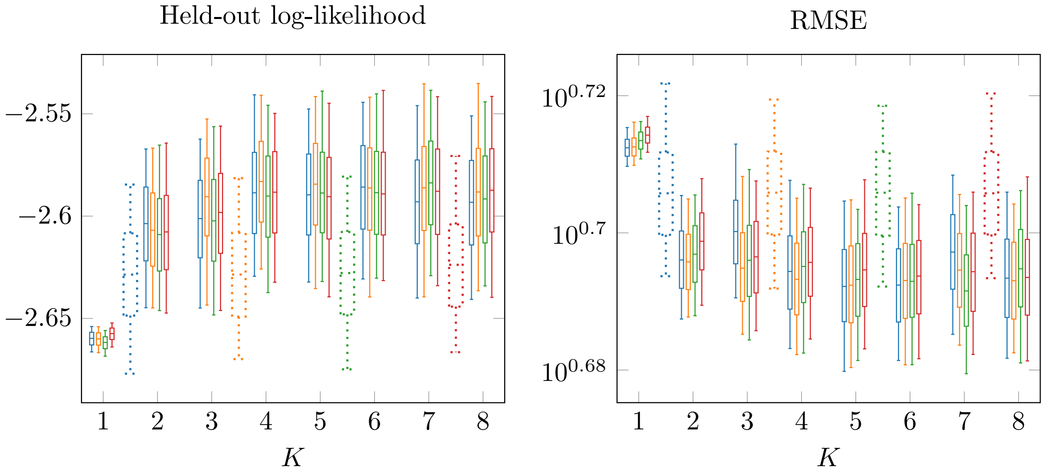


FIGURE 4.2: Evaluation of the performance of SAM-GLM (—), compared to LGCP (·····) for the one-year dataset. The box plot of the log-likelihood and root mean square error for the held-out data with respect to the posterior samples of model parameters $\phi^{(s)}$ (as described in Section 4.4.1.1) are shown for different model specifications: specification 1 (—), specification 2 (—), specification 3 (—), specification 4 (—). Blocking: MSOA, training data: burglary 2015, test data: burglary 2016. Note that the axis with the value of K does not apply to the LGCP results.

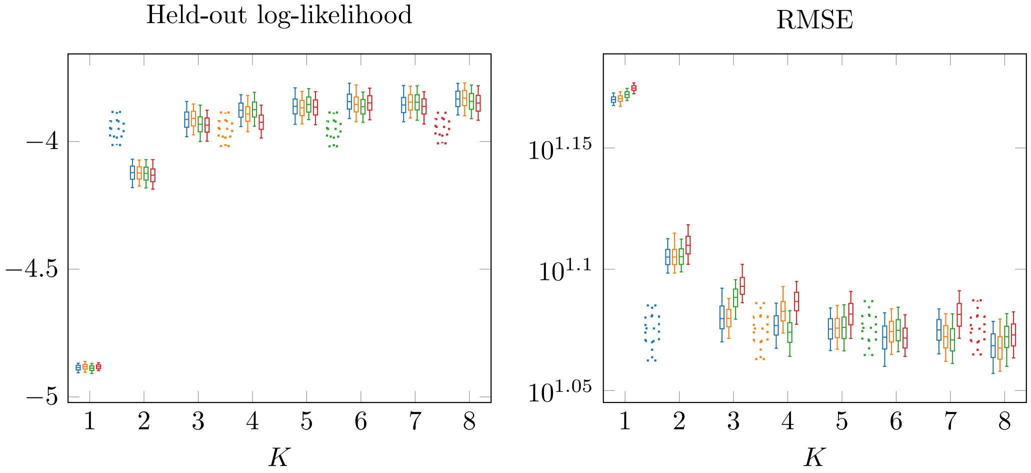


FIGURE 4.3: Evaluation of the performance of SAM-GLM (—), compared to LGCP (·····) for the three-year dataset. The box plot of the log-likelihood and root mean square error for the held-out data with respect to the posterior samples of model parameters $\phi^{(s)}$ (as described in Section 4.4.1.1) are shown for different model specifications: specification 1 (—), specification 2 (—), specification 3 (—), specification 4 (—). Blocking: MSOA, training data: burglary 2013-2015, test data: burglary 2016-2018. Note that the axis with the value of K does not apply to the LGCP results.

model specifications. Consequently, in the following sections, we limit our attention to specification 4 due to its parsimony.

While out-of-sample performance, measured by the held-out log-likelihood or RMSE, takes into account all locations, practitioners might only be interested in predicting crime hotspots. To this end, we evaluate PAI and PEI (see Section 4.4.1) as measures

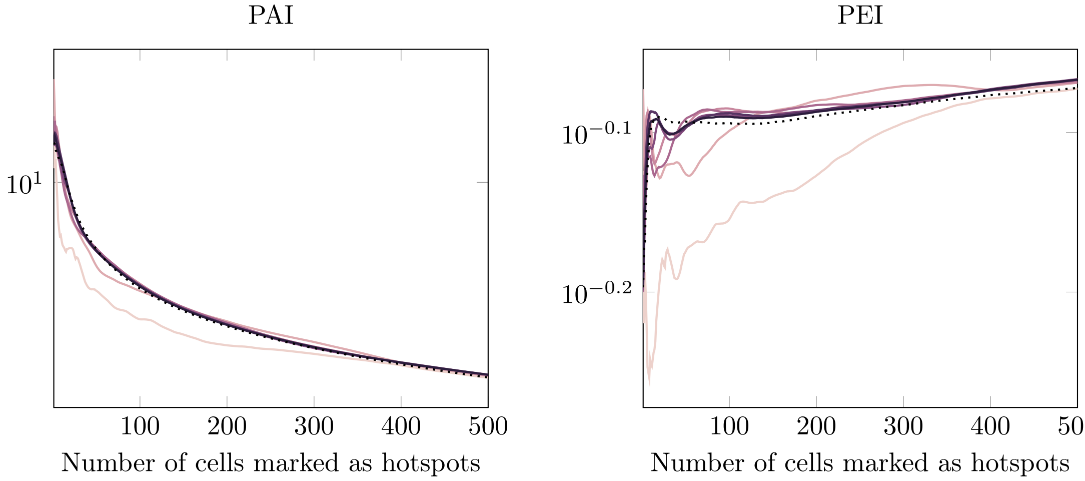


FIGURE 4.4: PAI/PEI performance SAM-GLM (—) and LGCP (·····) models, using specification 4. For the SAM-GLM results, the colour of the line represents the number of components: $K = 1$(—), $K = 2$(—), $K = 3$(—), $K = 4$(—), $K = 5$(—), $K = 6$(—), $K = 7$ (—). Blocking: MSOA, training data: burglary 2015, test data: burglary 2016, model specification: 4.

of hotspot prediction. Figures 4.4 and 4.5 show the plots of PAI and PEI measures for both models with specification 4, using the 2015 and 2013-2015 datasets, respectively. The plots show the score for when up to 500 cells (around 5% of the study region) are flagged as hotspots. Hotspots are chosen as the n cells with the highest expected value of burglaries. For the one-year dataset, the SAM-GLM model with $K = 2$ components is enough to outperform LGCP on PEI measure when between 50 and 500 cells are flagged as hotspots. For PAI measure, no significant difference can be seen for $K > 2$. The results based on the three-year data favour LGCP model when up to 150 cells are flagged as hotspots and $K < 5$. After adding more components, the SAM-GLM performance matches that of LGCP. When between 150 and 500 cells are flagged, $K \geq 3$ components is enough to outperform LGCP. These results are consistent with the previous finding that outperforming LGCP on the three-year dataset requires more components.

4.4.4 Block Size Sensitivity

The proposed model requires a specification of the blocking structure for the mixture weights prior. To assess sensitivity of this choice, we compare to local authority districts (LAD), as well as a single block for the whole study region. In the latter case, the model reduces to a non-spatial mixture of Poisson GLMs. There are 946 MSOAs, and 33 LADs

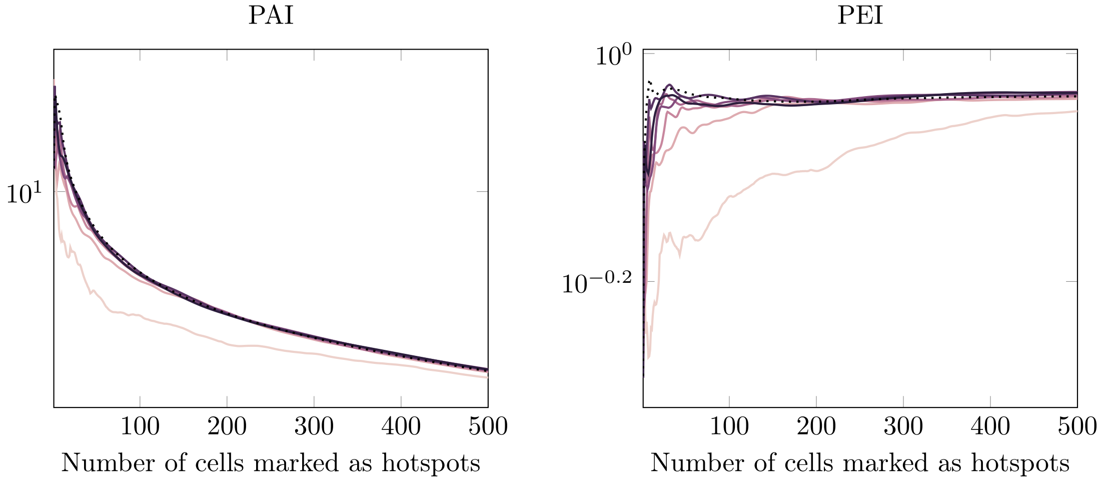


FIGURE 4.5: PAI/PEI performance SAM-GLM (—) and LGCP (·····) models, using specification 4. For the SAM-GLM results, the colour of the line represents the number of components: $K = 1$(—), $K = 2$(—), $K = 3$(—), $K = 4$(—), $K = 5$(—), $K = 6$(—), $K = 7$ (—). Blocking: MSOA, training data: burglary 2013-2015, test data: burglary 2016-2018, model specification: 4.

in the study region. The structure is hierarchical – multiple non-overlapping contiguous MSOAs constitute single LAD region.

Figures 4.6 and 4.7 show the box-plots of the held-out log-likelihood and RMSE for the one-year and the three-year datasets, respectively. The results for both metrics indicate that imposing spatial information using more localised prior results in better out-of-sample performance for the one-year dataset. To confirm that the difference is statistically significant, we performed an unpaired two-sample t-test comparing RMSE samples obtained using MSOA blocking structure to those obtained using the LAD and single blocks, respectively. Table 4.3 summarises the t-statistics and p-values. For the three-year dataset, there is no evident difference, and spatial prior does not improve predictive performance of the model. This is not surprising as the 3-year observation window will provide more information and thus the model is less likely to overfit even if we do not impose spatial dependence within the blocks.

4.4.5 Interpretation

For this analysis, we choose the three-year dataset because more data will lead to more robust inferences of the parameters. We choose specification 4 with $K = 3$ components because of its parsimony and the excellent performance shown above – for the three-year dataset and specification 4, there does not seem to be a significant improvement

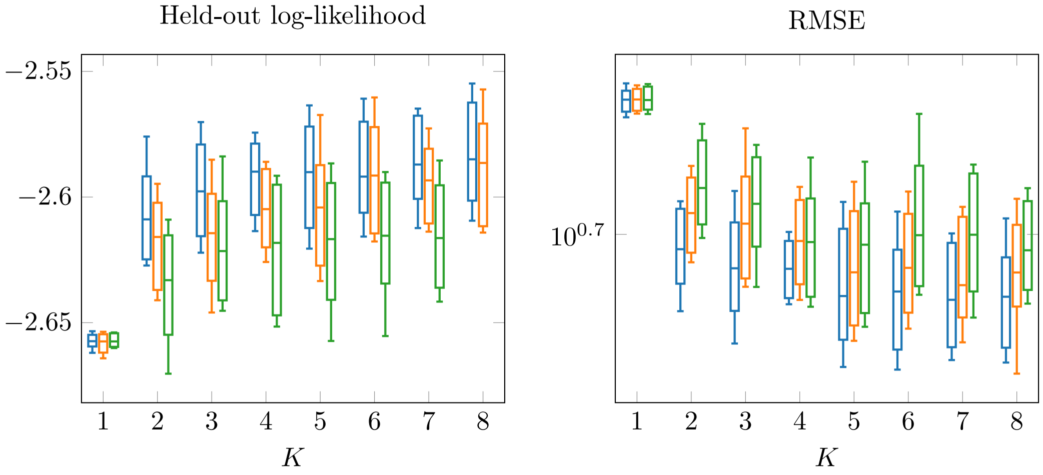


FIGURE 4.6: The box plot of the log-likelihood and root mean square error for the held-out data with respect to the posterior samples of model parameters $\phi^{(s)}$ (as described in Section 4.4.1.1) are shown for different block sizes: MSOA(—), LAD(—), single block(—). Training data: 2015, test data: 2016, model specification 4

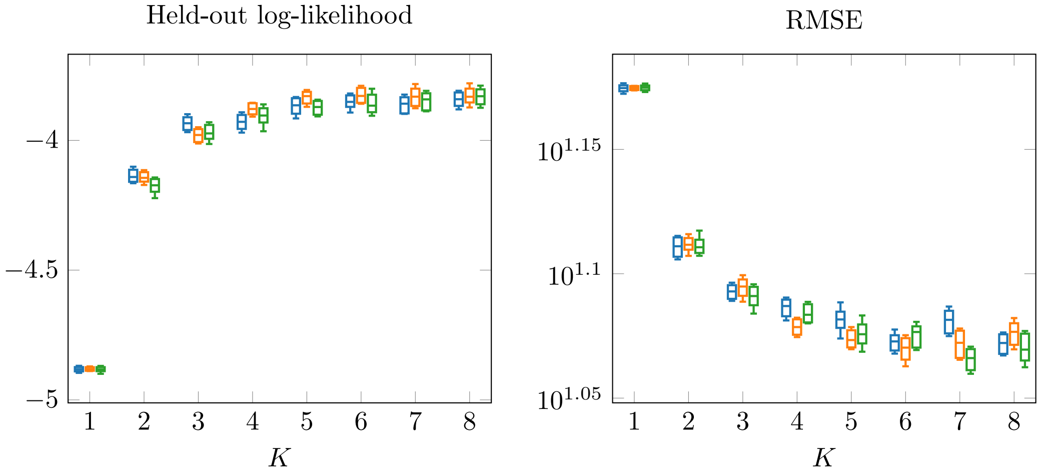


FIGURE 4.7: The box plot of the log-likelihood and root mean square error for the held-out data with respect to the posterior samples of model parameters $\phi^{(s)}$ (as described in Section 4.4.1.1) are shown for different block sizes: MSOA(—), LAD(—), single block(—). The error bars represent the standard deviation obtained from the respective MCMC samples. Training data: 2013-2015, test data: 2016-2018, model specification 4

after $K > 3$ components. Figure 4.8 shows the component allocation maps and the `IMP` measure with the effect sign $(+/-)$ for each covariate for all the three components. The allocation map for each component shows the proportion of the MCMC samples a cell is allocated to that component. The alphanumeric labels on the allocation plots are used in the discussion below when referring to specific locations. `IMP` is computed for each sample and component separately and then averaged over the MCMC samples. We also report the standard deviation of the `IMP` estimate in brackets.

Table 4.3: Sensitivity analysis of block sizes. p-values comparing whether the difference in RMSE performance is significant. Training data: burglary 2015, test data: burglary 2016, specification 4.

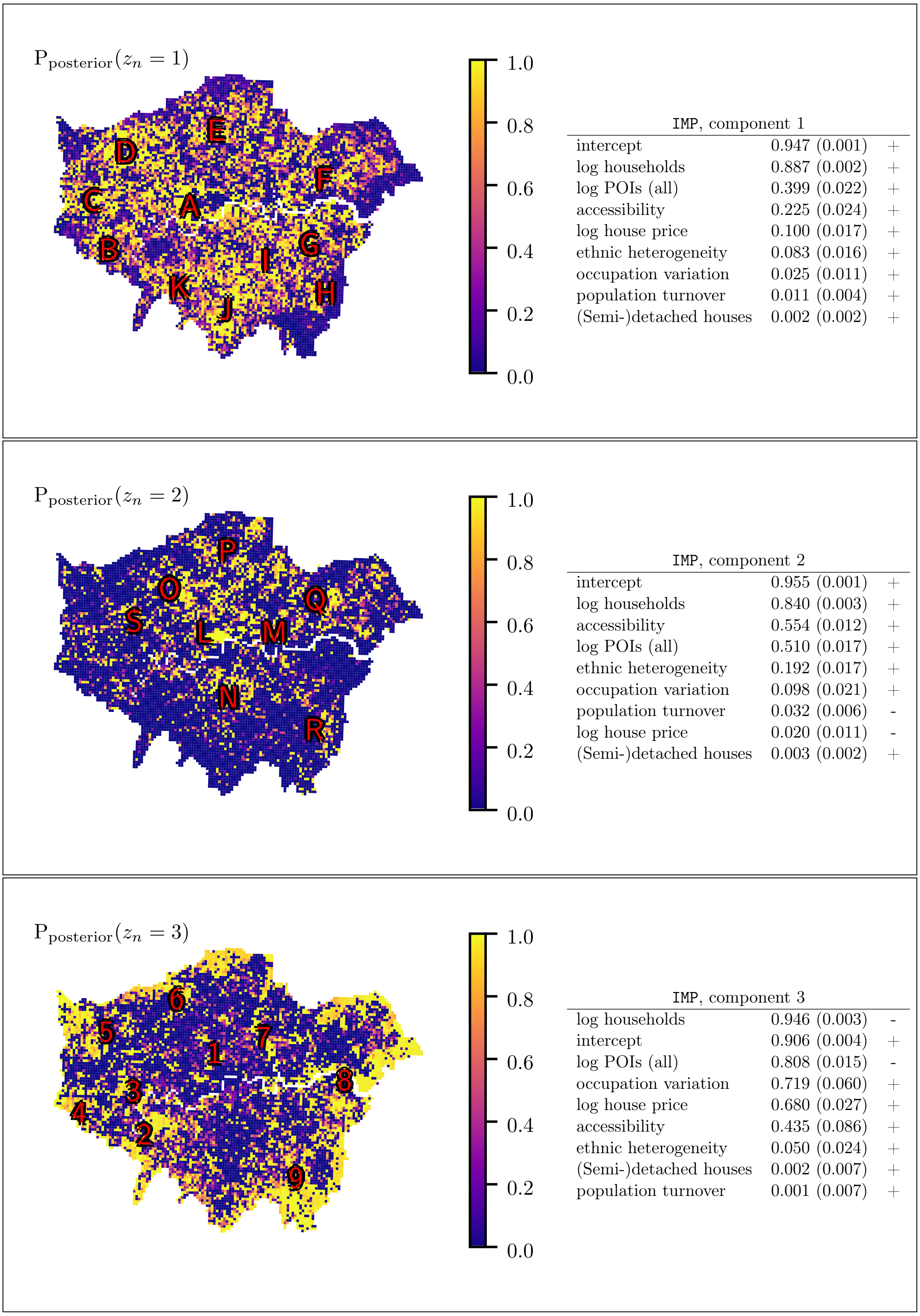
K	MSOA vs LAD		MSOA vs SINGLE	
	t-statistic	p-value	t-statistic	p-value
2	-68.732	$< 10^{-3}$	-115.042	$< 10^{-3}$
3	-76.260	$< 10^{-3}$	-87.534	$< 10^{-3}$
4	-39.016	$< 10^{-3}$	-35.207	$< 10^{-3}$
5	-26.858	$< 10^{-3}$	-52.991	$< 10^{-3}$
6	-41.913	$< 10^{-3}$	-76.152	$< 10^{-3}$
7	-12.173	$< 10^{-3}$	-56.847	$< 10^{-3}$
8	-31.547	$< 10^{-3}$	-66.688	$< 10^{-3}$

The first component is active throughout the study region, with large clusters around residential areas. These include areas around Kensington, Fulham, and Shepherd's Bush (A); Hounslow, Kingston, Richmond, and Twickenham (2); Hayes and Southall (C); Harrow and Edgware (D); East Barnet, Enfield, Walthamstow, Wood Green (E); Barking and Dagenham (F); Bexley (G); Orpington (H); Bromley (I); Croydon, and Purley (J); New Malden, and Morden (K). In this component, the number of households and points of interest have the strongest effect (excluding the intercept) – burglaries happen where targets are. Accessibility has also been inferred as an important covariate, consistent with the past criminological studies. In this component, house price is inferred as having a positive effect on the intensity of burglary, suggesting that offenders choose attractive targets. The positive effect of ethnic heterogeneity confirms the hypothesis from the social disorganisation theory. The other indicators of social disorganisation – occupation variation, residential turnover – are weaker but are consistent with the existing criminology literature. House price as a measure of reward and the proportion of houses that are detached and semi-detached have low `IMP` value.

Component 2 is active in the city centre and in the high streets of neighbourhoods: Soho, Mayfair, Covent Garden, Marylebone, Fitzrovia (L); Shoreditch and Stratford (M); Streatham and Tooting Bec (N); Wembley, and Brent (O); Enfield, Hampstead (P); Romford (Q); Orpington (R); Wembley, Harrow (S). Burglary rates in these locations are largely driven by points of interest and households. Compared to the first component (residential), the magnitudes of `IMP` values for these covariates are different - points of interest are more important for this component, and the number of households is more important for the first component. Accessibility measure is inferred to have high importance in this component. This measure is high in the city centre and around the high streets, which are usually well-connected to the public transport system. This confirms findings from crime pattern theory and routine activity theory which suggest that offenders choose locations that are part of their usual routine and in their awareness spaces. Ethnic heterogeneity and occupation variation have strong positive effect and signify the lack of social cohesion. Unexpectedly, our model infers a negative relationship between residential turnover and burglary intensity. Association of high residential turnover with the reduced risk of burglary apprehension has been shown as significant in only a few studies and was limited to *residential* burglary (Bernasco & Luykx 2003, Bernasco & Nieuwbeerta 2005, Townsley et al. 2015). Areas that are less residential such as high streets have a higher proportion of flats. Dwellings with shared premises such as flats have been shown to less likely become a target than one-household buildings (Beavon et al. 1994). Another possible reason could be the staleness of the data for the covariates which are taken from the 2011 census. Also, house price has been inferred to have a negative effect, i.e. more affluent locations are less likely to get targeted. This is contrary to the first component. A possible explanation mentioned in previous studies is that offenders often live in disadvantaged areas and choose targets within their awareness spaces, which are less likely to be affluent areas (Evans 1989, Rengert & Wasilchick 2010).

The last component is active in the areas of low intensity. These include Hyde Park, Regent's Park, Hampstead Heath (1); Richmond and Bushy parks (2); Osterley Park and Kew botanic gardens (3); Heathrow airport (4); RAF Northolt, and parks near Harrow (5); Edgware fields (6); Lee Valley (7); industrial zone in Barking and Rainham Marshes (8); parks around Bromley and Biggin Hill airport (9); and other non-urban areas located on the edges of the map. This component explains locations with little criminal activity, signified by negative `IMP` for the number of households and points of interest. Occupation variation, as a measure of socioeconomic heterogeneity, is strongly positive, which would support the hypothesis from social disorganisation theory. However, this is more likely due to the very low population in those areas which results in high occupation variation measure. Accessibility measure also has a positive effect on burglary rates in these locations. This is expected and in line with the hypotheses from the crime pattern theory. Other covariates have very small `IMP` values.

The allocation of cells partitions the map into three clusters. By aggregating the number


$\mathrm{P}_{\mathrm{posterior}}(z_n = 1)$
A
B
C
D
E
F
G
H
I
J
K
1.0
0.8
0.6
0.4
0.2
0.0
$\mathrm{P}_{\mathrm{posterior}}(z_n = 2)$
L
M
N
O
P
Q
R
S
1.0
0.8
0.6
0.4
0.2
0.0
$\mathrm{P}_{\mathrm{posterior}}(z_n = 3)$
1
2
3
4
5
6
7
8
9
1.0
0.8
0.6
0.4
0.2
0.0

IMP, component 1		
intercept	0.947 (0.001)	+
log households	0.887 (0.002)	+
log POIs (all)	0.399 (0.022)	+
accessibility	0.225 (0.024)	+
log house price	0.100 (0.017)	+
ethnic heterogeneity	0.083 (0.016)	+
occupation variation	0.025 (0.011)	+
population turnover	0.011 (0.004)	+
(Semi-)detached houses	0.002 (0.002)	+

IMP, component 2		
intercept	0.955 (0.001)	+
log households	0.840 (0.003)	+
accessibility	0.554 (0.012)	+
log POIs (all)	0.510 (0.017)	+
ethnic heterogeneity	0.192 (0.017)	+
occupation variation	0.098 (0.021)	+
population turnover	0.032 (0.006)	-
log house price	0.020 (0.011)	-
(Semi-)detached houses	0.003 (0.002)	+

IMP, component 3		
log households	0.946 (0.003)	-
intercept	0.906 (0.004)	+
log POIs (all)	0.808 (0.015)	-
occupation variation	0.719 (0.060)	+
log house price	0.680 (0.027)	+
accessibility	0.435 (0.086)	+
ethnic heterogeneity	0.050 (0.024)	+
(Semi-)detached houses	0.002 (0.007)	+
population turnover	0.001 (0.007)	+

FIGURE 4.8: Mixture model, allocations and IMP table for each mixture component. Training data: 2013-2015, specification 4.

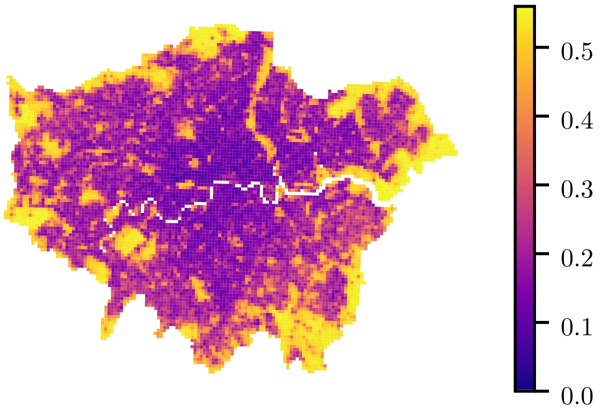
of observed crimes that occurred in each cluster we get that components 1, 2, and 3 cover 46%, 42%, 12% of all burglaries during the 2013-2015 period, respectively. Official aggregated police data for this period make the split of 64% and 36% for residential and non-residential burglary (police.uk 2019). Our inference agrees that there is more residential burglary than non-residential burglary and that approximately 35-45% of burglaries are non-residential. It is unclear whether the crime in low-count areas, which according to our model accounts for 12%, is residential or non-residential.

The support for the presence of spatial heterogeneity is further given by inspecting the inferences made by the LGCP model (for LGCP details see Section A.2 in the supplementary material). The left panel of Figure 4.9 shows standard deviations of the marginal posterior distributions of the Gaussian random field component (f). It is clear that the variance of the field component is clustered, where the regions with higher values are easily identifiable as those less urbanised. In contrast, SAM-GLM model has pickled up this heterogeneity by allowing a separate component for it (see component 3 in Figure 4.8). The right panel of Figure 4.9 shows `IMP` computed for all components of the LGCP model. `IMP` measure for the field component of the model is computed by treating it as a covariate with the coefficient equal to one. The `IMP` value for the latent field component is the third-highest, after the intercept and the number of households. A large contribution from the latent component indicates that the linear term in the Poisson regression model cannot on its own sufficiently explain the variation in the intensity of burglary.

4.4.6 Overdispersion, Excess of Zeros

The discussion of the inferences above shows that our model effectively handles excess of zeros by allocating low-count cells (non-urban areas) its own cluster, which has its own regression coefficients. Similarly, the proposed mixture model is able to reduce the overdispersion problem that is present in the standard Poisson GLM model (the special case of SAM-GLM, with $K = 1$). The mixture model may allocate each cell to a cluster that better describes the burglary count in that location. Inspecting the Pearson χ^2 statistic ($\chi^2 = \sum_{i=1}^{N} \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$) provides supporting evidence for this. Introduction of two extra components has resulted in the 81% decrease, from 106 942.43 to 20 028.99, showing a better model fit. This is further confirmed by a scatter

Standard deviation of the posterior of **f** (LGCP)



`IMP`, LGCP		
intercept	0.953 (0.001)	+
log households	0.937 (0.002)	+
field	0.810 (0.005)	+
log POIs (all)	0.728 (0.013)	+
accessibility	0.103 (0.058)	+
occupation variation	0.097 (0.052)	+
log house price	0.089 (0.043)	+
ethnic heterogeneity	0.061 (0.029)	+
population turnover	0.002 (0.027)	+
(Semi-)detached houses	0.000 (0.027)	-

FIGURE 4.9: Left: Standard deviation of the posterior distribution of the latent field, **f**, of the LGCP model. It is clear that, it is clustered and the elevated levels correspond to non-urban locations, airports, and parks (see the discussion above). Right: `IMP` measure for the component of the LGCP model. For both panels, training data: 2013-2015, model specification: 4.

plot of expected vs observed counts for the Poisson GLM model and the proposed model with $K = 3$ as shown in Figure A.2 in the supplementary material.

4.5 Conclusions

Spatial point patterns on large spatial regions, such as metropolitan areas, often exhibit localised behaviour. Motivated by this, we propose a mixture model that accounts for spatial heterogeneity as well as incorporates spatial dependence. Each component of the mixture is a model in itself, and thus allows for different locations to follow a different model, *e.g.*, in the urban context, less-urbanised locations can assume a different model from the city centre. Each component is an instance of the generalised linear model (GLM) which includes covariates. We account for spatial dependence through the mixture allocation part. The allocation of each location to one of the components is informed by both the data and the prior information. By utilising existing blocks structure, or defining a custom one, the prior supports locations within the same block to come from the same component. This formulation attempts to find the right balance between the ability to model sharp spatial variations and borrowing statistical strength for locations within the same block. Additionally, the model allows for spatial dependence between the blocks. Following the Bayesian framework, we present a Markov Chain Monte Carlo sampler to infer the posterior distributions. Inspection of the posterior distributions of the model parameters allows us to learn new insights about the underlying mechanisms of the point pattern.

Our results show that London burglary data are effectively modelled by the proposed method. Using out-of-sample and crime hotspot prediction evaluation measures, we show our model outperforms log-Gaussian Cox process (with Matèrn covariance function) that is the default model for point processes and is more computationally tractable.

The focus of this work on burglary crime does not limit the potential uses of the proposed model. We believe that the model can be applied in a wider setting of analysing spatial point patterns that may show localised behaviour and heterogeneity.

Future analysis could consider several directions not explored in this work. Firstly, our inference scheme for the model with block dependence produces an $\mathcal{O}\big(B^3 \times K\big)$ algorithm. To reduce this complexity, one could consider K level sets of a single Gaussian random field for mixture weights, instead of K Gaussian fields, thus reducing dimensionality (Hildeman et al. 2018, Fernández & Green 2002). Another approach is assuming Markovian structure of the Gaussian random fields, resulting in sparse computational methods(Rue & Held 2005). A different approach is considering inference schemes that are less computationally demanding than MCMC such as variational methods (Jordan et al. 1999). Secondly, different options for specifying the term that involves covariates could be explored. One could consider forcing certain covariates to share the coefficients across all components if there is a strong prior belief for doing so. Another possible area of investigation is spatially varying coefficient processes method, proposed by Gelfand et al. (2003).

4.6 Implementation and Supplementary Material

The source code that implements the methodology and reproduces the experiments is available at `https://github.com/jp2011/spatial-poisson-mixtures`. The supplementary material with mathematical derivations and supporting figures is available in the appendix at the end of this document.

Chapter 5

Spatial Modelling with Partial Differential Equations

This chapter gives an introduction to building models of spatial variation using partial differential equations (PDEs). Firstly, we give an overview of the models we consider in this thesis. We then proceed by discussing how one may assimilate observed data into this kind of models. Subsequently, we give a detailed and mathematically rigorous definition of what it means to solve a PDE, as part of which we elaborate on one of the most popular numerical schemes – the finite element method. Lastly, we discuss different options for introducing randomness into PDE models. This chapter, combined with Chapter 2, provides the necessary background for Chapter 6.

5.1 Models Based on Partial Differential Equations

For many phenomena in science and engineering, describing the rate of change in a quantity of interest u is more tractable or desirable than expressing the absolute value of the quantity itself. Many laws in natural sciences have been described in this manner. To name a few: Newton's laws, Hooke's law, heat equation. Partial differential equations (PDEs), which impose relations between the partial derivatives of a multivariable function, have been an invaluable tool in modelling and studying complex physical and natural systems. Such relations express how a quantity of interest, which is represented as a function, changes as we move through the domain D. In this thesis, we focus on spatial domains such that $D \subset \mathbb{R}^2$. As already outlined in Section 1.1.2.2, the general form of partial differential equations that we study is given as

$$\mathcal{L}(\kappa)u(\boldsymbol{x}) = f(\boldsymbol{x}), \tag{5.1}$$

where $\mathcal{L}$ is an elliptic operator parametrised by a *physical parameter* κ, f is the input term, and u is the solution of the given PDE. The operator $\mathcal{L}$ may be non-linear in κ and defines the relationship between the solution u and the input f.

As an illustration of how $\mathcal{L}$ may impose spatial dependence or heterogeneity, we consider the Laplace operator: $\mathcal{L} = (\nabla \cdot \nabla)$. It expresses how much the average value of u over the neighbourhood around a point $\boldsymbol{x} \in D$ differs from the value of u at $\boldsymbol{x}$. The magnitude of this change is controlled by the input f as follows:

$$\nabla \cdot \nabla u(\boldsymbol{x}) = f(\boldsymbol{x}). \tag{5.2}$$

It is evident that this formulation allows for imposing spatial correlation in u, and by letting f vary across the domain, we can incorporate spatial variability of the strength of spatial correlation.

We note that PDEs have also been used for time-dependent phenomena, however, in this thesis we assume steady state and focus only on the spatial variation of u.

5.1.1 Elliptic Partial Differential Equations

We restrict our attention to second-order linear PDEs with $\boldsymbol{x} = (x_1, x_2) \in D \subset \mathbb{R}^2$. For such class of PDEs, the equation in (5.1) can be written as

$$Au_{x_1x_1} + 2Bu_{x_1x_2} + Cu_{x_2x_2} + Du_{x_1} + Eu_{x_2} + Fu + G = 0, \tag{5.3}$$

where $u_{x_1} = \frac{\partial}{\partial x_1}$, $u_{x_1x_2} = \frac{\partial^2}{\partial x_1 \partial x_2}$. We omitted boundary conditions for brevity. Depending on A, B, C, D, E, F, G, we recognise three types of second-order linear PDEs: elliptic, parabolic, and hyperbolic. In this thesis, we further limit our focus to PDEs for which $B^2 - 4AC < 0$. This class of PDEs is called *elliptic*, inspired by the equation for an ellipse.

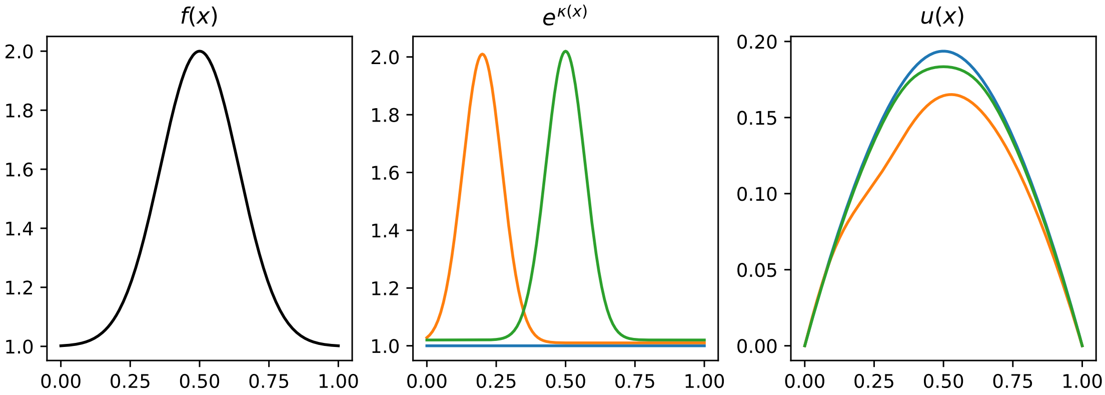


FIGURE 5.1: Solution of the Poisson equation (right) for the given forcing function $f(x)$ (left) and different diffusion coefficients $\exp(\kappa(x))$ (centre).

5.1.1.1 Running Example

As our main example, we consider the Poisson equation on domain D, as briefly introduced in Section 1.1.2.2:

$$-\nabla \cdot \big(\exp(\kappa(\boldsymbol{x}))\nabla u(\boldsymbol{x})\big) = -\sum_{j=1}^{2} \frac{\partial}{\partial x_j}\left(\exp(\kappa(x))\frac{\partial u(\boldsymbol{x})}{\partial x_j}\right) = f(\boldsymbol{x}) \tag{5.4}$$

where $\kappa(\boldsymbol{x}), f(\boldsymbol{x})\colon D \to \mathbb{R}$ are given functions, and Dirichlet boundary conditions are given as

$$u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial D. \tag{5.5}$$

For ease of notation, we set $\varkappa(\boldsymbol{x}) = \exp(\kappa(\boldsymbol{x}))$. We show the one-dimensional version of (5.4) on unit line interval with boundary conditions set as $u(0) = u(1) = 0$ in Figure 5.1. The Poisson equation on the unit line has the following form:

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(\exp(\kappa(x))\frac{\mathrm{d}u(x)}{\mathrm{d}x}\right) = f(x). \tag{5.6}$$

Figure 5.1 shows the solution $u(x)$ for a given $f(x)$ and different options of $\kappa(x)$ to illustrate how the solution changes when $\varkappa(x) = \exp(\kappa(x))$ changes.

5.1.2 Forward Problem and Inverse Problem

Once a model has been posited, a number of questions can be asked of that model. Firstly, we may use the model to determine the state of the system, u, for a given physical parameter κ and input f. For example, in structural mechanics, by knowing material properties which correspond to κ in (5.1), and Hooke's law, which corresponds to $\mathcal{L}(\kappa)$, one may predict the displacements of a material when subjected to force f. It can be used to predict what one would have observed had they done the physical experiment. Determining the state of the system u from the physical parameter κ and inputs f is termed a *forward problem*.

The second important class of problems entails determining κ from a finite number of observations of the state of the system u for a given input f. This problem is referred to as the *inverse problem*. For the structural mechanics example above, we may perform a tensile test in which we apply different force f to the material and take measurements of the resulting displacements, corresponding to the state of the system, u. We may then infer the material properties, κ, using those measurements and Hooke's law which models the relations between material properties, applied force, and displacements.

Experiments and observations play a crucial role in PDE models. Recent developments in sensor measurements and their wide availability enables study of systems that was not possible before. For example, plastic deformation in bridges (Lin et al. 2019, Febrianto et al. 2021). The obtained data is used to either calibrate the models (refining $\mathcal{L}$ to better represent the phenomena in question) or solve inverse problems, where we are interested in inferring a quantity of interest which is not directly observable (Biegler 2007, Stuart 2010). In this thesis, we restrict our attention to the inverse problem. We frame the problem in a probabilistic setting, which allows for principled uncertainty quantification. We discuss uncertainty pertaining to PDE models in the next section.

5.1.3 Uncertainty in PDE models

While a PDE is by itself a deterministic model, it is necessary that either due to the experimental nature of a task at hand, or due to observational noise, or due to the lack of knowledge, sources of uncertainty are accounted for and are appropriately quantified. As discussed in Section 2.2, uncertainty may enter models in different ways. For PDE models, uncertainty may come from the following sources:

1. *Uncertainty in the input* $f(\boldsymbol{x})$ may occur especially in an experimental setting, where input $f(\boldsymbol{x})$ is applied to a subject of interest through an actuating device. Due to noise, we may expect there to be variation in the input.

2. *Uncertainty in the parameter* $\kappa(\boldsymbol{x})$ often corresponds to the spatially varying properties of a system. For example, in a tensile test $\kappa(\boldsymbol{x})$ corresponds to material properties. Although these properties are physically deterministic (at a macroscopic level), due to the lack of knowledge about them, we may represent them using a probability measure to express our current beliefs about them (see the discussion on priors in Section 2.2).

3. *Misspecification error* refers to the mismatch between the posited model and the true data generating process. One may introduce a model component that will account for such misspecification. For example, Gaussian processes have been a common way of accounting for model misspecification (Kennedy & O'Hagan 2001, Girolami et al. 2021, Duffin et al. 2021).

4. *Observation error* is the uncertainty due to the random nature of how the data are recorded – imprecise measuring equipment, external factors such as weather impacting the readings, among others.

5.2 Solving Partial Differential Equations

The goal of this section is to discuss what it means to solve a PDE, and we make concrete the required properties of f, κ for the solution u to exist. Throughout this section, we closely follow Lord et al. (2014, ch. 2).

5.2.1 Variational Formulation

We use the running example of a 2D Poisson equation as defined in (5.4) with Dirichlet boundary conditions as given in (5.5).

If the derivatives are interpreted in the classical sense, we require that f, $\exp(\kappa)$ are continuous. The solution is then a smooth function with continuous first and second derivatives.

Definition 5.1 (classical solution)**.** Let $f \in C(\bar{D})$ and $\varkappa \in C^1(\bar{D})$. A function $u \in C^2(D) \cap C(\bar{D})$ that satisfies (5.4) for all $\boldsymbol{x} \in D$ and the boundary conditions in (5.5) is called a *classical solution*.

Note that we require that u itself has a well-defined limit on the boundary, but not necessarily its derivatives.

In practice, we want to solve problems where the choice of f is less restricted, *e.g.*, $f \in L^2(D)$[1]. In those situations, the notion of a classical solution may be too restrictive, so we interpret the derivatives in the weak sense and seek solutions in a Sobolev space.

Definition 5.2 (Weak derivative (Lord et al. 2014))**.** Consider a function $u : D \to Y$, where Y is a Banach space. Let $\mathcal{D}_j := \frac{\partial}{\partial x_j}$. Given a multi-index $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$, we define $|\boldsymbol{\alpha}| := \alpha_1 + \cdots + \alpha_d$ and $\mathcal{D}^{\alpha} := \mathcal{D}_1^{\alpha_1} \cdots \mathcal{D}_d^{\alpha_d}$, so that

$$\mathcal{D}^{\boldsymbol{\alpha}} u = \frac{\partial^{|\boldsymbol{\alpha}|} u}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$$

We say a measurable function $\mathcal{D}^{\boldsymbol{\alpha}} u : D \to \mathbb{R}$ is the $\boldsymbol{\alpha}$-th weak derivative of a measurable function $u : D \to Y$ if

$$\int_D \mathcal{D}^{\boldsymbol{\alpha}} u(\boldsymbol{x}) \phi(\boldsymbol{x}) d\boldsymbol{x} = (-1)^{|\boldsymbol{\alpha}|} \int_D u(\boldsymbol{x}) \mathcal{D}^{\boldsymbol{\alpha}} \phi(\boldsymbol{x}) d\boldsymbol{x}, \quad \forall \phi \in \mathrm{C}_\mathrm{c}^\infty(D),$$

where $C_c^\infty(D)$ is the space of infinitely differentiable functions with compact support.

Definition 5.3 (Sobolev spaces)**.** Let D be a domain and Y be a Banach space. For $p \geq 1$, the Sobolev space $W^{r,p}(D, Y)$ is the set of functions whose weak derivatives up to order $r \in \mathbb{N}$ are in $L^p(D, Y)$:

$$W^{r,p}(D, Y) := \{u : \mathcal{D}^{\boldsymbol{\alpha}} u \in L^p(D, Y) \text{ if } |\boldsymbol{\alpha}| \leq r\}. \tag{5.7}$$

If $p = 2$ and H is a Hilbert space, $H^r(D, H)$ is used to denote $W^{r,2}(D, H)$.

Proposition 5.4 (Sobolev norm and semi-norm)**.** *$W^{r,p}(D, Y)$ is a Banach space with the norm*

$$\|u\|_{W^{r,p}(D,Y)} := \left(\sum_{0 \leq |\boldsymbol{\alpha}| \leq r} \|\mathcal{D}^{\boldsymbol{\alpha}} u\|_{L^p(D,Y)}^p \right)^{1/p} \tag{5.8}$$

and $H^r(D, H)$ is a Hilbert space with inner product

$$\langle u, v \rangle_{H^r(D,H)} := \sum_{0 \leq |\boldsymbol{\alpha}| \leq r} \langle \mathcal{D}^{\boldsymbol{\alpha}} u, \mathcal{D}^{\boldsymbol{\alpha}} v \rangle_{L^2(D,H)}. \tag{5.9}$$

[1] $L^p(D)$ is the set of Borel measurable function $u \colon D \to \mathbb{R}$ with $\|u\|_{L^p(D)} < \infty$ and $|u|_{L^p(D)} := \left(\int_D |u(\boldsymbol{x})|^p \mathrm{d}\boldsymbol{x}\right)^{1/p}$.

It is also convenient to define the Sobolev semi-norm:

$$|u|_{H^r(D)} := \left(\sum_{|\boldsymbol{\alpha}|=r} \|\mathcal{D}^{\boldsymbol{\alpha}} u\|^2_{L^2(D)} \right)^{1/2}. \tag{5.10}$$

Due to boundary conditions, Sobolev spaces that incorporate boundary conditions are necessary. We therefore define a Sobolev space of functions that are zero on the Dirichlet boundary ∂D,

$$V := H^1_0(D) := \{v \in H^1 \colon v(\boldsymbol{x}) = 0, \quad \forall \boldsymbol{x} \in \partial D\}, \tag{5.11}$$

and a Sobolev space of functions on D which satisfy the boundary conditions $g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \partial D$:

$$W := H^1_g(D) := \{w \in H^1(D) \colon \gamma w(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \partial D\}, \tag{5.12}$$

where $\gamma \colon H^1(D) \to L^2(\partial D)$ is a *trace operator* that maps functions on D onto functions on the boundary ∂D. This implies that g must belong to the following subspace of $L^2(\partial D)$.

Definition 5.5 ($H^{1/2}(\partial D)$ space)**.** Let $D \subset \mathbb{R}^2$ be a bounded domain. The *trace space* is defined as

$$H^{1/2}(\partial D) := \gamma(H^1(D)) = \{\gamma w \colon w \in H^1(D)\}, \tag{5.13}$$

where γ is a trace operator on $H^1(D)$. $H^{1/2}(\partial D)$ is a Hilbert space and is equipped with the norm

$$\|g\|_{H^{1/2}(\partial D)} := \inf_{w} \left\{ \|w\|_{H^1(D)} \colon \gamma w = g \text{ and } w \in H^1(D) \right\}. \tag{5.14}$$

From now on, we will assume that the coefficients $\varkappa(\boldsymbol{x}) = \exp(\kappa(\boldsymbol{x}))$ satisfy the following assumption.

Assumption 5.6 (ellipticity condition)**.** *The diffusion coefficient* $\varkappa(\boldsymbol{x}) = \exp(\kappa(\boldsymbol{x}))$ *satisfies*

$$0 < \varkappa_{\min} \le \varkappa(\boldsymbol{x}) \le \varkappa_{\max} < \infty, \quad \textit{for almost all } \boldsymbol{x} \in D, \tag{5.15}$$

i.e., we have $\varkappa \in L^\infty(D)$.

Definition 5.7 (strong solution)**.** Let $f \in L^2(D)$. A function $u \in H^2(D) \cap H_0^1(D)$ that satisfies (5.4) for almost all $\boldsymbol{x} \in D$ with derivatives interpreted in the weak sense, and the boundary conditions in (5.5) is called a *strong solution*.

We can relax assumptions on f by considering $f \notin L^2(D)$, for example, the Dirac delta function. In those circumstances, it may still be possible to solve the PDE on separate intervals, as shown in the next one-dimensional example.

Example 5.8. *Let $D = [0,1]$, $\kappa(x) = 0$ and $f(x) = \delta(x - \frac{1}{2})$. The following $u(x)$ satisfies (5.6) on separate intervals $(0, 1/2)$ and $(1/2, 1)$:*

$$u(x) = \begin{cases} \frac{x}{2}, & 0 \leq x < \frac{1}{2}, \\ \frac{1}{2} - \frac{x}{2}, & \frac{1}{2} \leq x \leq 1. \end{cases} \tag{5.16}$$

For this solution, the weak derivative of $\mathcal{D}^1 u$ is not well-defined in $L^2(0,1)$, i.e., $u \notin H^2(0,1)$. We only have that $u \in H_0^1(0,1)$, so u is neither a classical nor a strong solution.

To overcome the limitations above, the PDE in (5.4) can be reformulated in a variational form by introducing a sufficiently smooth test function with compact support and by integrating over the domain D. This method has been introduced by Ritz (Strang & Fix 2008). For (5.4) and $\phi \in C_c^\infty(D)$, we have

$$-\int_D \nabla \cdot \big(\varkappa(\boldsymbol{x})\nabla u(\boldsymbol{x})\big)\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_D f(\boldsymbol{x})\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{5.17}$$

Applying the product rule for differentiation gives

$$\int_D \varkappa(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla\phi(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} - \int_D \nabla \cdot \big(\phi(\boldsymbol{x})\varkappa(\boldsymbol{x})\nabla u(\boldsymbol{x})\big)\mathrm{d}\boldsymbol{x} = \int_D f(\boldsymbol{x})\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{5.18}$$

and subsequently by the divergence theorem in two space dimensions we obtain

$$\int_D \varkappa(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla\phi(\boldsymbol{x}))\mathrm{d}\boldsymbol{x} - \int_{\partial D} \phi(\boldsymbol{x})\big(\varkappa(\boldsymbol{x})\nabla u(\boldsymbol{x})\big) \cdot \boldsymbol{n}\mathrm{d}s = \int_D f(\boldsymbol{x})\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \tag{5.19}$$

where $\boldsymbol{n}$ is the outwards pointing unit normal vector on ∂D and the second integral is a line integral. Since $\phi \in C_c^\infty(D)$ and $\phi(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \partial D$, we have

$$\int_D \varkappa(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_D f(\boldsymbol{x})\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{5.20}$$

Any solution to (5.4) therefore satisfies the variational problem

$$a(u, \phi) = \ell(\phi), \quad \forall \phi \in C_c^\infty(D), \tag{5.21}$$

where

$$a(u, \phi) := \int_D \varkappa(\boldsymbol{x}) \nabla u(\boldsymbol{x}) \cdot \nabla \phi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{5.22}$$

$$\ell(\phi) := \langle f, \phi \rangle_{L^2(D)}. \tag{5.23}$$

This derivation, together with the definitions of spaces V and W, allows us to write the definition of a weak solution.

Definition 5.9 (weak solution)**.** A *weak solution* to (5.4) is a function $u \in W$ that satisfies

$$a(u, v) = \ell(v), \quad \forall v \in V. \tag{5.24}$$

To show well-posedness of the variational formulation in Definition 5.9, one can show that the variational problem in (5.24) is equivalent to the following variational problem: find $u_0 \in V$ such that

$$a(u_0, v) = \ell(v) - a(u_g, v), \tag{5.25}$$

where $u_g \in H^1(D)$ such that $\gamma u_g = g$. The existence and uniqueness of $u_0 \in V$ satisfying (5.25) is a consequence of Lax-Milgram lemma.

Lemma 5.10 (Lax-Milgram lemma (Lax & Milgram 1955))**.** *Let H be a real Hilbert space with the norm $\|\cdot\|$ and let ℓ be a bounded linear functional on H. Let $a\colon H \times H \to \mathbb{R}$ be a bilinear form that is bounded and coercive ($a(x, x) \geq \beta \|x\|^2$ for all $x \in H$ for some $\beta > 0$). There exists a unique $u_\ell \in H$ such that $a(u_\ell, x) = \ell(x)$ for all $x \in H$.*

Due to the equivalence of variational problems in (5.25) and (5.24), the well-posedness of (5.24) is established through the well-posedness of (5.25) as summarised in the next theorem.

Theorem 5.11 (well-posedness of weak solution and upper bound)**.** *Let Assumption 5.6 hold, $f \in L^2(D)$ and $g \in H^{1/2}(\partial D)$. Then (5.24) has a unique solution $u \in W = H_g^1(D)$.*

Furthermore, we have

$$|u|_{H^1(D)} \leq K\big(\|f\|_{L^2(D)} + \|g\|_{H^{1/2}(\partial D)}\big), \tag{5.26}$$

where $K := \max\big\{K_p\varkappa_{\min}^{-1}, K_\gamma(1+\varkappa_{\max}\varkappa_{\min}^{-1}\big\}$. K_γ *is the bounding constant in* $\|u_g\|_{H^1(D)} \leq K_\gamma\|g\|_{H^{1/2}(\partial D)}$ *and* K_p *is the result of writing* $u = u_0 + u_g$ *with* $u_0 \in V$ *and* $u_g \in H^1(D)$ *such that* $\gamma u_g = g$*, and subsequently applying Poincaré's inequality to* u_0*:* $\|u_0\|_{L^2(D)} \leq K_p|u_0|_{H^1(D)}$*. For a proof, see Lord et al. (2014, sec. 2.2).*

5.2.2 Galerkin Approximation

Galerkin approximation technique gives appropriate methodology and error analysis for approximating spaces V and W with their finite-dimensional counterparts.

We introduce two finite-dimensional subspaces, $V^h \subset V = H_0^1(D)$ and $W^h \subset W = H_g^1(D)$, and solve (5.24) as follows.

Definition 5.12 (Galerkin approximation)**.** Let $W^h \subset W$ and $V^h \subset V$ and suppose that

$$v - w \in V^h, \quad \forall v, w \in W^h. \tag{5.27}$$

The *Galerkin approximation* for (5.4) and (5.5) is the function $u_h \in W^h$ satisfying

$$a(u_h, v) = \ell(v), \quad \forall v \in V^h. \tag{5.28}$$

Similarly to showing well-posedness of (5.24), we may apply Lax-Milgram lemma to $V^h \times V^h$ to show the existence and uniqueness of $u_h \in W^h$ as the solution of (5.28).

To quantify the approximation error between u and u_h, the result due to Céa gives the following error bound.

Theorem 5.13 (Galerkin approximation error)**.** *Suppose* $V^h \subset V$ *and* $W^h \subset W$ *and let* $u \in W$ *and* $u_h \in W^h$ *satisfy* (5.24) *and* (5.28)*, respectively. If* (5.27) *holds, then*

$$|u - u_h|_E = \inf_{w \in W^h} |u - w|_E, \tag{5.29}$$

where $|u|_E := a(u,u)^{1/2} = \big(\int_D \varkappa(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla u(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\big)^{1/2}$ *is the* energy norm*, which is equivalent to the Sobolev semi-norm* $|\cdot|_{H^1(D)}$ *(Lord et al. 2014, sec. 2.2), resulting in* *the following bound:*

$$|u - u_h|_{H^1(D)} \leq \sqrt{\frac{\varkappa_{\max}}{\varkappa_{\min}}}|u - w|_{H^1(D)} \quad \forall w \in W^h. \tag{5.30}$$

5.2.2.1 Finite Elements for Galerkin Approximation

Finite elements are a popular option for building function spaces V^h and W^h introduced above. In finite elements, $\bar{D}$ is partitioned into n_e elements, and W^h and V^h are chosen to be sets of piecewise polynomials. We limit the discussion to triangular elements due to their versatility – any domain with polygonal boundary can be partitioned with triangles. We now define what an admissible shape-regular mesh is.

Definition 5.14 (admissible shape-regular mesh)**.** Let $\mathcal{T} = \{\Delta_1, \dots \Delta_{n_e}\}$ be a set of non-overlapping triangles such that $\cup_{k=1}^{n_1} \bar{\Delta}_k = \bar{D}$. Distinct triangles must meet only at a vertex, or they must share an entire edge. Let h_k be the length of the longest edge of Δ_k and let $h := \max_k h_k$. A mesh is *shape-regular* if there exists a constant $C > 0$ independent of h such that

$$\frac{\rho_k}{h_k} \geq C, \quad \forall \Delta_k \in \mathcal{T}, \tag{5.31}$$

where ρ_k is the radius of the largest inscribed circle in Δ_k.

Given a triangular finite element mesh $\mathcal{T}_h$, we choose

$$V_h := \big\{v \in C(\bar{D}) \text{ with } v = 0 \text{ on } \partial D \text{ and } v|_{\Delta_k} \in \mathbb{P}_r(\Delta_k) \text{ for all } \Delta_k \in \mathcal{T}_h\big\}, \tag{5.32}$$

where $\mathbb{P}_r(\Delta_k)$ denotes polynomials in $\boldsymbol{x} = (x_0, x_1)$ of total degree r or less on the triangle Δ_k. We use nodal basis functions, $V^h = \text{span}\{\phi_1(\boldsymbol{x}_1), \dots, \phi_J(\boldsymbol{x}_J)\}$, where each ϕ_j is a continuous piecewise polynomial that satisfies

$$\phi_j(\boldsymbol{x}_i) = \delta_{ij}, \tag{5.33}$$

where δ_{ij} is the Kronecker delta function and $\{\boldsymbol{x}_1, \dots \boldsymbol{x}_J\}$ is a set of J nodes placed appropriately on D, depending on r. For example, for $r = 1$, the nodes are chosen to correspond to non-Dirichlet vertices of the mesh. To construct W^h, Dirichlet boundary nodes, $\boldsymbol{x}_{J+1}, \dots, \boldsymbol{x}_{J+J_b}$, need to be included, so we enrich the basis V^h with polynomials ϕ_j associated with J_b boundary nodes.

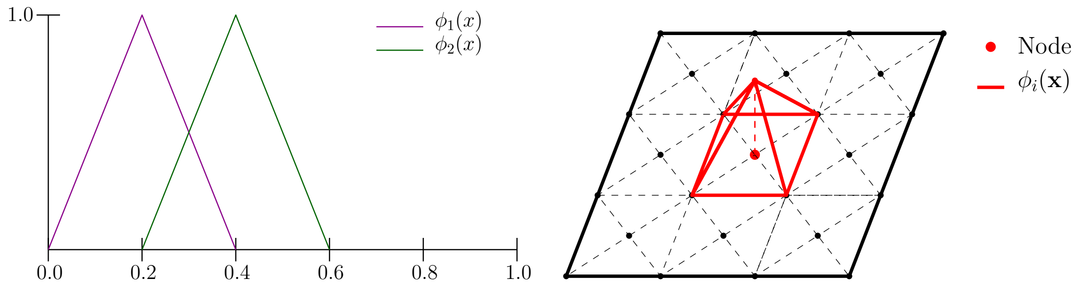


FIGURE 5.2: Examples of piecewise-linear nodal basis functions. On the left, two nodal basis functions defined on the unit line are shown. The domain is discretised into five elements. On the right, the domain is discretised using triangular elements and a single nodal basis function is shown.

In Figure 5.2, we show examples of piecewise-linear basis functions used in approximating a space of functions defined on subsets of $\mathbb{R}$ and $\mathbb{R}^2$, respectively. The domain is discretised into finite elements: triangles in 2D and line segments in 1D.

The Galerkin finite element approximation is then given as

$$u_h(\boldsymbol{x}) = \sum_{i=1}^{J} u_i \phi_i(\boldsymbol{x}) + \sum_{i=J+1}^{J+J_b} w_i \phi_i(\boldsymbol{x}) = u_0(\boldsymbol{x}) + w_g(\boldsymbol{x}), \tag{5.34}$$

where $w_i := g(\boldsymbol{x}_i)$ for $i = J+1, \ldots, J+J_b$. Substituting (5.34) into (5.25) and setting $v = \phi_j \in V^h$ gives

$$\sum_{i=1}^{J} u_i a(\phi_i, \phi_j) = \ell(\phi_j) - \sum_{i=J+1}^{J+J_b} w_i a(\phi_i, \phi_j), \quad j = 1, \ldots, J. \tag{5.35}$$

In matrix-vector notation, we gather the equations in the Galerkin matrix $\boldsymbol{A} \in \mathbb{R}^{(J+J_b)\times(J+J_b)}$ and the vector $\boldsymbol{b} \in \mathbb{R}^{J+J_b}$:

$$A_{ij} := a(\phi_i, \phi_j) \quad = \int_D \varkappa(\boldsymbol{x}) \nabla \phi_i(\boldsymbol{x}) \cdot \nabla \phi_j(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \qquad i, j = 1, \ldots, J+J_b, \tag{5.36}$$

$$b_i := \ell(\phi_i) \quad = \int_D f(\boldsymbol{x}) \phi_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \qquad i = 1, \ldots, J+J_b. \tag{5.37}$$

It is clear that a_{ij} is non-zero only when $\mathrm{supp}(\phi_i)$ and $\mathrm{supp}(\phi_j)$ intersect, implying that $\boldsymbol{A}$ is sparse. Partitioning $\boldsymbol{A}$ and $\boldsymbol{b}$ as

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_{II} & \boldsymbol{A}_{IB} \\ \boldsymbol{A}_{BI} & \boldsymbol{A}_{BB} \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} \boldsymbol{b}_I \\ \boldsymbol{b}_B \end{pmatrix}, \tag{5.38}$$

where I is the set of indices corresponding to the internal, i.e., non-Dirichlet boundary nodes, and B is the set of Dirichlet nodes indices. The Galerkin equations in (5.35) can be written as

$$\boldsymbol{A}_{II}\boldsymbol{u}_I = \boldsymbol{b}_I - \boldsymbol{A}_{IB}\boldsymbol{w}_B, \tag{5.39}$$

where $\boldsymbol{w}_B$ is the discrete boundary data.

One of the main advantages of the finite element formulation is that computation of integrals in (5.36) and (5.37) can be broken up over the elements Δ_k in $\mathcal{T}_h$, so that $\boldsymbol{A}$ and $\boldsymbol{b}$ are assembled using the element arrays $\boldsymbol{A}^k \in \mathbb{R}^{n_r \times n_r}$ and vectors $\boldsymbol{b}^k \in \mathbb{R}^{n_r}$, where n_r is the number of terms in $\mathcal{P}_r$ and also the number of basis functions. The components of $\boldsymbol{A}^k$ and $\boldsymbol{b}^k$ are defined as follows:

$$A^k_{pq} = \int_{\Delta_k} a(\boldsymbol{x})\nabla\phi^k_p(\boldsymbol{x}) \cdot \nabla\phi^k_q(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad p, q = 1, \ldots, n_r, \tag{5.40}$$

$$b^k_p = \int_{\Delta_k} f(\boldsymbol{x})\phi^k_p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad p = 1, \ldots, n_r, \tag{5.41}$$

where $\{\phi^k_1, \ldots, \phi^k_{n_r}\}$ are the *local basis functions* for Δ_k. To avoid integration over different triangular elements, one can perform integration on a reference element, by appropriately mapping from the reference element to each Δ_k, taking into consideration the Jacobian of the mapping. For more detail, see Lord et al. (2014, sec. 2.3). This ability to perform computation on a per-element basis, makes solving (5.39) amenable to parallelised implementation, and many successful software packages are available that perform this task in a parallel manner out of the box (Logg et al. 2012).

The approximation error analysis of finite elements follows from the result in (5.29), where the error energy norm[2] is bounded as

$$|u - u_h|_E \leq |u - w|_E, \quad \forall w \in W^h. \tag{5.42}$$

To show that the approximation error is of $\mathcal{O}(h)$, further regularity is assumed.

Assumption 5.15 (H^2-regularity)**.** *There exists a constant* $K_2 > 0$ *such that, for every* $f \in L^2(D)$*, the solution* u *to* (5.25) *belongs to* $H^2(D)$ *and satisfies*

$$|u|_{H^2(D)} \leq K_2 \|f\|_{L^2(D)}. \tag{5.43}$$

[2] $|u|_E := a(u, u)^{1/2} = \left(\int_D \varkappa(\boldsymbol{x})\nabla u \cdot \nabla u \mathrm{d}\boldsymbol{x}\right)^{1/2}$

Subsequently, using the equivalence of the Sobolev semi-norm and the energy norm, we have

$$|u - u_h|_E^2 \leq \varkappa_{\max} |u - w|_{H^1(D)}^2 = \varkappa_{\max} \sum_{k=1}^{n_e} |u - w|_{H^1(\Delta_k)}^2, \quad \forall w \in W^h. \tag{5.44}$$

Putting this analysis together, the following theorem holds (Lord et al. 2014, sec. 2.3).

Theorem 5.16 (finite elements error bound)**.** *Let u be the solution to* (5.25) *and let u_h be the piecewise linear finite element approximation satisfying* (5.28)*. If Assumption 5.15 holds and the finite element mesh $\mathcal{T}_h$ is shape regular, then*

$$|u - u_h|_E \leq K \sqrt{\varkappa_{\max}} h \|f\|_{L^2(D)}, \tag{5.45}$$

where $K > 0$ is a constant independent of h.

5.2.2.2 Finite Element Remarks

- The accuracy of the finite element method is increased by refining the mesh $\mathcal{T}_h$, on top of making the basis functions of V^h more complex.
- The per-element nature of assembling the Galerkin matrix $\boldsymbol{A}$ and the vector $\boldsymbol{b}$ induces sparse structures. The sparsity allows for the corresponding linear system to be solved in a parallel manner.
- Although the exposition in this thesis focuses on $D \subset \mathbb{R}^2$, complicated geometries – corresponding to real-life structures such as bridges – are amenable to finite elements discretisation and subsequent computations.

5.3 Stochastic PDEs

As discussed in Section 5.1.3, different kinds of uncertainty may be incorporated into a PDE model: uncertainty in the parameters, uncertainty in the input, misspecification error and the observation error. In this section, we discuss how uncertainty in the parameter and in the input may be accounted for in the Poisson equation. This discussion is based on Lord et al. (2014, chap. 9).

The main change compared to (5.4) is that functions $\kappa(\boldsymbol{x})$ and $f(\boldsymbol{x})$ now depend on $\omega \in \Omega$, where Ω is an abstract sample space used to define the appropriate probability space. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be such space, then the stochastic version of (5.4) requires that the following holds *almost surely* (a.s.):

$$-\nabla \cdot \big(\exp(\kappa(\boldsymbol{x}, \omega)\nabla u(\boldsymbol{x}, \omega)\big) = f(\boldsymbol{x}, \omega), \tag{5.46}$$

where Dirichlet boundary conditions may themselves depend on ω, but for simplicity we assume that they are zero a.s.:

$$u(\boldsymbol{x}, \omega) = g(\boldsymbol{x}, \omega) = 0, \quad \boldsymbol{x} \in \partial D. \tag{5.47}$$

It is assumed that $\exp(\kappa), f \in L^2(\Omega, L^2(D))$. The object of interest is to propagate the uncertainty in the parameter κ and in the input f to the solution u and any derived quantities. One may, for example, be interested in the expectation of the solution u taken over all possible values of $\omega \in \Omega$. To proceed, different approaches may be employed.

1. Study individual realisations of the solution u based on the realisations of the input $\kappa(\cdot, \omega)$. To prove the existence of the solution and finite elements error bound, Assumption 5.6 is replaced with the following assumption.

   Assumption 5.17. *For almost all $\omega \in \Omega$, realisations of $\varkappa(\cdot, \omega) = \exp(\kappa(\cdot, \omega)) \in L^\infty(D)$ and satisfy*

   $$0 < \text{ess inf}_{\boldsymbol{x} \in D} \varkappa(\boldsymbol{x}, \omega) \leq \varkappa(\boldsymbol{x}, \omega) \leq \text{ess sup}_{\boldsymbol{x} \in D} \varkappa(\boldsymbol{x}, \omega), \quad \textit{a.e. in } D. \tag{5.48}$$

   Such assumption allows for κ to be modelled as a Gaussian process, for example. To study uncertainty propagation, one typically employs sampling-based approaches, whereby for a set of samples of κ, the corresponding sample of solutions u is obtained by solving the PDE for each κ. Different summary statistics such as mean and variance may be obtained from that sample. See Lord et al. (2014, sec. 9.1) for proofs of existence of u and the finite element error analysis for Monte Carlo simulations.

2. The second approach proceeds by deriving a variational formulation on $\Omega \times D$, such that solution is sought in $W := L^2(\Omega, H^1_g(D))$ and satisfies

   $$a(u, v) = \ell(v), \quad \forall v \in V = L^2(\Omega, H^1_0(D)), \tag{5.49}$$

   where

   $$a(u, v) := \mathbb{E}\left[\int_D \varkappa(\boldsymbol{x}, \cdot)\nabla u(\boldsymbol{x}, \cdot) \cdot \nabla v(\boldsymbol{x}, \cdot)\mathrm{d}\boldsymbol{x}\right], \tag{5.50}$$

   $$\ell(v) := \mathbb{E}\left[\int_D f(\boldsymbol{x}, \cdot)v(\boldsymbol{x}, \cdot)\mathrm{d}\boldsymbol{x}\right]. \tag{5.51}$$

   Equations (5.50) and (5.51) involve integration over the abstract set Ω and probability measure $\mathbb{P}$. If instead $\kappa(\cdot, \omega)$ and $f(\cdot, \omega)$ are made to depend on M (finite) number of random variables $\{\xi_k \colon \Omega \to \Gamma_k \subset \mathbb{R}\}_{k=1}^M$, then the expectations in (5.50) and (5.51) are more tractable and can be approximated by a Galerkin method.

3. There are other approaches such as stochastic collocation method (Lord et al. 2014, sec. 9.6), but we omit the details.

In this thesis, we solely employ the first approach, where we solve a PDE with random parameter for each realisation of the physical parameter $\kappa(\boldsymbol{x})$ (further, we assume f is deterministic).

Chapter 6

Variational Bayesian Approximation of Inverse Problems using Sparse Precision Matrices

6.1 Introduction

The increased availability of measurements from engineering systems allows for the development of new and the improvement of existing computational models, which are usually formulated as partial differential equations. Inferring model parameters from observations of the physical system is termed the *inverse problem* (Tarantola 2005, Kaipio & Somersalo 2005, Stuart 2010). In this work, we consider the inverse problem where the quantities of interest (for example, some material properties) and the observations (*e.g.*, the displacement field) are related through elliptic PDEs. Most inverse problems are non-linear and ill-posed, meaning that the existence, uniqueness, and/or stability (continuous dependence on the parameters) of the solution are violated (Stuart 2010, Tarantola 2005, Kaipio & Somersalo 2005). These issues are often alleviated through some regularisation, like Tikhonov regularisation (Tikhonov & Arsenin 1977), that imposes assumptions on the regularity of the solution. Alternatively, the specification of the prior in the Bayesian formulation of inverse problems provides a natural choice for regularisation, and any given regularisation can be interpreted as a specific choice of priors in the Bayesian setting (Bishop 2006). Furthermore, the Bayesian formulation provides not only a qualitative but also a quantitative estimate of both epistemic and aleatoric uncertainty in the solution. In particular, the mean of the posterior probability distribution corresponds to the point estimate of the solution while the credible intervals capture the range of the parameters consistent with the observed measurements and prior assumptions. For these reasons, Bayesian methods have gained popularity in computational mechanics for experimental design and inverse problems with uncertainty quantification; see, *e.g.*, the recent papers by Abdulle & Garegnani (2021), Pandita et al. (2021), Pyrialakos et al. (2021), Ni et al. (2021), Sabater et al. (2021), Huang et al. (2021), Ibrahimbegovic et al. (2020), Tarakanov & Elsheikh (2020), Michelén Ströfer et al. (2020), Carlon et al. (2020), Wu et al. (2020), Uribe et al. (2020), Rizzi et al. (2019), Arnst & Soize (2019), Beck et al. (2018), Betz et al. (2018), Chen et al. (2017), Asaadi & Heyns (2017), Huang et al. (2017), Karathanasopoulos et al. (2017), Babuška et al. (2016), Girolami et al. (2021).

The Bayesian formulation of inverse problems is also the focal point of probabilistic machine learning, and in recent years significant progress has been made in adapting and scaling machine learning approaches to complex large-scale problems (Lu & Tang 2015, Solin et al. 2018). One of the leading models for Bayesian inverse problems are Gaussian processes (GPs) which define probability distributions over functions and allow for incorporating observed data to obtain posterior distributions. Given that most posterior distributions in Bayesian inference are analytically intractable, approximation methods need to be resorted to. Two classical approximation schemes are the Markov Chain Monte Carlo (MCMC) and the Laplace approximation (LA). The MCMC algorithm proceeds by creating a Markov Chain whose stationary distribution is the desired posterior distribution. Although MCMC provides asymptotic convergence in distribution, devising an efficient, finite-time sampling scheme is challenging, especially in higher dimensions (Gelman et al. 2013). Application-specific techniques such as parameter space reduction and state space reduction have been proposed in the literature to help scale up MCMC methods, but these low-rank approximations are not specific to MCMC methods only (Cui et al. 2016). Due to the asymptotic correctness of MCMC, we use it as a benchmark for the experimental studies in this thesis. Meanwhile, the Laplace approximation finds a Gaussian density centred around the mode of the true posterior, utilising the negative Hessian of the unnormalised posterior log-density (Bishop 2006). The Hessian is a large dense matrix, where forming each column requires multiple PDE solves; to make such calculations feasible, low-rank approximations are typically used (Villa et al. 2021, Bui-Thanh et al. 2013). Evidently, the Laplace approximation is not suitable for multi-modal posterior distributions due to the uni-modality of the Gaussian distribution.

6.1.1 Related Work

In recent years, advances in variational Bayes (VB) methods have allowed for Bayesian inference to be successfully applied to large data sets. Variational Bayes translates a sampling problem that arises from applying the Bayes rule into an optimisation problem (Jordan et al. 1999, Blei et al. 2017, Jordan & Wainwright 2007). The method finds a solution that minimises the Kullback-Leibler (KL) divergence between the true posterior distribution and a trial distribution from a chosen family of distributions, for instance, multivariate Gaussian distributions with a specific covariance structure. The strong appeal of VB is that one can explicitly choose the complexity of the trial distribution, i.e., its number of free parameters, such that the resulting optimisation problem is computationally tractable, and the approximate posterior adequately captures important aspects of the true posterior.

Further scalability of VB methods is due to advancements in sparse approximations and approximate inference. For instance, sparse GP methods such as Nyström approximation or fully independent training conditional method (FITC) rely on lower-dimensional representations that are defined by a smaller set of so-called inducing points to represent the full GP (Williams & Seeger 2001, Csató & Opper 2002, Seeger et al. 2003, Quiñonero-Candela & Rasmussen 2005, Snelson & Ghahramani 2006, Titsias 2009, 2008). Using this approximation for a data set of size N, algorithmic complexity is reduced from $\mathcal{O}(N^3)$ to $\mathcal{O}(NM^2)$, while storage demands go down from $\mathcal{O}(N^2)$ to $\mathcal{O}(NM)$, where M is a user selected number of inducing variables. To widen the applicability of VB to large datasets and non-conjugate models (combinations of prior distributions and likelihoods that do not result in a closed-form solution), *stochastic variational inference* (SVI) was proposed (Hensman et al. 2012, Hoffman et al. 2013, Hensman et al. 2013). Subsampling the original data and Monte Carlo estimation of the optimisation objective and its gradients, allows for calibrating complex models using large amounts of data. Multiple further extensions to the sparse SVI framework were proposed, leveraging the

Hilbert space formulation of VB (Cheng & Boots 2017), introducing parametric approximations (Jankowiak et al. 2020), applying the Lanczos algorithm to efficiently factorise the covariance matrix (Pleiss et al. 2018), transforming to an orthogonal basis (Salimbeni et al. 2018, Shi et al. 2020), and adapting to compositional models (Salimbeni & Deisenroth 2017).

The choice of prior is a central task in designing Bayesian models. If the prior is obtained from a domain expert, it is not necessarily less valuable than the data itself; one way of thinking about a prior is by considering how many observations one would be prepared to trade for a prior from an expert – if the expert is very knowledgeable, then one might be prepared to exchange a large part of a dataset to get access to that prior. Translating the expert knowledge into a prior probability distribution is a challenging task, and due to practical considerations, certain choices of priors are preferred for their simplicity and analytic tractability. When inferring values of parameters over a spatial domain, as is typically the case in finite elements, GP priors offer a natural way to incorporate the information about the smoothness and other known properties of the solution. We note that while other Bayesian models, such as Bayesian neural networks are gaining interest, it is very difficult to impose functional priors in such models, challenging the effective use of expert knowledge and leading to unrealistic uncertainty estimates (Sun et al. 2019, Burt et al. 2021).

6.1.2 Contributions

In this work, we advocate for the use of GP priors with stochastic variational inference as a principled and efficient way to solve the inverse problems arising in computational mechanics. We show, through an extensive empirical study, that variational Bayes methods provide a flexible and efficient alternative to MCMC methods in the context of Bayesian inverse problems based on elliptic PDEs while retaining the ability to quantify uncertainty. While similar directions have been explored in previous work, the focus there is on specific applications, such as parameter estimation problems in models of contamination (Tsilifis et al. 2016) or proof-of-concept on particular 1D inverse problems (Barajas-Solano & Tartakovsky 2019).

We extend the previous works in multiple aspects, focusing on improving the utility of VB in inverse problems arising from elliptic PDEs and providing a thorough discussion of the empirical results that can be used by practitioners to guide their use of VB in applications. Specifically, we argue that the efficiency of the VB algorithms for PDE based inverse problems can be improved by taking into account the structure of the problem, as encoded in the FEM discretisation of the PDE. Motivated by previous uses of precision matrices as a way of describing conditional independence (Tan & Nott 2018, Durrande et al. 2019), we leverage the sparse structure of the problems to impose conditional independence in the approximating posterior distribution. This choice of parametrisation results in sparse matrices, which improve the computational and the memory cost of the resulting algorithms. Such parametrisation, combined with stochastic optimisation techniques, allows the method to be scaled up to large problems on 2D domains. Through extensive empirical comparisons, we demonstrate that VB provides high-quality point estimates and uncertainty quantification comparable to the estimates attained by MCMC algorithms but with significant computational gains. Finally, we describe how the proposed framework can be seamlessly combined with existing solvers and optimisation algorithms in the finite element implementations.

The main concern related to VB in statistics stems from the fact that it is constrained by the chosen family of trial distributions, which may not approximate the true posterior distribution well. If the choice of the trial distributions is too restrictive, the estimate of the posterior mean is biased while the uncertainty may be underestimated (MacKay 2003, Wang & Titterington 2005, Turner & Sahani 2011). Furthermore, as noted in previous work, the commonly used mean-field factorisation of the trial distributions does not come with general guarantees on accuracy (Giordano et al. 2018). However, VB has been demonstrated to work well in practice in a variety of settings (Kingma & Welling 2014, Damianou et al. 2016, Blei et al. 2017, Zhang et al. 2019). Recent work on VB has provided some tools for assessing the robustness of the VB estimates (Giordano et al. 2018) .

6.1.3 Overview

The rest of the chapter is structured as follows. In Section 6.2, we define Bayesian inverse problems and detail some inference challenges related to their ill-posedness. In Section 6.3, we give a presentation of the variational Bayes framework, with strong focus on sparse parametrisation resulting from conditional independence. We give details of the experiments and the evaluation criteria, and discuss obtained results for each experiment in Section 6.4. Lastly, Section 6.5 concludes the chapter and discusses some promising directions for future work.

6.2 Bayesian Formulation of Inverse Problems

In this section, we review the Bayesian formulation of inverse problems by closely following Stuart (2010).

6.2.1 Forward Map and Observation Model

We are interested in finding $\kappa \in \mathcal{K}$, a model parameter, given $y \in \mathcal{Y}$, a noisy observation of the solution of the model, where $\mathcal{K}, \mathcal{Y}$ are Banach spaces[1]. The mapping, which is conditioned on model input f, is given by

$$y = \mathcal{G}(\kappa; f) + \eta, \tag{6.1}$$

where $\mathcal{G} : \mathcal{K} \to \mathcal{Y}$, $\eta \in \mathcal{Y}$ is additive observational noise. We focus on problems where $\mathcal{G}$ maps solutions of elliptic partial differential equations with parameter $\kappa \in \mathcal{K}$ and input f into the observation space $\mathcal{Y}$. For a suitable Hilbert space $\mathcal{U}$, which we make concrete later, let $\mathcal{A} \colon \mathcal{K} \to \mathcal{U}$ be a possibly non-linear solution operator of the PDE, conditioned on the input f. For a particular $\kappa \in \mathcal{K}$, the solution $u \in \mathcal{U}$ is

$$u = \mathcal{A}(\kappa; f). \tag{6.2}$$

[1] Respective norms for Banach spaces $\mathcal{K}$, $\mathcal{Y}$ are $\|\cdot\|_{\mathcal{K}}$ and $\|\cdot\|_{\mathcal{Y}}$.

To obtain observations y, we define a projection operator $\mathcal{P}\colon \mathcal{U} \to \mathcal{Y}$. Consequently, (6.1) can be written out in full as

$$y = \mathcal{P}(\mathcal{A}(\kappa; f)) + \eta. \tag{6.3}$$

6.2.2 Inference

We solve the inverse problem (6.1) for κ by finding κ such that the data misfit, $\|y - \mathcal{G}(\kappa; f)\|_{\mathcal{Y}}$, is minimised. As already mentioned in the introduction, this is typically an ill-posed problem: there may be no solution, it may not be unique, there may exist a dimensionality mismatch between the observations and the quantity being inferred, and it may depend sensitively on y. To proceed, we choose the Bayesian framework for regularising the problem to make it amenable to analysis and practical implementation. We describe our prior knowledge about κ in terms of a prior probability measure μ_0 on the subspace of $\mathcal{K}$ and use Bayes' formula to calculate the posterior probability measure, μ^y, for κ given y. The relationship between the posterior and prior is expressed as

$$\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}(\kappa) = \frac{1}{Z(y)} \exp(-\Phi(\kappa; y)), \tag{6.4}$$

where $\frac{\mathrm{d}\mu^y}{\mathrm{d}\mu_0}$ is the Radon-Nikodym derivative of μ^y with respect to μ_0, and Φ is the potential function which is determined by the forward problem (6.1), specifically $\mathcal{G}$ and η. To ensure that μ^y is a valid probability measure, we have $Z(y) = \int_{\mathcal{K}} \exp(-\Phi(\kappa; y)) \mathrm{d}\mu_0(\kappa)$.

From here on, we assume that pointwise measurements of the modelled quantity such that $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}}) = (\mathbb{R}^{n_y}, \|\cdot\|)$, where $\|\cdot\|$ is the Euclidean norm, and we treat data y and η as vectors, i.e. $\boldsymbol{y}$ and $\boldsymbol{\eta}$. We specify the additive noise vector $\boldsymbol{\eta}$ as Gaussian such that

$$\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \Gamma = \sigma_y^2 \mathbf{I}),$$

where σ_y is the standard deviation of the measurement noise and $\mathbf{I}$ is the identity matrix. We can write Φ conveniently as

$$\Phi(\kappa; \boldsymbol{y}) = \frac{1}{2} \|\mathcal{G}(\kappa; f) - \boldsymbol{y}\|^2_{\Gamma^{-1}}, \tag{6.5}$$

where $\|\cdot\|_{\Gamma^{-1}}$ is the norm induced by the weighted inner product[2].

[2] For any self-adjoint positive operator $\mathcal{T}$, weighted inner product is $\langle\cdot,\cdot\rangle_{\mathcal{T}} = \langle \mathcal{T}^{-1/2}\cdot, \mathcal{T}^{-1/2}\cdot\rangle$, and the induced norm is $\|\cdot\|_{\mathcal{T}} = \|\mathcal{T}^{-1/2}\cdot\|$

We restrict the space of solutions $\mathcal{K}$ to be a Hilbert space and place a Gaussian prior measure on κ with mean m and covariance operator $\mathcal{C}_\kappa$ such that

$$\mu_0(\kappa) \sim \mathcal{N}(m, \mathcal{C}_\kappa). \tag{6.6}$$

For detailed assumptions on μ_0, $\mathcal{G}$, and η that are required for deriving the posterior probability measure, we refer the reader to Stuart (2010, Sec. 2.4).

6.2.2.1 Algorithms

The objective is to find the posterior measure μ^y conditioned on the observations, as dictated by Bayes's rule. The forward map (6.1) and the respective functions must be discretised. In Bayesian inference there are two possible approaches for discretisation: 1) apply the Bayesian methodology first, discretise afterwards, or 2) discretise first, then apply the Bayesian methodology (Stuart 2010).

The first approach develops the solution of the inference problem in the function space before discretising it. A widely used algorithm of this form is the pre-conditioned Crank-Nicholson (pCN) MCMC scheme, where proposals are based on the prior measure μ_0 and the current state of the Markov chain. The pCN method is a standard choice for high-dimensional sampling problems, as its implementation is well-defined and is invariant to mesh refinement (Cotter et al. 2013, Pinski et al. 2015, Hairer et al. 2014). Since we will use this algorithm as one of the baselines, a summary of the algorithm is provided in Section 2.2.3.5. More recently, infinite-dimensional MCMC schemes that leverage the geometry of the posterior to improve the efficiency have been proposed, see Beskos et al. (2017), Rudolf & Sprungk (2018). Such approaches can account for anisotropy of the covariance of the posterior or the local curvature of Φ. Other than MCMC schemes, some variational Bayes formulations in function space have been proposed (for example, Minh (2017), Burt et al. (2021)), though currently they do not offer a viable computational alternative to the finite-dimensional formulation of variational inference.

The second approach proceeds by first discretising the problem and then deriving the inference method. This approach forms the basis of almost all inference procedures developed in engineering: MCMC algorithms such as Metropolis-Hastings (Metropolis et al. 1953, Hastings 1970) or Hamiltonian Monte Carlo (HMC) (Duane et al. 1987), the Laplace approximation, or variational Bayes (Jordan et al. 1998, 1999) are used to approximate the posterior. In the discretised formulation, HMC has achieved recognition as the *gold standard* for its good convergence properties, favourable performance on high-dimensional and poorly conditioned problems, and universality of implementation that enables its generic use in many applications through probabilistic programming languages (*e.g.*, Stan (Carpenter et al. 2017)). Therefore, along with the pCN scheme mentioned above, our baseline for inference methods includes the HMC method, and we provide a summary of the HMC scheme in Section 2.2.3.3.

For the rest of the exposition in this chapter, we will focus on algorithms in the finite-dimensional case, where we discretise κ to yield a vector $\boldsymbol{\kappa}$. In finite dimensions, probability densities with respect to the Lebesgue measure can be defined, thus leading to a more familiar form of the Bayes's rule:

$$p(\boldsymbol{\kappa} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\kappa})\ p(\boldsymbol{\kappa})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y} \mid \boldsymbol{\kappa})\ p(\boldsymbol{\kappa}), \tag{6.7}$$

where $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$ is the posterior density, $p(\boldsymbol{y} \mid \boldsymbol{\kappa})$ is the likelihood of the observed data $\boldsymbol{y}$ for a given discretised $\boldsymbol{\kappa}$ and is determined by the discretised forward problem (6.1) and noise $\boldsymbol{\eta}$. The prior density for $\boldsymbol{\kappa}$, which itself may depend on some (hyper-) parameters ψ, is denoted by $p(\boldsymbol{\kappa})$. Next two sections focus on discussing $p(\boldsymbol{y} \mid \boldsymbol{\kappa})$ and $p(\boldsymbol{\kappa})$, respectively.

6.2.3 Poisson Equation and Likelihood

Let us consider a specific forward problem where $u = \mathcal{A}(\kappa; f)$ is the solution to the Poisson problem:

$$-\nabla \cdot (\exp(\kappa(\boldsymbol{x}))\nabla u(\boldsymbol{x})) = f(\boldsymbol{x}), \tag{6.8}$$

where $\boldsymbol{x} \in \Omega \subset \mathbb{R}^d$, with $d \in \{1, 2, 3\}$, $\kappa(\boldsymbol{x}) \in \mathbb{R}$ is the log-diffusion coefficient, $u(\boldsymbol{x}) \in \mathbb{R}$ is the unknown, and $f(\boldsymbol{x}) \in \mathbb{R}$ is a deterministic forcing term. The boundary conditions have been omitted for brevity. We are given n_y noisy observations $\boldsymbol{y} \in \mathbb{R}^{n_y}$ of the solution u at a finite set of points, $\{\boldsymbol{x}_i\}_{i=1}^{n_y}$. The observation points are collected in the matrix $\mathbf{X} \in \mathbb{R}^{n_y \times d}$. Although this PDE is linear in u for a given κ, the methodology in this work applies to non-linear cases and can be extended for time-dependent cases such as the inverse problem of inferring initial conditions of a system given observations of the system at a later time.

We discretise the weak form of the Poisson problem (6.8) with a standard finite element approach (see Chapter 5 for background on solving PDEs). Specifically, the domain of interest Ω is subdivided into a set $\{\omega_e\}$ of non-overlapping elements of size $h = \max_e \operatorname{diam}(\omega_e)$ such that:

$$\Omega = \bigcup_{e=1}^{n_e} \omega_e \,. \tag{6.9}$$

The unknown field $u(\boldsymbol{x})$ is approximated with Lagrange basis functions $\phi_i(\boldsymbol{x})$ and the respective nodal coefficients $\mathbf{u} = (u_1, \dots, u_{n_u})^\top$ of the n_u non-Dirichlet boundary mesh nodes by

$$u_h(\boldsymbol{x}) = \sum_{i=1}^{n_u} \phi_i(\boldsymbol{x}) \mathbf{u}_i \,. \tag{6.10}$$

The discretisation of the weak form of the Poisson equation yields the linear system of equations

$$\mathbf{A}(\boldsymbol{\kappa})\mathbf{u} = \mathbf{f} \,, \tag{6.11}$$

where $\mathbf{A}(\boldsymbol{\kappa}) \in \mathbb{R}^{n_u \times n_u}$ is the stiffness matrix, $\boldsymbol{\kappa} \in \mathbb{R}^{n_{\boldsymbol{\kappa}}}$ is the vector of log-diffusion coefficients, $\mathbf{f} \in \mathbb{R}^{n_u}$ is the nodal source vector. The stiffness matrix of an element with label e is given by

$$A^e_{ij}(\kappa_e) = \int_{\omega_e} \exp(\kappa_e) \frac{\partial \phi_i(\boldsymbol{x})}{\partial \boldsymbol{x}} \cdot \frac{\partial \phi_j(\boldsymbol{x})}{\partial \boldsymbol{x}} \mathrm{d}\boldsymbol{x} \,, \tag{6.12}$$

where the log-diffusion coefficient κ_e of the element is assumed to be *constant* within the element. The source vector is discretised as:

$$f_i = \int_\Omega f(\boldsymbol{x}) \phi_i(\boldsymbol{x}) \mathrm{d}x \,. \tag{6.13}$$

Hence, according to the observation model (6.5) the likelihood is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\kappa}) = p(\boldsymbol{y} \mid \mathbf{u}(\boldsymbol{\kappa})) = \mathcal{N}(\mathbf{P}\mathbf{A}(\boldsymbol{\kappa})^{-1}\mathbf{f}, \sigma_y^2 \mathbf{I}) \,, \tag{6.14}$$

where the matrix $\mathbf{P}$ represents the discretisation of the observation operator $\mathcal{P}$.

The mapping from the coefficients $\boldsymbol{\kappa}$ to the solution $\mathbf{u}$ is $\mathbf{u}(\boldsymbol{\kappa}) = \mathbf{A}(\boldsymbol{\kappa})^{-1}\mathbf{f}$. The unconditional distribution of $\mathbf{u}$ is given by:

$$p(\mathbf{u}) = \int p(\mathbf{u} \mid \boldsymbol{\kappa}) p(\boldsymbol{\kappa}) \mathrm{d}\boldsymbol{\kappa} \,, \tag{6.15}$$

where $p(\mathbf{u} \mid \boldsymbol{\kappa})$ is deterministic as defined in (6.11) but $\boldsymbol{\kappa}$ appears in it non-linearly, implying that the inference is not analytically tractable.

Throughout the experiments in the later sections, we either set Dirichlet (essential) boundary conditions everywhere (for example $u(\boldsymbol{x}) = 0$ on $\partial\Omega$), or assume Neumann (natural) boundary conditions on parts of the boundary. The choice will be made explicit in each experiment. To compute the likelihood, we solve the Poisson problem (6.8) for $u(\boldsymbol{x})$ using the finite element method (FEM).

6.2.4 Prior

As discussed above, we place a Gaussian measure on κ, $\mu_0(\kappa) \sim \mathcal{N}(m, \mathcal{C}_\kappa)$. Properties of samples from the measure depend on mean m and on the spectral properties of the covariance operator $\mathcal{C}_\kappa$. We restrict the space of prior functions to $L^2(\Omega;\mathbb{R})$. Then, operator $\mathcal{C}_\kappa$ can be constructed from the covariance function, $k(\boldsymbol{x},\boldsymbol{x}') = \mathbb{E}\big[\big(\kappa(\boldsymbol{x}) - m(\boldsymbol{x})\big)\big(\kappa(\boldsymbol{x}') - m(\boldsymbol{x}')\big)\big]$ as:

$$(\mathcal{C}_\kappa \varphi)(\boldsymbol{x}) = \int_\Omega k(\boldsymbol{x},\boldsymbol{x}')\varphi(\boldsymbol{x}')\mathrm{d}\boldsymbol{x}', \tag{6.16}$$

for any $\varphi \in L^2(\Omega;\mathbb{R})$. This formulation is what is commonly referred to as a Gaussian process (GP) with mean function $m(\cdot)$, which we assume to be zero, and covariance function $k(\cdot,\cdot)$ such that

$$\kappa \sim \mathcal{GP}\big(m(\cdot), k(\cdot,\cdot)\big). \tag{6.17}$$

Even though the process is infinite-dimensional, an instantiation of the process is finite and reduces to a multivariate Gaussian distribution by definition. The covariance function is typically parametrised by a set of hyperparameters ψ. One popular option, which satisfies assumptions about μ_0 as per Stuart (2010), is the squared exponential kernel:

$$k_{\mathrm{SE}}(\boldsymbol{x},\boldsymbol{x}') = \sigma_\kappa^2 \exp\left(-\frac{r^2}{2\ell_\kappa^2}\right), \tag{6.18}$$

where $r = \|\boldsymbol{x} - \boldsymbol{x}'\|_2$ is the Euclidean distance between the inputs. It depends on two hyper-parameters $\psi = \{\sigma_\kappa, \ell_\kappa\}$, the scaling parameter σ_κ, and the length-scale ℓ_κ. Note that, $k_{\mathrm{SE}}(\cdot,\cdot)$ is an infinitely smooth function, which implies that so is $\kappa(\cdot)$. The RBF kernel imposes smoothness and stationarity assumptions on the solution; in addition, such choice of kernel offers a way to regularise the resulting optimisation problem. This particular choice is driven more by convenience than real data: depending on the expert knowledge of the true solution, other kernels may be used to impose other assumptions such as periodicity.

Both conditioning and marginalisation of the GP can be done in closed form. In particular, consider the joint model of the values $\boldsymbol{\kappa}$ at training locations $\mathbf{X}$ and the unknown test values $\boldsymbol{\kappa}^*$ at test locations $\mathbf{X}^*$:

$$\begin{bmatrix} \boldsymbol{\kappa} \\ \boldsymbol{\kappa}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_\psi(\mathbf{X},\mathbf{X}) & \mathbf{K}_\psi(\mathbf{X},\mathbf{X}^*) \\ \mathbf{K}_\psi(\mathbf{X}^*,\mathbf{X}) & \mathbf{K}_\psi(\mathbf{X}^*,\mathbf{X}^*) \end{bmatrix}\right), \tag{6.19}$$

where $\mathbf{K}_\psi(\mathbf{X},\mathbf{X}^*)$ is the matrix resulting from evaluating $k(\cdot,\cdot)$ at all pairs of training and test points. The conditional distribution of the function values $\boldsymbol{\kappa}^*$ given the values $\boldsymbol{\kappa}$ at $\mathbf{X}$ is:

$$\boldsymbol{\kappa}^* \mid \boldsymbol{\kappa} \sim \mathcal{N}\left(\tilde{\boldsymbol{\kappa}}^*,\, \tilde{\mathbf{K}}\right), \tag{6.20}$$

where

$$\begin{aligned} \tilde{\boldsymbol{\kappa}}^* &= \mathbf{K}\left(\mathbf{X}^*,\mathbf{X}\right)\left[\mathbf{K}(\mathbf{X},\mathbf{X})\right]^{-1}\boldsymbol{\kappa} \\ \tilde{\mathbf{K}} &= \mathbf{K}\left(\mathbf{X}^*,\mathbf{X}^*\right) - \mathbf{K}\left(\mathbf{X}^*,\mathbf{X}\right)\left[\mathbf{K}(\mathbf{X},\mathbf{X})\right]^{-1}\mathbf{K}\left(\mathbf{X},\mathbf{X}^*\right). \end{aligned} \tag{6.21}$$

The marginal distribution can be recovered by finding the relevant part of the covariance matrix; for example, the marginal of $\boldsymbol{\kappa}$ given $\mathbf{X}$ is $\boldsymbol{\kappa} \sim \mathcal{N}\left(0, \mathbf{K}_\psi(\mathbf{X},\mathbf{X})\right)$.

In this work, we place a zero-mean Gaussian process prior on $\kappa(\boldsymbol{x})$ and assume the squared exponential kernel with length-scale ℓ_κ and fixed variance $\sigma_\kappa^2 = 1$. As mentioned in the previous section, we assume that $\kappa(\boldsymbol{x})$ is constant on each element of the mesh (we use the same mesh as for discretising $u(\boldsymbol{x})$ and $f(\boldsymbol{x})$). We place the prior on $\boldsymbol{\kappa}$ so that the centroids of the elements are the training points of the GP:

$$p(\boldsymbol{\kappa}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_\psi(\mathbf{X},\mathbf{X})). \tag{6.22}$$

6.3 Variational Bayes Approximation

We consider a variational Bayes approximation of the posterior distribution of $\boldsymbol{\kappa}$. We give the details below.

6.3.1 Variational Bayes

We assume that any hyper-parameters ψ of the prior are fixed, and are only interested in the posterior distribution of $\boldsymbol{\kappa}$. The variational approach proceeds by approximating the true posterior $p(\boldsymbol{\kappa} \mid \boldsymbol{y})$ according to (6.7) with a trial density $q(\boldsymbol{\kappa})$, which is the minimiser of the Kullback-Leibler (KL) divergence between a chosen family of trial densities $\mathcal{D}_q$ and the true posterior distribution $p(\boldsymbol{\kappa}|\boldsymbol{y})$, as discussed in Section 2.2.4. To find the approximate posterior distribution we have:

$$q^*(\boldsymbol{\kappa}) = \underset{q(\boldsymbol{\kappa})\in\mathcal{D}_q}{\arg\min}\ \mathrm{KL}(q(\boldsymbol{\kappa}) \parallel p(\boldsymbol{\kappa} \mid \boldsymbol{y})). \tag{6.23}$$

Following the derivation in Section 2.2.4, the task can be reformulated as:

$$q^*(\boldsymbol{\kappa}) = \underset{q(\boldsymbol{\kappa})\in\mathcal{D}_q}{\arg\max}\ \mathbb{E}_q\big[\log p(\boldsymbol{y} \mid \boldsymbol{\kappa})\big] - \mathrm{KL}(q(\boldsymbol{\kappa}) \parallel p(\boldsymbol{\kappa})). \tag{6.24}$$

To proceed, we compute $\mathbb{E}_q\big[\log p(\boldsymbol{y} \mid \boldsymbol{\kappa})$ using a Monte Carlo approximation using N_{SVI} samples, as described in Section 2.2.4. To compute the gradients qith respect to the parameters of $q(\boldsymbol{\kappa})$ we leverage the reparametrisation trick (see Section 2.2.4.1). Our empirical tests in later sections of this chapter show that the value of N_{SVI} in the range of 2–5 provides fast convergence of the optimisation, agreeing with previous literature (Kingma & Welling 2014). The Monte Carlo approximation of $\mathbb{E}_q\big[\log p(\boldsymbol{y} \mid \boldsymbol{\kappa})$ is in line with the work in Barajas-Solano & Tartakovsky (2019) but in contrast with the analytic approximation based on the Hessian calculations proposed in Tsilifis et al. (2016).

Throughout this chapter, we assume that $\mathcal{D}_q$ is the family of multivariate Gaussian distributions (see the discussion on the choice of $\mathcal{D}_q$ below). This assumption and the fact that $p(\boldsymbol{\kappa})$ is Gaussian implies that the second term of (6.24), $\mathrm{KL}(q(\boldsymbol{\kappa}) \parallel p(\boldsymbol{\kappa}))$, is available in closed form.

6.3.2 Specification of Trial Distribution

The specification of the approximating family of distributions determines how much structure of the true posterior distribution is captured by the variational approximation.

To model complex relationships between the components of the posterior, a more complex approximating family of distributions is needed. As the richer family of distributions is likely to require more parameters, the optimisation of the usually non-convex ELBO becomes harder. A balance must be struck in this trade-off: the family should be rich enough, but the optimisation task should still be computationally tractable.

A practical and widely used variational family is the multivariate Gaussian distribution, parametrised by the mean vector and the covariance matrix. One of the key benefits of this choice is that the KL divergence term of the ELBO in (2.36) is available in closed form for a GP prior. The choice of the parametrisation of the covariance matrix determines how much structure, other than the mean estimate, is captured by the variational family. We discuss this in more detail in the next section.

Numerous approaches have been proposed to extend the trial distribution beyond the Gaussian family. A standard approach in situations when the true posterior distribution is likely to be multimodal is to consider mixtures of variational densities (Bishop et al. 1998). A more recent development is embedding parameters of a mean-field approximation in a hierarchical model to induce variational dependencies between latent variables (Tran et al. 2015, Ranganath et al. 2016).

6.3.2.1 Gaussian Trial Distribution

Choosing the trial distribution $q(\boldsymbol{\kappa})$ as a multivariate Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ requires optimisation over the mean $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. The flexibility in choosing how we specify both of these parameters, especially the covariance matrix, enables us to balance the trade-off between the expressiveness of the approximating distribution and the computational efficiency.

The richest specification corresponds to parametrising the covariance matrix $\boldsymbol{\Sigma}$ using its full Cholesky factor $\boldsymbol{L}$, i.e.,

$$q(\boldsymbol{\kappa}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{L}\boldsymbol{L}^\top). \tag{6.25}$$

This choice results in a dense covariance matrix that may be able to capture the full covariance structure between the inputs (*i.e.* each input may be correlated with every other input). Parametrising the components of $\boldsymbol{L}$ automatically ensures that the covariance matrix $\boldsymbol{\Sigma}$ is positive definite as necessary. The number of parameters to optimise grows as $\mathcal{O}(n_\kappa^2)$ and this leads to a difficult optimisation task that needs to be carefully initialised and parametrised. We refer to this parametrisation as full-covariance variational Bayes (FCVB).

A much more efficient choice is a diagonal covariance matrix, which is often referred to as mean-field variational Bayes (MFVB). By limiting the number of parameters that need to be optimised, the optimisation task becomes simpler and the number of parameters grows only as $\mathcal{O}(n_\kappa)$. While more computationally efficient and easier to initialise, MFVB ignores much of the dependence structure of the posterior distribution.

6.3.3 Conditional Independence and Sparse Precision Matrices

Instead of parametrising the covariance matrix $\boldsymbol{\Sigma}$, or its Cholesky decomposition $\boldsymbol{L}$, in physical systems it is often advantageous to parametrise the precision matrix, $\boldsymbol{Q}$, where $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. While a component of the covariance matrix $\boldsymbol{\Sigma}$ expresses *marginal* dependence between the two corresponding random variables, the elements of the precision matrix reflect their *conditional independence* (Rue et al. 2009). Or, more specifically, for two components κ_i and κ_j of the random vector $\boldsymbol{\kappa}$ we note

$$p(\kappa_i,\, \kappa_j) = p(\kappa_i)p(\kappa_j) \quad \Leftrightarrow \quad \Sigma_{ij} = 0\,, \tag{6.26}$$

where Σ_{ij} denotes the respective component of $\boldsymbol{\Sigma}$. Furthermore, defining the vector $\boldsymbol{\kappa}_{-\{i,j\}}$ from the random vector $\boldsymbol{\kappa}$ by removing its i-th and j-th component, we note

$$p(\kappa_i,\, \kappa_j \mid \boldsymbol{\kappa}_{-\{i,j\}}) = p(\kappa_i \mid \boldsymbol{\kappa}_{-\{i,j\}})p(\kappa_j \mid \boldsymbol{\kappa}_{-\{i,j\}}) \quad \Leftrightarrow \quad Q_{ij} = 0\,. \tag{6.27}$$

That is, $Q_{ij} = 0$ if and only if κ_i is independent from κ_j, *conditional* on all other components of $\boldsymbol{\kappa}$.

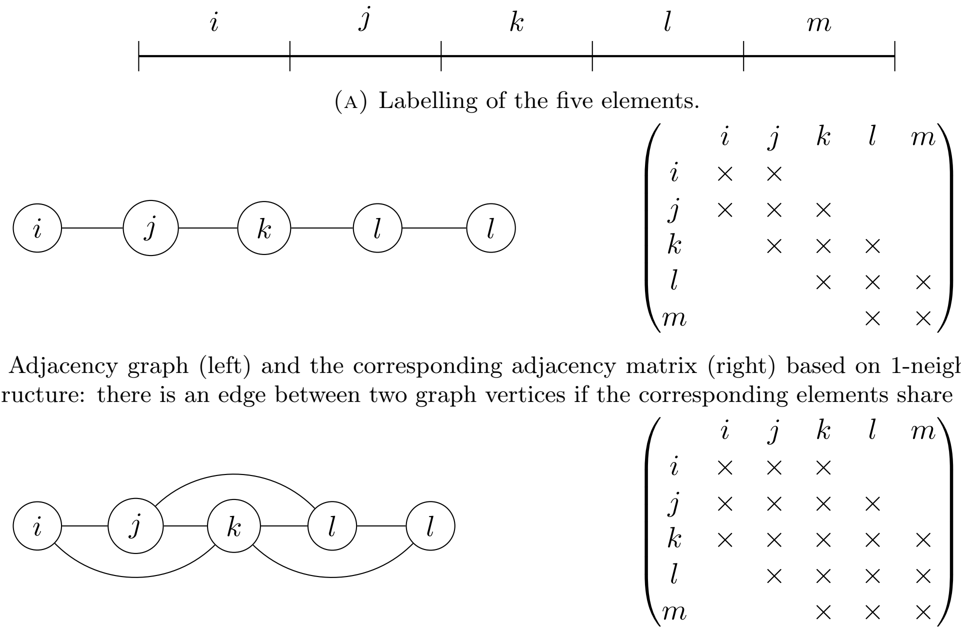
A succinct way to represent conditional independence is using an undirected graph whose nodes correspond to the random variables (Bishop 2006). A graph edge is present between two graph vertices i and j if the corresponding random variables are *not* conditionally independent from each other, given all the other random variables. Or, expressed differently, the edges between the graph vertices correspond to non-zeros in the precision matrix. In our context, each graph vertex represents a finite element and graph edges are introduced according to geometric adjacency of the finite elements as determined by the mesh. To this end, we define the 1-neighbourhood of a finite element as the union of the element itself and of elements sharing a node with the element. The n-neighbourhood is defined recursively as the union of all 1-neighbourhoods of all the elements in the $(n-1)$-neighbourhood. We introduce an edge between two graph vertices when the respective elements are in the same n-neighbourhood.

Figure 6.1 shows examples of adjacency graphs and the structure of the corresponding precision matrices $\boldsymbol{Q}$ for 5 random variables resulting from a discretisation of a 1D domain with 5 finite elements. In the considered examples the random variables represent the constant log-diffusion coefficient in the elements. As shown in Figures 6.1b and 6.1c choosing a larger n-neighbourhood for graph construction leads to a denser precision matrix. For instance, from the structure of the precision matrix in Figure 6.1b, which assumes a 1-neighbourhood structure, we can read for the log-diffusion coefficient of element j the following conditional independence relationship:

$$Q_{ik} = 0 \wedge Q_{il} = 0 \wedge Q_{im} = 0 \Rightarrow p(\kappa_i \mid \kappa_j,\, \kappa_k,\, \kappa_l,\, \kappa_m) = p(\kappa_i \mid \kappa_j)\,. \tag{6.28}$$

When the coefficient of element j is given, the coefficient of the neighbouring element i is independent from all the remaining coefficients. This is intuitively plausible and in line with physical observations. Clearly, the covariance matrices corresponding to the given sparse precision matrices are dense. Hence, in the considered case the coefficient of element i may still be correlated to the coefficient of element m, i.e. $p(\kappa_i \mid \kappa_m) \neq p(\kappa_i)$. This correlation will most likely be relatively weak given the large distance between the two elements, but knowing the coefficient of element m will certainly restrict the range of possible values for the coefficient of element i.

After obtaining the structure of the precision matrix, which is sparse but, in general, not banded, one can reorder the numbering of the elements in the finite element mesh to reduce its bandwidth. This allows for efficient linear algebra operations. See Cuthill & McKee (1969) for an example of a reordering algorithm. Once a minimum bandwidth ordering with b_{min} has been established, we use the property that the bandwidth of the Cholesky factor $\boldsymbol{L}_{\boldsymbol{Q}}$ of matrix $\boldsymbol{Q}$ is less than or equal to the bandwidth of $\boldsymbol{Q}$ (Rue & Held 2005). Finally, the parameters we optimise are the components of the lower band



(A) Labelling of the five elements.

$$\begin{pmatrix} & i & j & k & l & m \\ i & \times & \times & & & \\ j & \times & \times & \times & & \\ k & & \times & \times & \times & \\ l & & & \times & \times & \times \\ m & & & & \times & \times \end{pmatrix}$$

(B) Adjacency graph (left) and the corresponding adjacency matrix (right) based on 1-neighbourhood structure: there is an edge between two graph vertices if the corresponding elements share a node.

$$\begin{pmatrix} & i & j & k & l & m \\ i & \times & \times & \times & & \\ j & \times & \times & \times & \times & \\ k & \times & \times & \times & \times & \times \\ l & & \times & \times & \times & \times \\ m & & & \times & \times & \times \end{pmatrix}$$

(C) Adjacency graph (left) and the corresponding adjacency matrix (right) based on 2-neighbourhood structure: there is an edge between two vertices if the corresponding elements are in each others 2-neighbourhoods.

FIGURE 6.1: An example of a 1D bar discretised with five elements and two different conditional independence assumptions.

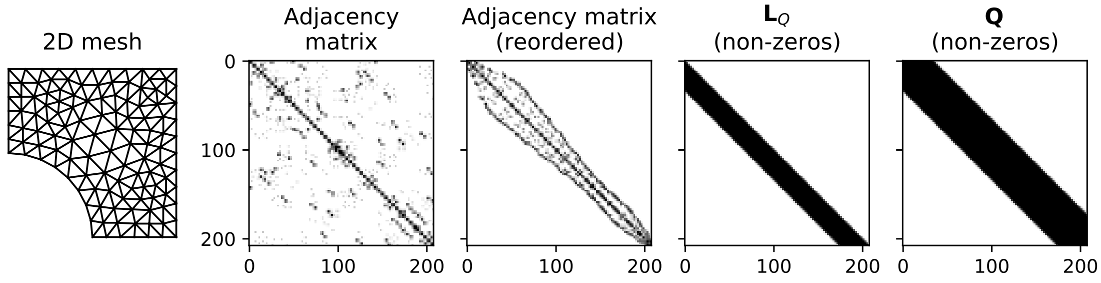


FIGURE 6.2: Sparse precision matrix parametrisation for a 2D problem. An order-2 neighbourhood structure is assumed for conditional independence. The structure of the adjacency matrix depends on the specific element numbering. By renumbering the elements, one can obtain a banded adjacency matrix, which is then used to parametrise the Cholesky factor of the precision matrix, as described in Section 6.3.2.1.

of size $b_{\min}$ of matrix $\boldsymbol{L}_Q$, so that the approximating distribution reads

$$q(\boldsymbol{\kappa}) \sim \mathcal{N}\Big(\boldsymbol{\mu}, (\boldsymbol{L}_Q \boldsymbol{L}_Q^\top)^{-1}\Big). \tag{6.29}$$

This process of devising a parametrisation for the precision matrix for a more complex mesh in 2D is illustrated in Figure 6.2. This approach is computationally efficient – the number of parameters grows as $\mathcal{O}(n_\kappa)$ – and is able to capture dependencies between all the random variables.

6.3.4 Stochastic Optimisation

To maximise the ELBO in (6.24), we use the ADAM algorithm (Kingma & Ba 2015). ADAM is a member of a larger class of stochastic optimisation methods that have become popular as tools for maximising non-convex cost functions. These methods construct a stochastic estimate of the gradient to perform gradient descent-based optimisation. ADAM, a stochastic gradient descent algorithm with an adaptive step size is one popular algorithm that exhibits a stable behaviour on many problems and is easy to use without significant tuning. The algorithm uses a per-parameter step size, which is based on the first two moments of the estimate of the gradient for each parameter. Specifically, the step size is proportional to the ratio of the exponential moving average of the 1st moment to the square root of the exponential moving average of the non-centred 2nd moment. At any point, the exponential moving average is computed with decay parameters β_1 and β_2 for the 1st and 2nd moment, respectively. We adopt the parameter values suggested in Kingma & Ba (2015): $\beta_1 = 0.9$ and $\beta_2 = 0.99$. The speed of convergence is further controlled by the learning parameter α which is used to regulate the step size for all parameters in the same way. In our experiments, we set it to 0.01 and let it decay exponentially every 2,500 steps (1,000 for MFVB), with the decay rate of 0.96. While the ADAM algorithm performs well on a variety of problems, it has been shown that the convergence of this algorithm is poor on some problems (Reddi et al. 2018). We discuss alternative approaches as potential future work in Section 6.5.

To monitor convergence, we use a rule that tracks an exponentially weighted moving average of the decrease in the loss values between successive steps, and stops when that average drops below a threshold. The use of such an adaptive rule gives us a way to track the convergence of the algorithm and provides a conservative estimate for the time it takes for the optimisation to converge. This rule can be adapted based on the available computational budget.

6.3.5 The Algorithm

The maximisation of the ELBO in (2.36) involves finding the parameters of the trial distribution $q(\boldsymbol{\kappa})$, i.e. its mean $\boldsymbol{\mu}$ and Cholesky factor $\boldsymbol{L_Q}$, that minimise KL between $q(\boldsymbol{\kappa})$ and the posterior $p(\boldsymbol{\kappa}|\boldsymbol{y})$. Algorithm 3 shows the required steps to compute the ELBO and its gradients with respect to the parameters of the trial distribution. Different from the discussion so far, in Algorithm 3 it is assumed that there are multiple independent observation vectors $\mathbf{y}_i$ with $i \in \{1, 2, \dots, N_y\}$.

Algorithm 3: ELBO estimation and its gradient with respect to the parameters of the trial distribution.

```
Input: Current parameters $\boldsymbol{\mu}$ and $\boldsymbol{L_Q}$ of $q(\boldsymbol{\kappa})$
Output: ELBO and its gradients with respect to the parameters of $q(\boldsymbol{\kappa})$
Sample $[\boldsymbol{\kappa}^{(1)}, \boldsymbol{\kappa}^{(2)}, \dots, \boldsymbol{\kappa}^{(N_{\mathrm{SVI}})}]$ from $q(\boldsymbol{\kappa})$
for each $\boldsymbol{\kappa}^{(i)}$ do
    Solve for $\mathbf{u}(\boldsymbol{\kappa}^{(i)})$ and obtain gradients with respect to $\boldsymbol{\kappa}$ using the FEM
    $p(\boldsymbol{y} \mid \boldsymbol{\kappa}^{(i)}) \leftarrow \prod_{j=1}^{N_y} p(\boldsymbol{y}_j; \mathbf{u}(\boldsymbol{\kappa}^{(i)}), \sigma_y^2)$ and propagate its gradient with respect to
      $\boldsymbol{\kappa}^{(i)}$
ELBO $\leftarrow N_{\mathrm{SVI}}^{-1} \sum_{i=1}^{N_{\mathrm{SVI}}} \log p(\boldsymbol{y} \mid \boldsymbol{\kappa}^{(i)}) + \mathrm{KL}(q(\boldsymbol{\kappa}) \parallel p(\boldsymbol{\kappa}))$ and propagate the gradient
  with respect to the parameters of $q(\boldsymbol{\kappa})$ using the reparametrisation trick
  (see Section 2.2.4.1 and Kingma & Welling (2014))
return ELBO, $\nabla$ELBO
```

6.4 Examples

We evaluate the efficacy of variational inference first for 1D and 2D Poisson equation examples and then a benchmark proposed by Aristoff & Bangerth (2021). We discretise the examples with a standard finite element method using linear Lagrange basis functions. We perform inference over $\boldsymbol{\kappa}$ and we keep the hyper-parameter of the prior, $\psi = \{\sigma_\kappa, \ell_\kappa\}$, fixed at a chosen value, described in each of the experiments. We compare variational Bayes methods against two sampling-based inference schemes, Hamiltonian Monte Carlo (HMC) and pre-conditioned Crank-Nicholson Markov Chain Monte Carlo (pCN); both are known to be asymptotically correct as the number of samples increases. The evaluation criteria we use focus on three aspects of an inference scheme: the accuracy with respect to capturing the mean and the variance of the solution; propagation of uncertainty in derived quantities of interest; and the time until convergence of the solution.

To assess the propagation of uncertainty in derived quantities of interest, we consider a summary quantity for which a point estimate alone may not be informative enough for downstream tasks. In particular, we compute the log of total boundary flux through the boundary Γ_b:

$$r(\kappa) = \log \int_{\Gamma_b} e^{\kappa(s)} \nabla u(s) \cdot \boldsymbol{n} \, \mathrm{d}s, \tag{6.30}$$

where $\boldsymbol{n}$ is a unit vector normal to the boundary Γ_b.

To quantitatively assess the inference of $\boldsymbol{\kappa}$, we obtain S samples from the posterior distribution of $\boldsymbol{\kappa}$, $\{\boldsymbol{\kappa}^{(s)}\}_{s=1}^{S}$. For synthetic experiments, where we know the true $\boldsymbol{\kappa}$ which generated the observations, we compute the mean $\boldsymbol{\kappa}$ error norm. The computation is the Euclidean norm of the error between the true value, $\boldsymbol{\kappa}_{\text{true}}$, and the mean of the obtained samples:

$$\text{Mean } \boldsymbol{\kappa} \text{ error} = \left\| \frac{1}{S} \sum_{s=1}^{S} \boldsymbol{\kappa}^{(s)} - \boldsymbol{\kappa}_{\text{true}} \right\|_2 . \tag{6.31}$$

Further, we compute the expected error in the solution space. This measures how close the solutions corresponding to the samples of $\boldsymbol{\kappa}$ are to the true solution $\mathbf{u}(\boldsymbol{\kappa}_{\text{true}})$. Specifically, we compute

$$\text{Mean } \mathbf{u}(\boldsymbol{\kappa}) \text{ error} = \frac{1}{S} \sum_{s=1}^{S} \left\| \mathbf{u}(\boldsymbol{\kappa}^{(s)}) - \mathbf{u}(\boldsymbol{\kappa}_{\text{true}}) \right\|_2 . \tag{6.32}$$

6.4.1 One-dimensional Poisson Equation Experiments

For this experiment, we assume the unit-line domain, which is discretised into 32 equal-length elements. We impose Dirichlet boundary conditions on both boundaries, specifically we set $u(0) = u(1) = 0$; the forcing is constant everywhere $f(\mathbf{x}) = 1$. Unless specified otherwise, all experiments in this section use $N_y = 5$ observations per sensor and the sensor noise $\sigma_y = 0.01$. Sensors are located on each of the discretisation nodes. For the prior on κ, we choose a zero-mean Gaussian process with squared exponential kernel (see Section 6.2.4 for details). We compare the results for three specifications of the prior length-scale, $\ell_\kappa \in \{0.1, 0.2, 0.3\}$. The length-scale used to generate the data is $\ell_\kappa = 0.2$. For inferences made using data generated by a shorter length-scale, see Section B.1.

6.4.1.1 VB Performs Competitively Based on Error Norms

Figure 6.3 shows the mean $\boldsymbol{\kappa}$ error norm (6.31) and the expected solution error norm (6.32) obtained from 10,000 posterior samples of $\boldsymbol{\kappa}$ from Hamiltonian Monte Carlo (HMC),

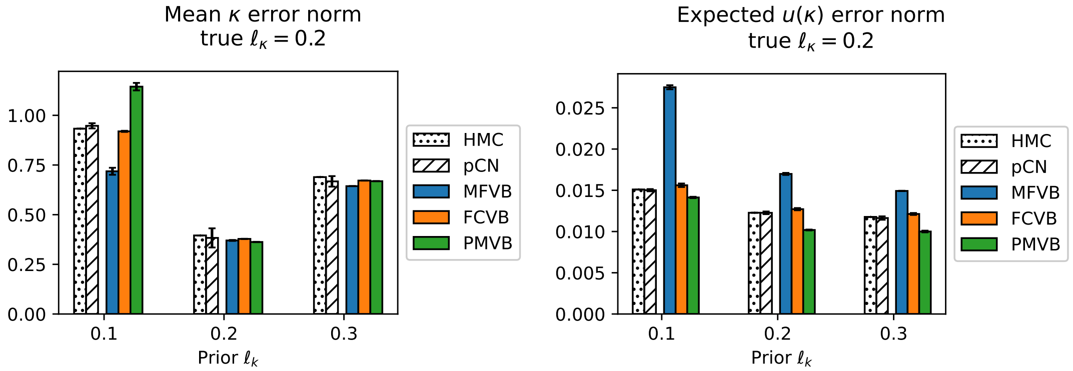


FIGURE 6.3: Mean $\boldsymbol{\kappa}$ error norm for the Poisson 1D problem (left), as defined in (6.31), and expected solution error norm (right), as defined in (6.32). Both quantities are estimated using 10,000 samples from the inferred posterior distribution of $\boldsymbol{\kappa}$. Quantitatively, the sampling methods (HMC and pCN) and VB produce comparable results in both metrics. One can observe that the lower error in the parameter $\boldsymbol{\kappa}$ space does not necessarily imply lower error in the solution $\boldsymbol{u}$ space. This is likely due to non-linear dependence of $\boldsymbol{u}$ on $\boldsymbol{\kappa}$. For a qualitative comparison, see Figure 6.4 where each row of results corresponds to a different value of the true prior length-scale ℓ_κ.

pre-conditioned Crank-Nicholson MCMC (pCN), as well as VB inference with different parametrisations of the covariance/precision matrix. It is evident that for prior length-scales $\ell_\kappa \in \{0.2, 0.3\}$, the mean $\boldsymbol{\kappa}$ error norms computed by the variational Bayes methods are very close to the estimates from HMC and pCN. For prior $\ell_\kappa = 0.1$, the mean $\boldsymbol{\kappa}$ error norm computed by MFVB is lower than other VB methods and MCMC methods. This is most likely due to MFVB being a much easier optimisation task compared to other VB methods with more optimisation parameters that capture dependencies. For the expected solution error norm, MFVB posterior consistently underestimates the uncertainty in $\boldsymbol{\kappa}$, thus ignoring possible values of $\boldsymbol{\kappa}$ which are consistent with the data. This is further confirmed in the qualitative assessment of uncertainty in the next section. While MCMC methods are asymptotically correct, in practice, devising efficient samplers for high-dimensional problems within a limited computational budget is still a challenging task and requires substantial hand-tuning. To affirm that all the VB methods provide a good estimate of the mean of $\boldsymbol{\kappa}$, as compared to MCMC methods, is better demonstrated by inspecting Figure 6.4 which shows not only the mean but also the posterior uncertainty, which we discuss next.

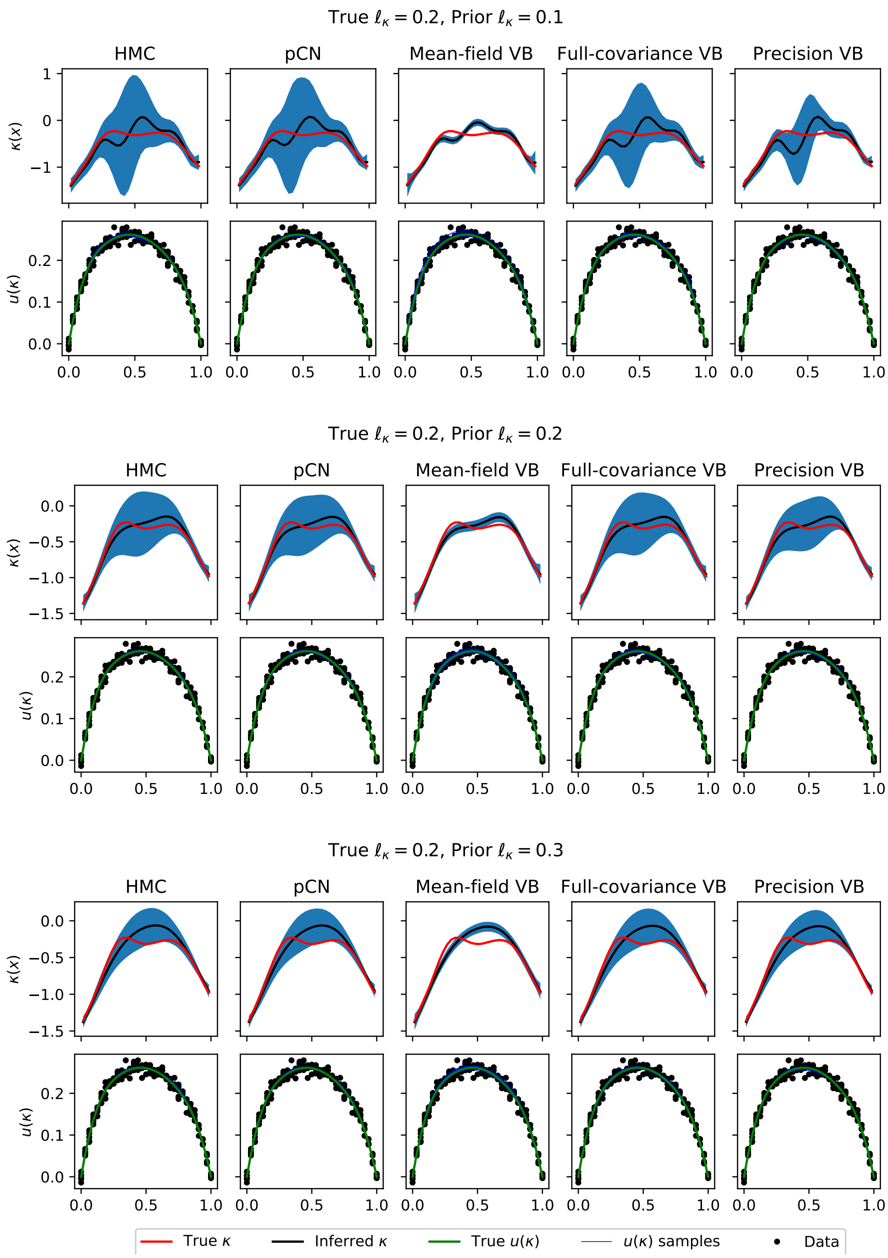


FIGURE 6.4: Top row in each of the three panels show true values of $\boldsymbol{\kappa}$ (red), posterior means (black) and plus and minus two times the standard deviation (blue shaded regions) for HMC, pCN, and VB variants for different values of prior length-scales ℓ_κ. The bottom rows show the data (black), true solution $\mathbf{u}$ (green), solutions for different samples of κ (blue). For the PMVB estimate, the bandwidth is set to 10.

6.4.1.2 VB Adequately Estimates Posterior Variance

Figure 6.4 shows the true values of $\boldsymbol{\kappa}$ (red), the posterior means (black) and plus and minus two times the standard deviation (blue shaded regions) estimated by HMC, pCN, and variational inference with mean-field (MFVB), full covariance (FCVB), and precision matrix (PMVB) parametrisations for different values of prior length-scales. We observe

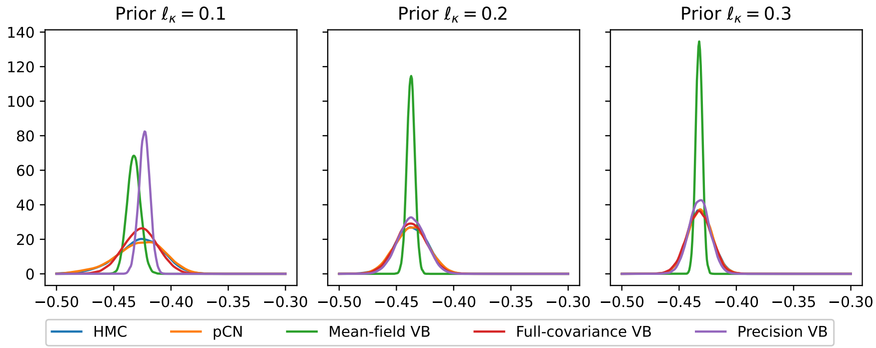


FIGURE 6.5: Log of the boundary flux at the left boundary node ($x = 0$) for the 1D Poisson example. For PMVB, the precision matrix bandwidth of 10 is used.

that the posterior variance estimates computed by HMC, pCN, and full covariance VB are qualitatively very similar, with the estimated uncertainty increasing with increasing distance from the fixed boundary. However, the MFVB solution greatly underestimates posterior variance while computing a reasonable estimate of the posterior mean. The over-confidence of MFVB means that values of $\boldsymbol{\kappa}$ that are consistent with the observed data are ignored; this may lead to poor calibration if the MFVB posterior is used as the true $\boldsymbol{\kappa}$ in downstream tasks or in other contexts. For the PMVB parametrisation, the uncertainty is underestimated to a much lesser extent.

The observations above are further confirmed by the density plot of our quantity of interest: the log of the total flux on the boundary, shown in Figure 6.5. For this example, we compute the flux on the left boundary at $x = 0$ and show the posterior distribution of this quantity. For longer prior length-scales, FCVB and PMVB agree with the estimates obtained from pCN and HMC, whereas mean-field VB underestimates the uncertainty. For the short prior length-scale ($\ell_\kappa = 0.1$), both PMVB and MFVB underestimate the uncertainty as compared with HMC, pCN, and FCVB schemes. The posterior distribution of FCVB approximately agrees with the MCMC schemes.

For the results obtained using the PMVB scheme, we used the 10-neighborhood structure to define the adjacency matrix and the non-zero elements of the precision matrix, $\boldsymbol{Q}$ (see Section 6.3.3). The order of the neighbourhood structure, which corresponds to the precision matrix bandwidth, determines how much dependence within $\boldsymbol{\kappa}$ is captured by the approximating posterior distribution. In Figure 6.6, we show how the estimate of the mean and the variance of $\boldsymbol{\kappa}$ changes for different orders of neighbourhood structure. As expected, with the increasing bandwidth, the posterior estimate of $\boldsymbol{\kappa}$ gets closer to the

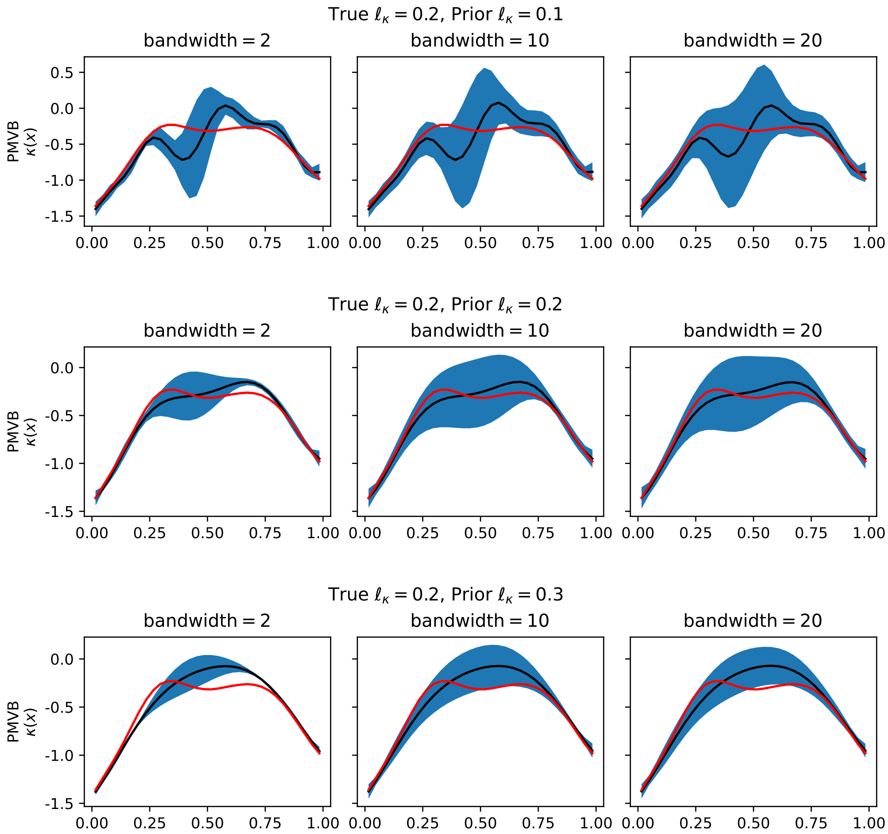


FIGURE 6.6: True values of $\boldsymbol{\kappa}$ (red), posterior means (black) and plus and minus two times the standard deviation (blue shaded region) for different matrix bandwidths of the precision matrix parametrisation of VB. Bandwidth corresponds to the order of neighbourhood structure considered when parametrising $\boldsymbol{Q}$.

estimate of FCVB, HMC, and pCN (shown in Figure 6.4). While there is a significant change in the uncertainty estimate when we increase the bandwidth from 2 to 10, it is less pronounced when we change it from 10 to 20. For this reason, we choose the value of 10 for the PMVB parametrisation in 1D.

6.4.1.3 VB Estimates Improve with More Observations and Decreasing Observational Noise

The consistency of the posterior refers to the contraction of the posterior distribution to the truth as the data quality increases, *i.e.* either the number of observations increases or observation noise tends to zero. A recent line of work (Abraham & Nickl 2020, Monard et al. 2020, Giordano & Nickl 2020) showed the posterior consistency for the estimates obtained using popular MCMC schemes such as pCN or unadjusted discretised Langevin algorithm for Bayesian inverse problems based on PDE forward mappings. While similar results are not available for VB methods in infinite-dimensional case, consistency and Bernstein-von Mises type results have been shown for the finite-dimensional case, including Bayesian inverse problems (Wang & Blei 2019, Lu et al. 2017). Empirically, our experiments show that for the given family of trial distributions the VB posterior distribution contracts to the true $\boldsymbol{\kappa}$.

Firstly, we show that increasing the number of observations, N_y, results in a more accurate estimate. Given that the observations, $\{\boldsymbol{y}_i\}_{i=1}^{N_y}$, are independent of each other, the likelihood term of the ELBO (see (2.36)) is the product of the individual likelihood terms:

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{N_y} \mid \boldsymbol{\kappa}) = \prod_{i}^{N_y} p(\boldsymbol{y}_i \mid \boldsymbol{\kappa}). \tag{6.33}$$

Secondly, by decreasing the observational noise σ_y we expect the posterior distribution to get closer to the ground truth and with lower uncertainty. Figure 6.7 shows the true values of $\boldsymbol{\kappa}$ (red), the posterior mean estimates (black) and plus and minus two times the standard deviation (blue shaded regions) obtained by different variants of variational Bayes for varying numbers of observations (top panel) and different values of observational noise (bottom panel). We can see that MFVB underestimates the posterior variances and these estimates do not depend on the number of observations (top panel in Figure 6.7) or the amount of observational noise (bottom panel in Figure 6.7). However, the FCVB and PMVB uncertainty estimates get narrower with increasing number of observations and with decreasing observational noise, which is a desirable behaviour that should be exhibited by any consistent uncertainty estimation method. We can also see that the true solution is contained within the uncertainty bounds for all numbers of observations and noise levels for the full covariance parametrisation. This is not the case for the mean-field VB, providing another indication of uncertainty underestimation for this parametrisation.

6.4.1.4 VB Is an Order of Magnitude Faster than HMC

For HMC estimates, we obtain 200,000 samples out of which the first 100,000 are used to calibrate the sampling scheme and are subsequently discarded. Table 6.1 provides the run-times for HMC, MFVB, FCVB, and PMVB. For the HMC column, we also report (shown in brackets) the range of effective sample sizes (ESS) across different components

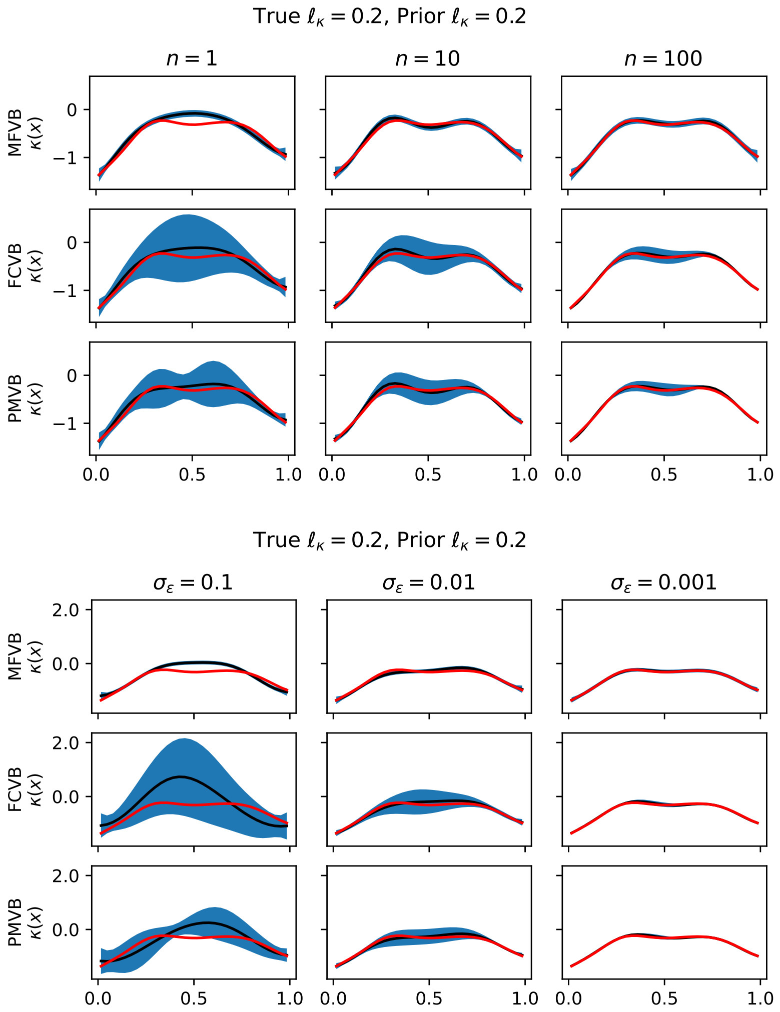


FIGURE 6.7: True values of $\boldsymbol{\kappa}$ (red), posterior means (black) and plus and minus two times the standard deviation (blue shaded regions) for VB with different parametrisations for different number of observations per sensor, $N_y \in \{1, 10, 100\}$ (top panel), and for different values of sensor noise $\sigma_\epsilon \in \{0.1, 0.01, 0.001\}$ (bottom panel).

of $\boldsymbol{\kappa}$. For details on ESS, we refer the reader to (Gelman et al. 2013, Ch. 11). Even with conservative convergence criteria (described in Section 6.3.4), the computational cost of VB algorithms is up to 25 times lower than that of HMC. To emphasise the computational efficiency of the variational inference, we show the posterior estimates for different number of Monte Carlo samples in the estimation of ELBO. Figure 6.8 shows that on a qualitative level, a low number of samples is sufficient to obtain a good estimate. In particular, even with 2 Monte Carlo samples, the estimates are very similar to the case where $N_{\text{SVI}} = 20$. However, a lower number of samples may result in slower convergence of the optimisation scheme. Figure 6.9 shows that for the FCVB and PMVB parametrisations, where the number of optimised parameters is larger than for MFVB, increasing the number of SVI samples may speed up the convergence of the optimisation.

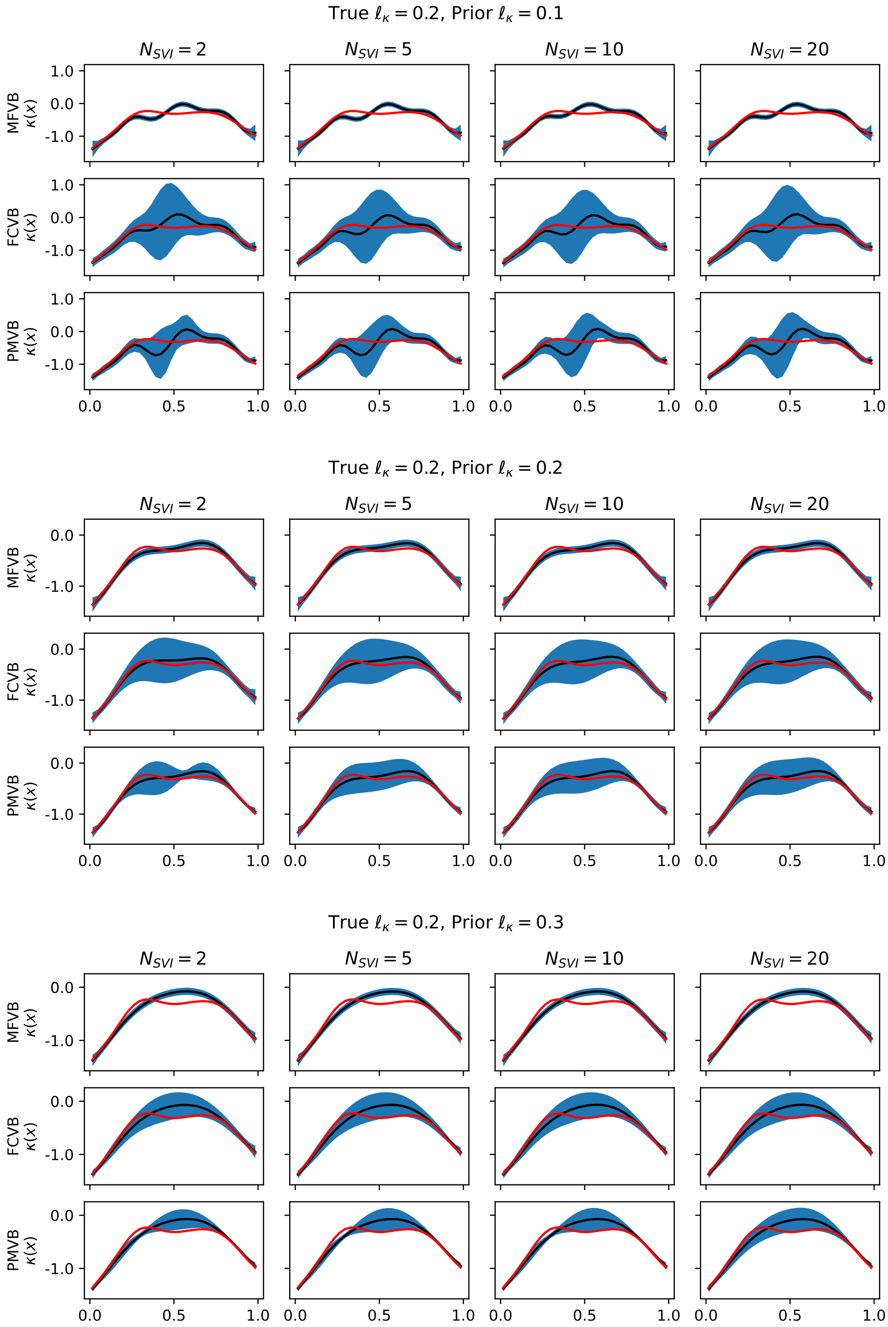


FIGURE 6.8: True values of $\boldsymbol{\kappa}$ (red), posterior means (black) and plus and minus two times the standard deviation (blue shaded regions) of VB with different parametrisations for varying number of Monte Carlo samples when computing ELBO. Three different length-scales for the prior are shown: 0.1, 0.2, 0.3.

true ℓ_κ	prior ℓ_κ	Time (hours)				
		HMC		MFVB	FCVB	PMVB
0.1	0.1	15.2	(871–3244)	1.1	3.6	2.1
	0.2	11.1	(1043–4006)	0.7	2.7	2.1
	0.3	7.2	(1130–5408)	0.6	2.3	2.0
0.2	0.1	15.2	(1600–4700)	0.6	2.2	1.8
	0.2	10.4	(1067–3468)	0.6	2.3	2.0
	0.3	7.0	(1487–3969)	0.5	1.7	1.8

TABLE 6.1: Run-times for different inference schemes in hours for the Poisson 1D problem. For VB methods, $N_{\text{SVI}} = 3$. The column for HMC includes the range of effective sample sizes (ESS) across different components of $\boldsymbol{\kappa}$.

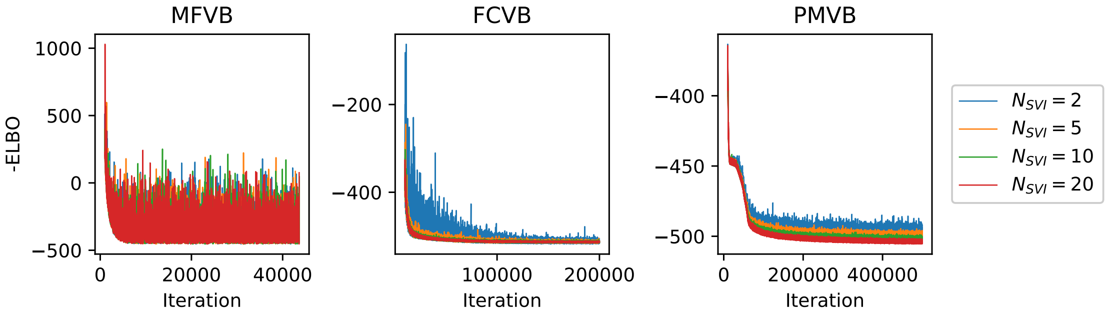


FIGURE 6.9: Negative ELBO trace plot for both MFVB and FCVB for different values of N_{SVI}. For this example, true $\ell_\kappa = 0.2$ and prior $\ell_\kappa = 0.1$.

The effect is not as strong for the MFVB parametrisation.

6.4.2 Two-dimensional Poisson Equation Experiments

We consider a 2D Poisson problem on the unit-square domain with a circular hole as shown in Figure 6.10, with boundary conditions as indicated in the same figure. The problem is discretised with 208 linear triangular elements and 125 nodes. The forcing term is assumed to be constant throughout the domain, $f(\boldsymbol{x}) = 1$. Unless specified otherwise, all experiments in this section use $N_y = 5$ observations per sensor and the sensor noise $\sigma_y = 0.001$ (note that for the 1D example we used $\sigma_y = 0.01$). The sensors are located at each node of the mesh. As in the 1D example, we assume a zero-mean GP prior on κ with square exponential kernel with varying length-scale, ℓ_κ, as discussed in Section 6.2.4.

Firstly, the results in Figure 6.11 show that the mean $\boldsymbol{\kappa}$ error of VB methods is very similar to the sampling methods (pCN and HMC). Similarly to the 1D case, the expected solution error norm is highest for MFVB estimate, indicating the lack of capturing the

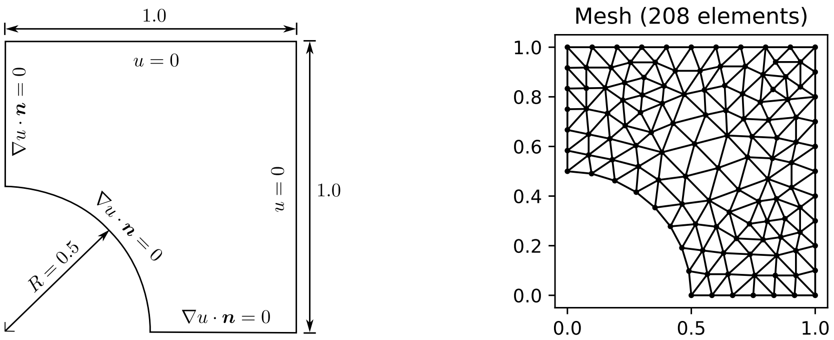

1.0
$u = 0$
$\nabla u \cdot \boldsymbol{n} = 0$
$u = 0$
1.0
$\nabla u \cdot \boldsymbol{n} = 0$
$R = 0.5$
$\nabla u \cdot \boldsymbol{n} = 0$
Mesh (208 elements)
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0

FIGURE 6.10: Left: Specification of the domain for the 2D Poisson problem. Note that we impose Dirichlet boundary conditions $u(x, y) = 0$ when $x = 1$ or $y = 1$. We impose Neumann boundary conditions on the rest of the boundary. Right: a triangular discretisation of the domain.

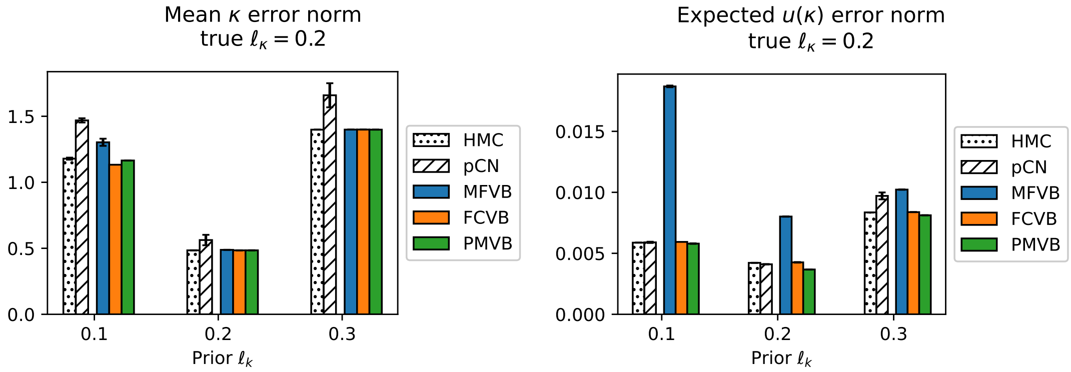


FIGURE 6.11: Mean $\boldsymbol{\kappa}$ error norm for the Poisson 2D problem (left), as defined in (6.31), and expected solution error norm (right), as defined in (6.32). Both quantities are estimated using 10,000 samples from the inferred posterior distribution of $\boldsymbol{\kappa}$. Quantitatively, the sampling methods (HMC and pCN) and VB produce comparable results in both metrics. One can observe that the lower error in the parameter $\boldsymbol{\kappa}$ space does not necessarily imply lower error in the solution $\boldsymbol{u}$ space. This is likely due to non-linear dependence of $\boldsymbol{u}$ on $\boldsymbol{\kappa}$. For a qualitative comparison, see Figures 6.12 – 6.14.

possible values of $\boldsymbol{\kappa}$ for which the solutions, $\mathbf{u}(\boldsymbol{\kappa})$, are consistent with the observed data. The results also show that both errors are lowest when the prior ℓ_κ matches the length-scale used to generate the data.

Figures 6.12 – 6.14 show the results for the posterior mean and the standard deviation of $\boldsymbol{\kappa}$, and the solution $\mathbf{u}(\boldsymbol{\kappa})$ corresponding to the mean of the posterior. We consider three configurations with prior length-scale $\ell_\kappa \in \{0.1, 0.2, 0.3\}$, where the length-scale used to generate the data is $\ell_\kappa = 0.2$. In all cases, the estimates of the posterior mean of $\boldsymbol{\kappa}$ and the corresponding solutions $\mathbf{u}$ are very close to the true values. Similarly to the 1D case discussed in Section 6.4.1, the variance estimates between HMC and FCVB are consistent, especially for longer prior length-scales. There seems to be a discrepancy

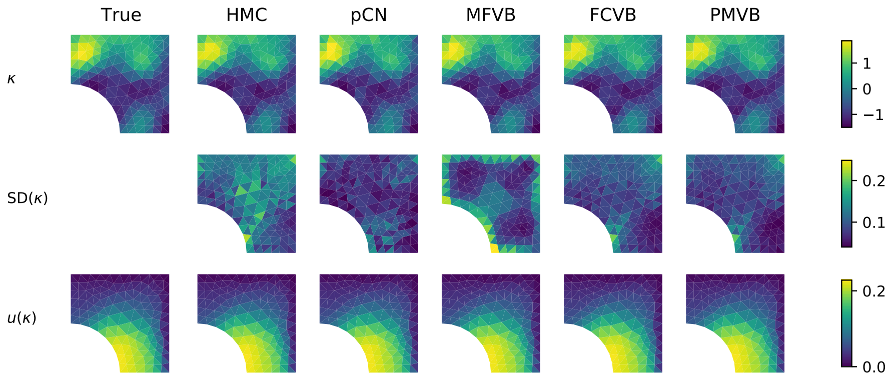


FIGURE 6.12: Posterior mean and standard deviation for $\boldsymbol{\kappa}$ and the corresponding $\mathbf{u}$ for 2D Poisson example with prior length-scale $\ell_\kappa = 0.1$.

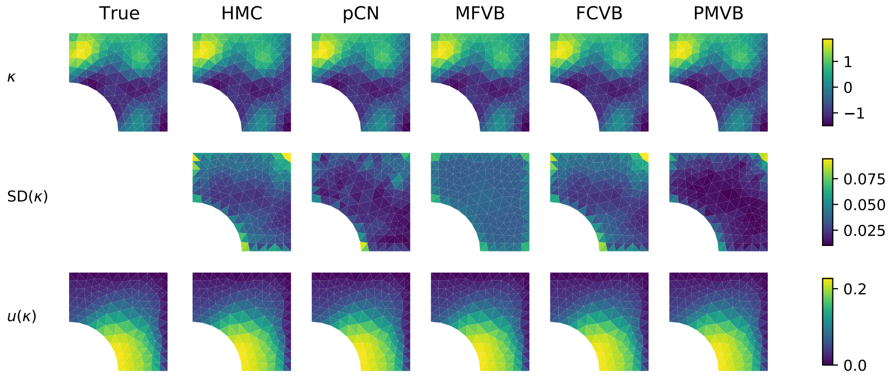


FIGURE 6.13: Posterior mean and standard deviation for $\boldsymbol{\kappa}$ and the corresponding $\mathbf{u}$ for 2D Poisson example with prior length-scale $\ell_\kappa = 0.2$.

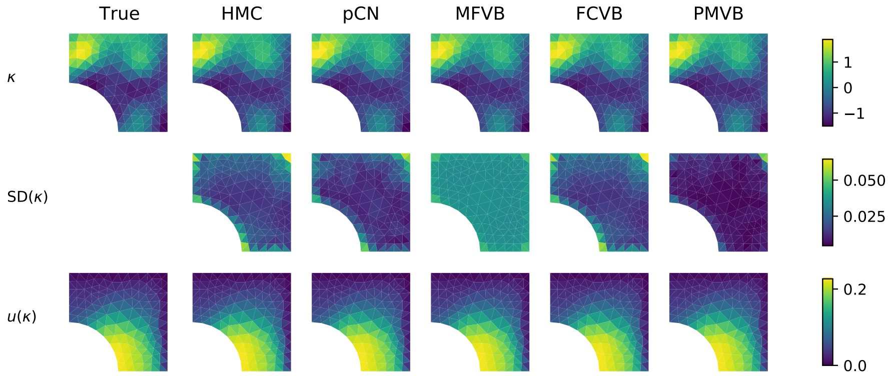


FIGURE 6.14: Posterior mean and standard deviation for $\boldsymbol{\kappa}$ and the corresponding $\mathbf{u}$ for 2D Poisson example with prior length-scale $\ell_\kappa = 0.3$.

between the estimates obtained using MFVB and those obtained by other methods. The estimates obtained using precision-matrix parametrisation are qualitatively very close to the FCVB and MCMC estimates.

For the quantity of interest, we compute the log of the total flux along the right boundary of the domain ($x = 1$), and the results are shown in Figure 6.15. Unlike the 1D case,

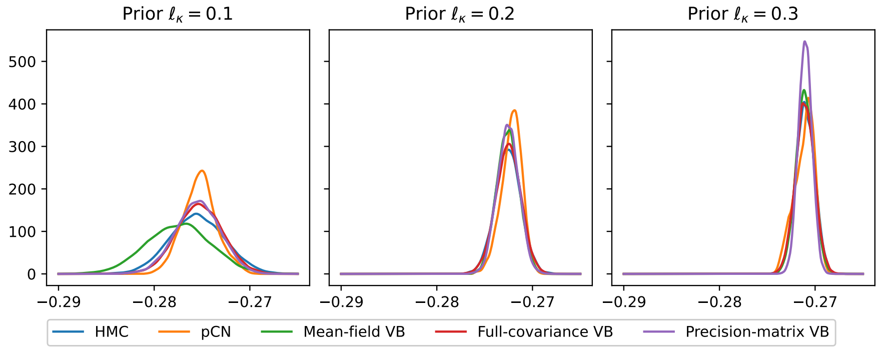


FIGURE 6.15: Log of the total flux computed along the right boundary ($x = 1$). For PMVB, the precision matrix is parametrised using the second-order neighbourhood structure, as shown in Figure 6.2.

true ℓ_κ	prior ℓ_κ	Time (hours)				
		HMC		MFVB	FCVB	PMVB
0.1	0.1	240.6	(930–11200)	6.4	29.6	28.1
	0.2	295.5	(1537–11067)	6.6	32.6	28.9
	0.3	242.0	(1057–6068)	7.3	27.3	30.6
0.2	0.1	242.7	(1102–18235)	6.2	34.3	27.2
	0.2	264.3	(1304–9848)	7.4	33.7	34.0
	0.3	221.9	(1192–6356)	7.8	31.3	34.0

TABLE 6.2: Run-times for different inference schemes in seconds. The number of Monte Carlo samples is $N_{\text{SVI}} = 5$ for all MFVB, FCVB, and PMVB. The column for HMC includes the range of effective sample sizes (ESS) across different components of $\boldsymbol{\kappa}$.

the posterior estimates of the boundary flux are approximately the same for all the considered methods, except for the mean-field estimate when prior $\ell_\kappa = 0.1$, where the MFVB estimate is biased as compared to the other methods.

The empirical computational cost for these experiments is given in Table 6.2. For the HMC experiments, we obtained 250,000 samples, out of which the first 125,000 were used to calibrate the sampling scheme and discarded afterwards. The timing results show that HMC takes an order of magnitude longer than variational Bayes, with some variation that depends on the parametrisation.

6.4.3 Inverse Problem Benchmark

We evaluate the effectiveness of VB methods on a recently proposed benchmark for Bayesian inverse problems (Aristoff & Bangerth 2021). The benchmark aims to provide a test case that reflects practical applications, but at the same time is easy to replicate.

Like above, the test case is a Poisson inverse problem where the task is to recover log-diffusion, κ, from a finite set of noisy observations. The problem domain is a unit square, the forcing function $f(\boldsymbol{x}) = 10$ is constant throughout the domain, and the solution of the PDE is imposed to be zero on all four boundaries.

The benchmark discretises κ using 64 quadrilateral elements, such that κ is constant for each individual element as shown in Figure 6.16. The forward solution of the PDE is obtained after discretising u using 32×32 bilinear quadrilateral elements. The locations where the solution is observed are placed on a uniform grid of 169 points (13×13). The measurements are corrupted by the Gaussian noise with standard deviation $\sigma_y = 0.05$. The authors of the benchmark provide the measurements as well as the true log-diffusion coefficient κ which generated the observations. The true log-diffusion coefficient, shown in Figure 6.16, is zero throughout the domain, except two regions, where the value is $\log(10)$ and $\log(0.1)$. It is these two jumps that make it a non-trivial test case.

Unlike in the previous examples, we place a prior on κ which does not induce any spatial correlation between any of the $\boldsymbol{\kappa}$ coefficients. The role of the prior is to express our belief about the ranges of the coefficients, rather than any dependencies. Although authors place $\mathcal{N}(\mu = 4, \sigma^2 = 4)$ for each component of $\boldsymbol{\kappa}$ independently, we choose $\mathcal{N}(\mu = 0, \sigma^2 = 1)$ as most of the coefficients of the true $\boldsymbol{\kappa}$ are at the baseline level equal to zero, and the fact that the $\boldsymbol{\kappa}$ corresponds to the diffusion parameter on the log-scale, a priori we do not expect such high variance.

We performed the inference using HMC, MFVB, FCVB, and PMVB. The means and standard deviations of inferred log-diffusion coefficients, together with the PDE solutions corresponding to the inferred means, are shown in Figure 6.16. The results suggest that the mean estimates of all three methods do capture the jumps and the overall structure of $\boldsymbol{\kappa}$. Specifically, the FCVB estimate of the mean of $\boldsymbol{\kappa}$ is closest to the true value. As for uncertainty quantification, the MFVB and PMVB estimates are closer to the HMC estimate (our assumed ground truth for the uncertainty) than the FCVB estimate. The FCVB estimate seems to overestimate the uncertainty at a few locations. This is potentially due to being stuck in a local optimum during the optimisation procedure, which for FCVB involves high-dimensional exploration.

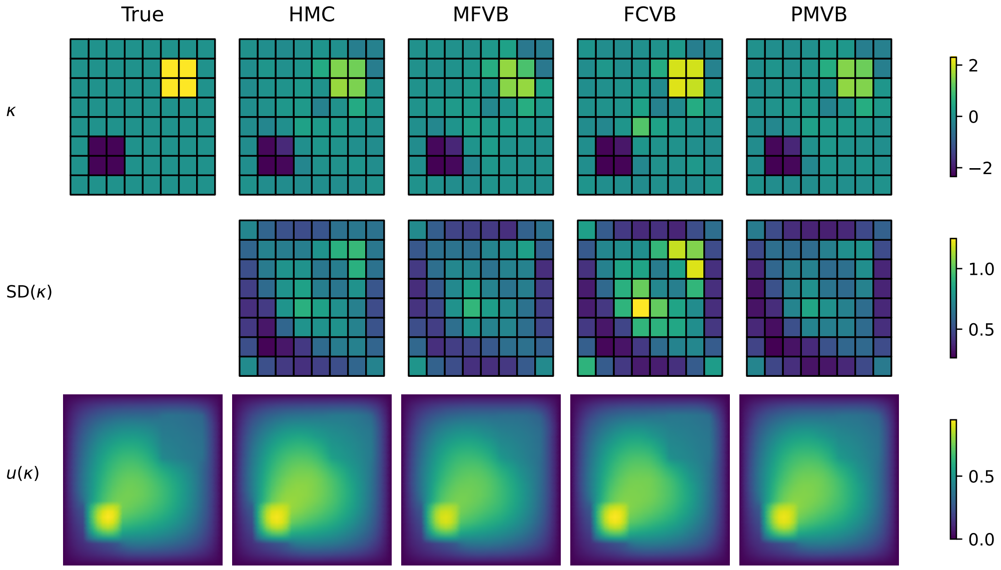


FIGURE 6.16: Posterior mean and standard deviation for $\boldsymbol{\kappa}$ and the corresponding $\mathbf{u}$ for the benchmark example with independent prior for each coefficient of $\boldsymbol{\kappa}$: $\kappa_i \sim \mathcal{N}(0, 1)$.

6.5 Conclusions

In this chapter, we have presented the variational inference framework for Bayesian inverse problems and investigated its efficacy on problems based on elliptic PDEs. Computationally, variational Bayes offers a tractable alternative to the intractable MCMC methods, and provides consistent mean and uncertainty estimates on the problems inspired by questions in computational mechanics. VB recasts the integration problem associated with Bayesian inference into an optimisation problem. As such, it is naturally integrated with existing FEM solvers, using the gradient calculations from the FEM solvers to optimise the ELBO in VB. Furthermore, the geometry of the problem encoded in the FEM mesh is utilised through the use of a sparse precision matrix that defines the conditional independence structure of the problem. Our results on the 1D and 2D Poisson problems support the claims of accuracy and scalability of VB. We note that the inferred variance is important in uncertainty quantification with a probabilistic forward model (for a different load case).

More specifically, our results show that

- the mean of the variational posterior provides an accurate point estimate irrespective of the choice of the parametrisation of the covariance structure,

- the variational approximation with a full-covariance or precision matrix structure adequately estimates posterior uncertainty when compared to HMC and pCN which are known to be asymptotically correct,
- parametrising the multivariate Gaussian distribution using a sparse precision matrix provides a way to balance the trade-off between computational complexity and the ability to capture dependencies in the posterior distribution,
- variational Bayes provides a good estimate for the mean and the variance of the posterior distribution in a time that is an order of magnitude faster than HMC or pCN,
- the multivariate Gaussian variational family is flexible enough to capture the true posterior distribution with high accuracy,
- the VB estimates may be used effectively in downstream tasks to estimate various quantities of interest, and
- variational Bayes method is flexible enough to model multimodal posteriors, as illustrated on a preliminary truss example, see Section B.2.

Our work may be extended in a number of natural ways that allows for greater adaptivity to the specific problems encountered in applications and integration within existing frameworks. Firstly, taking advantage of fast implementations of sparse linear algebra routines would further improve the scalability of VB with the structured precision matrix, as proposed in our work. Secondly, casting the inverse problem in a multi-level setting and taking advantage of low-dimensional projections has potential to further improve computational efficiency (Nagel & Sudret 2016, Ghattas & Willcox 2021). Thirdly, the results provided in this work use standard off-the-shelf optimisation routines; further computational improvements may be achieved using customised algorithms. As a further extension, in some applications it may be informative to consider the uncertainty in the forcing function so that the forward mapping is stochastic, as discussed in (Girolami et al. 2021). Finally, one of the aims of our work is to take advantage of the advances in Bayesian inference and adapt the novel algorithms to inverse problems in computational mechanics. As such, any further developments in VB as applied to machine learning and computational statistics problems may be directly applied using the framework proposed in this chapter.

6.6 Implementation

Codes for performing all forms of variational Bayes inference presented in this chapter are available on Github at `https://github.com/jp2011/bip-pde-vi`. The user must provide their own PDE solver which accepts $\boldsymbol{\kappa}$ as input parameter and computes $\log p(\boldsymbol{y} \mid \boldsymbol{\kappa})$, together with its gradient with respect to $\boldsymbol{\kappa}$.

Chapter 7

Summary and Further Work

Spatial information has played a crucial role in the development of models in science and engineering. In spatial statistics, spatial information is used to improve models by leveraging spatial context of observations. In natural sciences, many laws are formulated with respect to spatial location. Motivated by the importance of spatial information in modelling, we have made two methodological contributions, which we summarise in the next section. Afterwards, we discuss limitations and potential directions for future work.

7.1 Summary

In Chapter 4, we developed methodology for effectively modelling heterogeneous spatial point patterns over large domains such as cities. We considered estimation of the intensity of point patterns and analysed the factors contributing to its variation. Motivated by the application of burglary crime in Greater London, we proposed a model that accounts for spatial heterogeneity and imposes spatial dependence effectively. Events can refer to either residential or commercial burglaries, each with different empirical spatial patterns: commercial burglary occurrences cluster around the city centre, high streets and industrial parks, and residential burglary clusters around residential areas. This motivated our choice of a mixture model, which allows for different locations to be modelled by different components. The proposed Bayesian model is a finite mixture of Poisson generalised linear models such that each location is probabilistically assigned to one of the mixture components. Each component is characterised by the regression coefficients, which we used to interpret the localised effects of the covariates. By using a block structure of the study region, our approach allows specifying spatial dependence between nearby locations. We estimated the proposed model using Markov Chain Monte Carlo (MCMC) method, giving the posterior distribution of quantities of interest. Compared to log-Gaussian Cox processes, which are the go-to model for point patterns, the proposed model has better predictive performance, including the ability to predict hotspots, and can be estimated at a lower computational cost. In addition, the interpretability of the model components can provide operational insights.

The second contribution, presented in Chapter 6, concerns the inverse problem for models involving partial differential equations (PDEs). PDEs provide a mechanistic way for incorporating spatial dependence and spatial heterogeneity into models. The inverse problem involves assimilating observations, such as sensor measurements, into a PDE model to infer a physical parameter of the PDE. Such problems are generally ill-posed and must be regularised. Having chosen the Bayesian approach for the regularisation, we advocated for the use of variational Bayes methods as an alternative to Markov Chain Monte Carlo (MCMC) methods for inferring the posterior distribution of the physical parameter of a PDE. We showed that variational Bayes methods provide scalable inference that adequately propagates uncertainty. We proposed a family of Gaussian trial distributions parametrised by precision matrices, thus taking advantage of the inherent sparsity of the inverse problem encoded in its finite element discretisation. We utilised stochastic optimisation to efficiently estimate the variational objective and assess not only the error in the solution mean but also the uncertainty of the estimate. We performed an extensive empirical assessment on examples based on the Poisson equation, which is a fundamental model in science and engineering. The experiments included different regimes, such as different prior assumptions or varying the number of observations and measurement noise levels.

7.2 Limitations and Future work

This thesis has provided several new modelling approaches to leveraging spatial information. As with many modelling methods, the challenge lies in trading off model quality and cost. To reduce the cost, measured in computational terms, we had to make choices that imply a restricted or simplified model. Below, we summarise these choices, discuss their implications, and suggest several directions for future work.

In Chapter 4, we presented an inference algorithm for inferring point pattern intensities and their constitutive parts. One of the main limitations of our approach is the restriction to spatial-only settings. Although this choice was dictated by the low temporal resolution of the data, from an algorithmic standpoint, future work could develop the use of mixture models for modelling heterogeneous phenomena over spatio-temporal domains. Related to spatial heterogeneity, an approach different from mixture models could be taken. One possible candidate are spatially varying processes, proposed by Gelfand et al. (2003), but their scalability properties would have to be further investigated. To reduce the computational cost of our proposed method, we suggest several options. Firstly, one could impose Markovian structure in the mixture allocation component which is currently driven by K Gaussian processes, where K is the number of mixtures. The mixture allocation component could alternatively be modelled by K level sets of a single Gaussian process, thereby significantly reducing the dimension of the model. Lastly, alternative Bayesian inference methods such as variational Bayes methods could be explored.

Similarly to above, the methodology we presented in Chapter 6 is applied to linear elliptic PDEs in 2D. A natural extension would be to consider non-elliptic or non-linear PDEs, such as time-dependent heat equation. Although the methodology we developed is applicable in general, solving more complex PDEs which include non-linearity and time dimension poses its challenges. To further increase the flexibility of the proposed methodology so that it is applicable to a wider range of applications, one could assume that the input, function f, is itself a stochastic process. To tackle the problem of computational efficiency, several directions are possible. One could cast the inverse problem in a multi-level setting and solve the inference problem at multiple resolutions. This would reduce the computational cost at fine-resolution levels. Further improvements to the proposed methodology could come from alternative parametrisations of the approximation distribution. Another possible direction to take this work further is to characterise the bias introduced by the finite element discretisation, similarly to Papandreou et al. (2022). Lastly, one could investigate using more tailored optimisation schemes and linear algebra routines that better leverage sparse structure arising from the discretisation of PDEs using finite elements.

7.3 Concluding Remarks

Incorporating spatial information into models remains an important problem in science and engineering. Several challenges are at the forefront of the spatial modelling research. Firstly, the increasingly numerous and non-homogeneous sources of data, such as from the internet of things devices, call for novel methods to be developed to leverage the data. Secondly, as we have seen in this thesis, the computational cost of many spatial modelling methods is prohibitive, which renders the methods unable to cope with real-world problems. The ever-increasing computing capabilities will hopefully mean that many of the existing problems, and also new ones as a result of new data sources, will become more tractable. Recent successes of artificial intelligence research in which computational power played a crucial role suggest that we can expect that to be the case (Vinyals et al. 2019, Jumper et al. 2021). This is not to imply that domain expert knowledge is not essential to address the challenges above. On the contrary, expert knowledge should guide the modelling task at every step.

Bibliography

Abdulle, A. & Garegnani, G. (2021), 'A probabilistic finite element method based on random meshes: A posteriori error estimators and Bayesian inverse problems', *Computer Methods in Applied Mechanics and Engineering* **384**.

Abraham, K. & Nickl, R. (2020), 'On statistical Calderón problems', *Mathematical Statistics and Learning* **2**(2), 165–216.

Abramowitz, M. & Stegun, I. A. (1965), *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables*, Vol. 55, Courier Corporation.

Agresti, A. & Agresti, B. F. (1978), 'Statistical analysis of qualitative variation', *Sociological Methodology* **9**, 204–237.

Aldor-Noiman, S., Brown, L. D., Fox, E. B. & Stine, R. A. (2016), 'Spatio-temporal low count processes with application to violent crime events', *Statistica Sinica* **26**, 1587–1610.

Alvares, D., Armero, C. & Forte, A. (2018), 'What does objective mean in a Dirichlet-multinomial process?', *International Statistical Review* **86**(1), 106–118.

Andresen, M. A. (2010), The place of environmental criminology within criminological thought, *in* 'Classics in Environmental Criminology', CRC Press, pp. 21–44.

Anselin, L. (2010), 'Thirty years of spatial econometrics: Thirty years of spatial econometrics', *Papers in Regional Science* **89**(1), 3–25.

Anselin, L., Cohen, J., Cook, D., Gorr, W. & Tita, G. (2000), 'Spatial analyses of crime', *Criminal justice* **4**(2), 213–262.

Aristoff, D. & Bangerth, W. (2021), 'A benchmark for the Bayesian inversion of coefficients in partial differential equations'.

Arnst, M. & Soize, C. (2019), 'Identification and sampling of Bayesian posteriors of high-dimensional symmetric positive-definite matrices for data-driven updating of computational models', *Computer Methods in Applied Mechanics and Engineering* **352**, 300–323.

Asaadi, E. & Heyns, P. S. (2017), 'A computational framework for Bayesian inference in plasticity models characterisation', *Computer Methods in Applied Mechanics and Engineering* **321**, 455–481.

Babuška, I., Sawlan, Z., Scavino, M., Szabó, B. & Tempone, R. (2016), ‘Bayesian inference and model comparison for metallic fatigue data’, *Computer Methods in Applied Mechanics and Engineering* **304**, 171–196.

Baddeley, A. & Turner, R. (2005), ‘**Spatstat** : An *R* Package for Analyzing Spatial Point Patterns’, *Journal of Statistical Software* **12**(6).

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2015), *Hierarchical Modeling and Analysis for Spatial Data*, number 135 *in* ‘Monographs on Statistics and Applied Probability’, second edn, CRC Press, Taylor & Francis Group, Boca Raton.

Barajas-Solano, D. A. & Tartakovsky, A. M. (2019), ‘Approximate Bayesian model inversion for PDEs with heterogeneous and state-dependent coefficients’, *Journal of Computational Physics* **395**, 247–262.

Bauchau, O. A. & Craig, J. I. (2009), Basic equations of linear elasticity, *in* O. A. Bauchau, J. I. Craig & G. M. L. Gladwell, eds, ‘Structural Analysis’, Vol. 163, Springer Netherlands, Dordrecht, pp. 3–51.

Bayes, T. (1763), ‘An essay towards solving a problem in the doctrine of chances.’, *Philosophical Transactions of the Royal Society of London* **53**, 370–418.

Beavon, D. J., Brantingham, P. L. & Brantingham, P. J. (1994), ‘The influence of street networks on the patterning of property offenses’, *Crime prevention studies* **2**, 115–148.

Beck, J., Dia, B. M., Espath, L. F., Long, Q. & Tempone, R. (2018), ‘Fast Bayesian experimental design: Laplace-based importance sampling for the expected information gain’, *Computer Methods in Applied Mechanics and Engineering* **334**, 523–553.

Bernasco, W. (2014), Residential burglary, *in* G. Bruinsma & D. Weisburd, eds, ‘Encyclopedia of Criminology and Criminal Justice’, Springer New York, New York, NY, pp. 4381–4391.

Bernasco, W., Johnson, S. D. & Ruiter, S. (2015), ‘Learning where to offend: Effects of past on future burglary locations’, *Applied Geography* **60**, 120–129.

Bernasco, W. & Luykx, F. (2003), ‘Effects of attractiveness, opportunity, and accessibility to burglars on residential burglary rates of urban neighbourhoods’, *Criminology* **41**(3), 981–1002.

Bernasco, W. & Nieuwbeerta, P. (2005), ‘How do residential burglars select target areas?’, *The British Journal of Criminology* **45**(3), 296–315.

Beskos, A., Girolami, M., Lan, S., Farrell, P. E. & Stuart, A. M. (2017), ‘Geometric MCMC for infinite-dimensional inverse problems’, *Journal of Computational Physics* **335**, 327–351.

Betz, W., Papaioannou, I., Beck, J. L. & Straub, D. (2018), ‘Bayesian inference with subset simulation: Strategies and improvements’, *Computer Methods in Applied Mechanics and Engineering* **331**, 72–93.

Biegler, L. T., ed. (2007), *Real-Time PDE-constrained Optimization*, Computational Science & Engineering, Society for Industrial and Applied Mathematics, Philadelphia.

Bishop, C., Lawrence, N., Jaakkola, T. & Jordan, M. (1998), Approximating posterior distributions in belief networks using mixtures, *in* M. Jordan, M. Kearns & S. Solla, eds, 'Advances in Neural Information Processing Systems', Vol. 10, MIT Press.

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, Information Science and Statistics, Springer, New York.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017), 'Variational inference: A review for statisticians', *Journal of the American Statistical Association* **112**(518), 859–877.

Bowers, K. & Hirschfield, A. (1999), 'Exploring links between crime and disadvantage in north-west england: An analysis using geographical information systems', *International Journal of Geographical Information Science* **13**(2), 159–184.

Brantingham, P. & Brantingham, P. (1981), Notes on the geometry of crime, *in* 'Environmental Criminology', Sage Publications, Beverly Hills, CA.

Brantingham, P. J. & Brantingham, P. L. (1975), 'The spatial patterning of burglary', *The Howard Journal of Criminal Justice* **14**(2), 11–23.

Brantingham, P. L. & Brantingham, P. J. (1993), 'Nodes, paths and edges: Considerations on the complexity of crime and the physical environment', *Journal of Environmental Psychology* **13**(1), 3–28.

Breslow, N. E. (1984), 'Extra-Poisson variation in log-linear models', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **33**(1), 38–44.

Brunsdon, C., Fotheringham, A. S. & Charlton, M. E. (1996), 'Geographically weighted regression: A method for exploring spatial nonstationarity', *Geographical Analysis* **28**(4), 281–298.

Bui-Thanh, T., Ghattas, O., Martin, J. & Stadler, G. (2013), 'A computational framework for infinite-dimensional bayesian inverse problems part I: The linearized case, with application to global seismic inversion', *SIAM Journal on Scientific Computing* **35**(6), A2494–A2523.

Burt, D. R., Ober, S. W., Garriga-Alonso, A. & van der Wilk, M. (2021), Understanding variational inference in function-space, *in* 'Third Symposium on Advances in Approximate Bayesian Inference'.

Cameron, A. C. & Trivedi, P. K. (1990), 'Regression-based tests for overdispersion in the poisson model', *Journal of Econometrics* **46**(3), 347–364.

Carlon, A. G., Dia, B. M., Espath, L., Lopez, R. H. & Tempone, R. (2020), 'Nesterov-aided stochastic gradient methods using Laplace approximation for Bayesian design optimization', *Computer Methods in Applied Mechanics and Engineering* **363**.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. & Riddell, A. (2017), 'Stan : A probabilistic programming language', *Journal of Statistical Software* **76**(1), 1–32.

Celeux, G., Hurn, M. & Robert, C. P. (2000), 'Computational and inferential difficulties with mixture posterior distributions', *Journal of the American Statistical Association* **95**(451), 957–970.

Chainey, S., Tompson, L. & Uhlig, S. (2008), 'The utility of hotspot mapping for predicting spatial patterns of crime', *Security Journal* **21**(1-2), 4–28.

Chen, P., Villa, U. & Ghattas, O. (2017), 'Hessian-based adaptive sparse quadrature for infinite-dimensional Bayesian inverse problems', *Computer Methods in Applied Mechanics and Engineering* **327**, 147–172.

Cheng, C.-A. & Boots, B. (2017), Variational inference for Gaussian process models with linear complexity, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc.

Clare, J., Fernandez, J. & Morgan, F. (2009), 'Formal evaluation of the impact of barriers and connectors on residential burglars' macro-level offending location choices', *Australian & New Zealand Journal of Criminology* **42**(2), 139–158.

Clarke, R. V. & Cornish, D. B. (1985), 'Modeling offenders' decisions: A framework for research and policy', *Crime and Justice* **6**, 147–185.

Cohen, L. E. & Felson, M. (1979), 'Social change and crime rate trends: A routine activity approach', *American Sociological Review* **44**, 588–608.

Cotter, S. L., Roberts, G. O., Stuart, A. M. & White, D. (2013), 'MCMC methods for functions: Modifying old algorithms to make them faster', *Statistical Science* **28**(3), 424–446.

Cox, D. R. (1955), 'Some statistical methods connected with series of events', *Journal of the Royal Statistical Society. Series B (Methodological)* **17**(2), 129–164.

Cox, D. R. (2006), *Principles of Statistical Inference*, Cambridge University Press, Cambridge; New York.

Cressie, N. A. C. (1993), *Statistics for Spatial Data*, Wiley Series in Probability and Mathematical Statistics Applied Probability and Statistics, rev. edn, Wiley, New York.

Csató, L. & Opper, M. (2002), 'Sparse on-line Gaussian processes', *Neural Computation* **14**(3), 641–668.

Cui, T., Marzouk, Y. & Willcox, K. (2016), 'Scalable posterior approximations for large-scale Bayesian inverse problems via likelihood-informed parameter and state reduction', *Journal of Computational Physics* **315**, 363–387.

Cuthill, E. & McKee, J. (1969), Reducing the bandwidth of sparse symmetric matrices, *in* 'Proceedings of the 1969 24th National Conference', ACM '69, Association for Computing Machinery, New York, NY, USA, pp. 157–172.

Daley, D. J. & Vere-Jones, D. (2003), *An Introduction to the Theory of Point Processes*, 2nd edition edn, Springer-Verlag New York, New York.

Damianou, A. C., Titsias, M. K. & Lawrence, N. D. (2016), 'Variational inference for latent variables and uncertain inputs in Gaussian processes', *Journal of Machine Learning Research* **17**(42), 1–62.

Diggle, P. (1985), 'A kernel method for smoothing point process data', *Applied Statistics* **34**(2), 138.

Diggle, P. J., Moraga, P., Rowlingson, B. & Taylor, B. M. (2013), 'Spatial and spatio-temporal log-Gaussian Cox processes: Extending the geostatistical paradigm', *Statistical Science* **28**(4), 542–563.

Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987), 'Hybrid Monte Carlo', *Physics Letters B* **195**(2), 216–222.

Duffin, C., Cripps, E., Stemler, T. & Girolami, M. (2021), 'Statistical finite elements for misspecified models', *Proceedings of the National Academy of Sciences* **118**(2).

Durrande, N., Adam, V., Bordeaux, L., Eleftheriadis, S. & Hensman, J. (2019), Banded matrix operators for Gaussian Markov models in the automatic differentiation era, *in* K. Chaudhuri & M. Sugiyama, eds, 'Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics', Vol. 89 of *Proceedings of Machine Learning Research*, PMLR, pp. 2780–2789.

Efron, B. (1998), 'R. A. Fisher in the 21st century', *Statistical Science* **13**(2), 95–114.

Evans, D. J. (1989), Geographical analyses of residential burglary, *in* D. J. Evans & D. T. Herbert, eds, 'The Geography of Crime', Routledge, London, pp. 86–107.

Febrianto, E., Butler, L., Girolami, M. & Cirak, F. (2021), 'Digital twinning of self-sensing structures using the statistical finite element method'.

Felson, M. & Clarke, R. V. (1998), Opportunity makes the thief, Technical Report 98, Home Office, London.

Fernández, C. & Green, P. J. (2002), 'Modelling spatially correlated data via mixtures: A Bayesian approach', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 805–826.

Fisher, R. A. (1922), 'On the mathematical foundations of theoretical statistics', *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **222**(594-604), 309–368.

Fisher, R. A. (1934), 'Two new properties of mathematical likelihood', *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **144**(852), 285–307.

Flaxman, S., Chirico, M., Pereira, P. & Loeffler, C. (2019), 'Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ "real-time crime forecasting challenge"', *Ann. Appl. Stat.* **13**(4), 2564–2585.

Frühwirth-Schnatter, S., Celeux, G. & Robert, C. P., eds (2019), *Handbook of Mixture Analysis*, CRC Press, Boca Raton, Florida.

Gelfand, A. E., ed. (2010), *Handbook of Spatial Statistics*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, Boca Raton.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F. & Banerjee, S. (2003), 'Spatial modeling with spatially varying coefficient processes', *Journal of the American Statistical Association* **98**(462), 387–396.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian Data Analysis*, Chapman and Hall/CRC.

Geman, S. & Geman, D. (1984), 'Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**(6), 721–741.

Ghattas, O. & Willcox, K. (2021), 'Learning physics-based models from data: Perspectives from inverse problems and model reduction', *Acta Numerica* **30**, 445–554.

Giordano, M. & Nickl, R. (2020), 'Consistency of Bayesian inference with Gaussian process priors in an elliptic inverse problem', *Inverse Problems* **36**(8).

Giordano, R., Broderick, T. & Jordan, M. I. (2018), 'Covariances, robustness, and variational Bayes', *Journal of Machine Learning Research* **19**, 1981–2029.

Girolami, M. (2020), 'Introducing data-centric engineering: An open access journal dedicated to the transformation of engineering design and practice', *Data-Centric Engineering* **1**, e1.

Girolami, M. & Calderhead, B. (2011), 'Riemann manifold Langevin and Hamiltonian Monte Carlo methods', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(2), 123–214.

Girolami, M., Febrianto, E., Yin, G. & Cirak, F. (2021), 'The statistical finite element method (statFEM) for coherent synthesis of observation data and model predictions', *Computer Methods in Applied Mechanics and Engineering* **375**.

gov.uk (2017), 'Crime against businesses: Findings from the 2017 commercial victimisation survey'.

Green, P. J. (2010), Introduction to finite mixtures, *in* S. Frühwirth-Schnatter, G. Celeux & C. P. Robert, eds, 'Handbook of Spatial Statistics', Chapman & Hall/CRC Handbooks of Modern Statistical Methods, CRC Press, Boca Raton, Florida.

Green, P. J. & Richardson, S. (2002), 'Hidden Markov models and disease mapping', *Journal of the American Statistical Association* **97**(460), 1055–1070.

Grieves, M. (2015), Digital twin: Manufacturing excellence through virtual factory replication, Technical report, Department of Engineering Systems, Florida Institute of Technology.

Grün, B. & Leisch, F. (2008), Finite mixtures of generalized linear regression models, *in* Shalabh & C. Heumann, eds, 'Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg', Physica-Verlag HD, Heidelberg, pp. 205–230.

Hairer, M., Stuart, A. M. & Vollmer, S. J. (2014), 'Spectral gaps for a Metropolis–Hastings algorithm in infinite dimensions', *The Annals of Applied Probability* **24**(6).

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

Hawkes, A. G. (1971), 'Spectra of some self-exciting and mutually exciting point processes', *Biometrika* **58**(1), 83–90.

Hensman, J., Fusi, N. & Lawrence, N. D. (2013), Gaussian processes for big data, *in* 'Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence', UAI'13, AUAI Press, Arlington, Virginia, USA, pp. 282–290.

Hensman, J., Rattray, M. & Lawrence, N. D. (2012), Fast variational inference in the conjugate exponential family, *in* 'Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2', NIPS'12, Curran Associates Inc., Red Hook, NY, USA, pp. 2888–2896.

Hildeman, A., Bolin, D., Wallin, J. & Illian, J. B. (2018), 'Level set Cox processes', *Spatial Statistics* **28**, 169–193.

Hoffman, M. D., Blei, D. M., Wang, C. & Paisley, J. (2013), 'Stochastic variational inference', *Journal of Machine Learning Research* **14**, 1303–1347.

Hoffman, M. D. & Gelman, A. (2014), 'The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo', *Journal of Machine Learning Research* **15**(47), 1593–1623.

Huang, Y., Beck, J. L. & Li, H. (2017), 'Bayesian system identification based on hierarchical sparse Bayesian learning and Gibbs sampling with application to structural damage assessment', *Computer Methods in Applied Mechanics and Engineering* **318**, 382–411.

Huang, Y., Beck, J. L., Li, H. & Ren, Y. (2021), 'Sequential sparse Bayesian learning with applications to system identification for damage assessment and recursive reconstruction of image sequences', *Computer Methods in Applied Mechanics and Engineering* **373**.

Hunt, J. M. (2016), Do Crime Hot Spots Move? Exploring the Effects of the Modifiable Areal Unit Problem and Modifiable Temporal Unit Problem on Crime Hot Spot Stability, PhD thesis, American University, Washington, D.C.

Ibrahimbegovic, A., Matthies, H. G. & Karavelić, E. (2020), 'Reduced model of macro-scale stochastic plasticity identification by Bayesian inference: Application to quasi-brittle failure of concrete', *Computer Methods in Applied Mechanics and Engineering* **372**.

Jackson, J. D. (1999), *Classical Electrodynamics*, 3rd ed edn, Wiley, New York.

Jankowiak, M., Pleiss, G. & Gardner, J. (2020), Parametric Gaussian process regressors, *in* H. D. III & A. Singh, eds, 'Proceedings of the 37th International Conference on Machine Learning', Vol. 119 of *Proceedings of Machine Learning Research*, PMLR, pp. 4702–4712.

Jeffreys, H. (1998), *Theory of Probability*, Oxford Classic Texts in the Physical Sciences, 3rd ed edn, Clarendon Press ; Oxford University Press, Oxford [Oxfordshire] : New York.

Johnson, S. D. & Bowers, K. J. (2004), 'The stability of space-time clusters of burglary', *The British Journal of Criminology* **44**(1), 55–65.

Johnson, S. D. & Bowers, K. J. (2010), 'Permeability and burglary risk: Are cul-de-sacs safer?', *Journal of Quantitative Criminology* **26**(1), 89–111.

Johnson, S. D. & Summers, L. (2015), 'Testing ecological theories of offender spatial decision making using a discrete choice model', *Crime & Delinquency* **61**(3), 454–480.

Jones, D., Snider, C., Nassehi, A., Yon, J. & Hicks, B. (2020), 'Characterising the digital twin: A systematic literature review', *CIRP Journal of Manufacturing Science and Technology* **29**, 36–52.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1999), 'An introduction to variational methods for graphical models', *Machine Learning* **37**(2), 183–233.

Jordan, M. I. & Wainwright, M. J. (2007), 'Graphical models, exponential families, and variational inference', *Foundations and Trends® in Machine Learning* **1**(1–2), 1–305.

Jordan, R., Kinderlehrer, D. & Otto, F. (1998), 'The variational formulation of the Fokker-Planck equation', *Siam Journal on Mathematical Analysis* **29**(1), 1–17.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021), 'Highly accurate protein structure prediction with AlphaFold', *Nature* **596**(7873), 583–589.

Kaipio, J. & Somersalo, E. (2005), *Statistical and Computational Inverse Problems*, Springer, New York.

Karathanasopoulos, N., Angelikopoulos, P., Papadimitriou, C. & Koumoutsakos, P. (2017), 'Bayesian identification of the tendon fascicle's structural composition using finite element models for helical geometries', *Computer Methods in Applied Mechanics and Engineering* **313**, 744–758.

Kennedy, M. C. & O'Hagan, A. (2001), 'Bayesian calibration of computer models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63**(3), 425–464.

Kingma, D. P. & Ba, J. (2015), Adam: A method for stochastic optimization, *in* Y. Bengio & Y. LeCun, eds, 'International Conference on Learning Representations', San Diego, CA, USA.

Kingma, D. P. & Welling, M. (2014), Auto-encoding variational Bayes, *in* '2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings'.

Kleiber, C. & Zeileis, A. (2008), *Applied Econometrics with R*, Springer-Verlag, New York.

Knorr-Held, L. & Raßer, G. (2000), 'Bayesian detection of clusters and discontinuities in disease maps', *Biometrics* **56**(1), 13–21.

Krige, D. G. (1951), 'A statistical approach to some basic mine valuation problems on the witwatersrand', *Journal of the Southern African Institute of Mining and Metallurgy* **52**(6), 119–139.

Laplace, P.-S. (1812), *Théorie Analytique Des Probabilités*, Courcier, Paris.

Laub, P. J., Taimre, T. & Pollett, P. K. (2015), Hawkes processes, Technical report, University of Queensland, Brisbane, Australia.

Lax, P. D. & Milgram, A. N. (1955), *IX. Parabolic Equations*, Princeton University Press, pp. 167–190.

Lin, W., Butler, L. J., Elshafie, M. Z. E. B. & Middleton, C. R. (2019), 'Performance assessment of a newly constructed skewed half-through railway bridge using integrated sensing', *Journal of Bridge Engineering* **24**(1), 04018107.

Lindgren, F. & Rue, H. (2015), 'Bayesian spatial modelling with R-INLA', *Journal of Statistical Software* **63**(19).

Lindgren, F., Rue, H. & Lindström, J. (2011), '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.

Logg, A., Mardal, K.-A. & Wells, G., eds (2012), *Automated Solution of Differential Equations by the Finite Element Method*, Vol. 84 of *Lecture Notes in Computational Science and Engineering*, Springer Berlin Heidelberg, Berlin, Heidelberg.

Lord, G. J., Powell, C. E. & Shardlow, T. (2014), *An Introduction to Computational Stochastic PDEs*, number 50 *in* 'Cambridge Texts in Applied Mathematics', Cambridge University Press, New York, NY, USA.

Lu, C. & Tang, X. (2015), Surpassing human-level face verification performance on LFW with Gaussian face, *in* 'Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence', AAAI'15, AAAI Press, pp. 3811–3819.

Lu, Y., Stuart, A. & Weber, H. (2017), 'Gaussian approximations for probability measures on R^{d}', *SIAM/ASA Journal on Uncertainty Quantification* **5**(1), 1136–1165.

MacKay, D. J. C. (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.

McCullagh, P. & Nelder, J. A. (1998), *Generalized Linear Models*, number 37 *in* 'Monographs on Statistics and Applied Probability', 2nd ed edn, Chapman & Hall/CRC, Boca Raton.

Menting, B., Lammers, M., Ruiter, S. & Bernasco, W. (2019), 'The influence of activity space and visiting frequency on crime location choice: Findings from an online self-report survey', *The British Journal of Criminology* **60**(2), 303–322.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953), 'Equation of state calculations by fast computing machines', *The Journal of Chemical Physics* **21**(6), 1087–1092.

Michelén Ströfer, C. A., Zhang, X.-L., Xiao, H. & Coutier-Delgosha, O. (2020), 'Enforcing boundary conditions on physical fields in Bayesian inversion', *Computer Methods in Applied Mechanics and Engineering* **367**.

Minh, H. Q. (2017), 'Infinite-dimensional log-determinant divergences between positive definite trace class operators', *Linear Algebra and its Applications* **528**, 331–383.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. & Tita, G. E. (2011), 'Self-exciting point process modeling of crime', *Journal of the American Statistical Association* **106**(493), 100–108.

Møller, J., Syversveen, A. R. & Waagepetersen, R. P. (1998), 'Log Gaussian Cox processes', *Scandinavian Journal of Statistics* **25**(3), 451–482.

Møller, J. & Waagepetersen, R. P. (2004), *Statistical Inference and Simulation for Spatial Point Processes*, Chapman & Hall/CRC, Boca Raton.

Møller, J. & Waagepetersen, R. P. (2007), 'Modern statistics for spatial point processes*', *Scandinavian Journal of Statistics* **34**(4), 643–684.

Monard, F., Nickl, R. & Paternain, G. P. (2020), 'Statistical guarantees for Bayesian uncertainty quantification in non-linear inverse problems with Gaussian process priors'.

Nagel, J. B. & Sudret, B. (2016), 'A unified framework for multilevel uncertainty quantification in Bayesian inverse problems', *Probabilistic Engineering Mechanics* **43**, 68–84.

Neal, R. (2011), MCMC using Hamiltonian dynamics, *in* S. Brooks, A. Gelman, G. Jones & X.-L. Meng, eds, 'Handbook of Markov Chain Monte Carlo', Vol. 20116022, Chapman and Hall/CRC.

Ni, P., Li, J., Hao, H., Han, Q. & Du, X. (2021), 'Probabilistic model updating via variational Bayesian inference and adaptive Gaussian process modeling', *Computer Methods in Applied Mechanics and Engineering* **383**.

Office for National Statistics (2019), 'Census geography - Office for National Statistics'.

Ogata, Y. (1988), 'Statistical models for earthquake occurrences and residual analysis for point processes', *Journal of the American Statistical Association* **83**(401), 9–27.

Ordnance Survey (2018), 'Points of interest [CSV geospatial data], scale 1:1250, items: 670887'.

Pandita, P., Tsilifis, P., Awalgaonkar, N. M., Bilionis, I. & Panchal, J. (2021), 'Surrogate-based sequential Bayesian experimental design using non-stationary Gaussian processes', *Computer Methods in Applied Mechanics and Engineering* **385**.

Papandreou, Y., Cockayne, J., Girolami, M. & Duncan, A. B. (2022), 'Theoretical guarantees for the statistical finite element method'.

Pinski, F. J., Simpson, G., Stuart, A. M. & Weber, H. (2015), 'Algorithms for Kullback–Leibler approximation of probability measures in infinite dimensions', *SIAM Journal on Scientific Computing* **37**(6), A2733–A2757.

Piquero, A. R. & Weisburd, D. (2010), *Handbook of Quantitative Criminology*, Springer, New York ; London.

Pleiss, G., Gardner, J., Weinberger, K. & Wilson, A. G. (2018), Constant-time predictive distributions for Gaussian processes, *in* J. Dy & A. Krause, eds, 'Proceedings of the 35th International Conference on Machine Learning', Vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 4114–4123.

police.uk (2018), 'About | data.police.uk'.

police.uk (2019), 'Data downloads | data.police.uk'.

Povala, J., Kazlauskaite, I., Febrianto, E., Cirak, F. & Girolami, M. (2022), 'Variational Bayesian approximation of inverse problems using sparse precision matrices', *Computer Methods in Applied Mechanics and Engineering* **393**, 114712.

Povala, J., Virtanen, S. & Girolami, M. (2020), 'Burglary in London: Insights from statistical heterogeneous spatial point processes', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **69**(5), 1067–1090.

PredPol (2019), 'PredPol mission | about us | aiming to reduce victimization keep communities safer', https://www.predpol.com/.

Pyrialakos, S., Kalogeris, I., Sotiropoulos, G. & Papadopoulos, V. (2021), 'A neural network-aided Bayesian identification framework for multiscale modeling of nanocomposites', *Computer Methods in Applied Mechanics and Engineering* **384**.

Quiñonero-Candela, J. & Rasmussen, C. E. (2005), 'A unifying view of sparse approximate Gaussian process regression', *Journal of Machine Learning Research* **6**(65), 1939–1959.

Ranganath, R., Tran, D. & Blei, D. (2016), Hierarchical variational models, *in* M. F. Balcan & K. Q. Weinberger, eds, 'Proceedings of the 33rd International Conference on Machine Learning', Vol. 48 of *Proceedings of Machine Learning Research*, PMLR, New York, New York, USA, pp. 324–333.

Rasmussen, C. E. & Williams, C. K. I. (2006), *Gaussian Processes for Machine Learning*, Adaptive Computation and Machine Learning, MIT Press, Cambridge, Mass.

Reddi, S. J., Kale, S. & Kumar, S. (2018), On the convergence of Adam and beyond, *in* '6th International Conference on Learning Representations (ICLR)', Vancouver, BC,.

Rengert, G. F. & Wasilchick, J. (2010), The use of space in burglary, *in* M. A. Andresen, P. J. Brantingham & B. J. Kinney, eds, 'Classics in Environmental Criminology', CRC Press, pp. 257–272.

Ripley, B. D. (1976), 'The second-order analysis of stationary point processes', *Journal of Applied Probability* **13**(02), 255–266.

Ripley, B. D. (1977), 'Modelling spatial patterns', *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(2), 172–192.

Rizzi, F., Khalil, M., Jones, R., Templeton, J., Ostien, J. & Boyce, B. (2019), 'Bayesian modeling of inconsistent plastic response due to material variability', *Computer Methods in Applied Mechanics and Engineering* **353**, 183–200.

Roberts, G. O. & Rosenthal, J. S. (2004), 'General state space Markov chains and MCMC algorithms', *Probability Surveys* **1**(0), 20–71.

Rousseau, J. & Mengersen, K. (2011), 'Asymptotic behaviour of the posterior distribution in overfitted mixture models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(5), 689–710.

Rudolf, D. & Sprungk, B. (2018), ‘On a generalization of the preconditioned Crank–Nicolson Metropolis algorithm’, *Foundations of Computational Mathematics* **18**(2), 309–343.

Rue, H. & Held, L. (2005), *Gaussian Markov Random Fields: Theory and Applications*, number 104 *in* ‘Monographs on Statistics and Applied Probability’, Chapman & Hall/CRC, Boca Raton.

Rue, H., Martino, S. & Chopin, N. (2009), ‘Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2), 319–392.

Saatçi, Y. (2012), Scalable Inference for Structured Gaussian Process Models, PhD thesis, Citeseer.

Sabater, C., Le Maître, O., Congedo, P. M. & Görtz, S. (2021), ‘A Bayesian approach for quantile optimization problems with high-dimensional uncertainty sources’, *Computer Methods in Applied Mechanics and Engineering* **376**.

Salimbeni, H., Cheng, C.-A., Boots, B. & Deisenroth, M. (2018), Orthogonally decoupled variational Gaussian processes, *in* S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi & R. Garnett, eds, ‘Advances in Neural Information Processing Systems’, Vol. 31, Curran Associates, Inc.

Salimbeni, H. & Deisenroth, M. (2017), Doubly stochastic variational inference for deep Gaussian processes, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, eds, ‘Advances in Neural Information Processing Systems’, Vol. 30, Curran Associates, Inc.

Sampson, R. J. & Groves, W. B. (1989), ‘Community structure and crime: Testing social-disorganization theory’, *American Journal of Sociology* **94**(4), 774–802.

Sampson, R. J., Raudenbush, S. W. & Earls, F. (1997), ‘Neighborhoods and violent crime: A multilevel study of collective efficacy’, *Science* **277**(5328), 918–924.

Seeger, M. W., Williams, C. K. I. & Lawrence, N. D. (2003), Fast forward selection to speed up sparse Gaussian process regression, *in* C. M. Bishop & B. J. Frey, eds, ‘Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics’, Vol. R4 of *Proceedings of Machine Learning Research*, PMLR, pp. 254–261.

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Díaz-Ávalos, C. & Rue, H. (2014), ‘Spatio-temporal log-Gaussian Cox processes for modelling wildfire occurrence: The case of Catalonia, 1994–2008’, *Environmental and Ecological Statistics* **21**(3), 531–563.

Shaw, C. R. & McKay, H. D. (1942), *Juvenile Delinquency and Urban Areas : A Study of Rates of Delinquents in Relation to Differential Characteristics of Local Communities in American Cities*, Chicago, Ill. : The University of Chicago Press.

Shi, J., Titsias, M. & Mnih, A. (2020), Sparse orthogonal variational inference for Gaussian processes, *in* S. Chiappa & R. Calandra, eds, ‘Proceedings of the Twenty Third International

Conference on Artificial Intelligence and Statistics', Vol. 108 of *Proceedings of Machine Learning Research*, PMLR, pp. 1932–1942.

Simpson, D., Illian, J. B., Lindgren, F., Sørbye, S. H. & Rue, H. (2016), 'Going off grid: Computationally efficient inference for log-Gaussian Cox processes', *Biometrika* **103**(1), 49–70.

Smith, K., Taylor, P. & Elkin, M. (2013), Crimes detected in england and wales 2012/13, Statistical Bulletin 02/13, Home Office, London.

Snelson, E. & Ghahramani, Z. (2006), Sparse Gaussian processes using pseudo-inputs, *in* Y. Weiss, B. Schölkopf & J. Platt, eds, 'Advances in Neural Information Processing Systems', Vol. 18, MIT Press.

Solin, A., Cortes, S., Rahtu, E. & Kannala, J. (2018), PIVO: Probabilistic Inertial-Visual Odometry for Occlusion-Robust Navigation, *in* '2018 IEEE Winter Conference on Applications of Computer Vision (WACV)', IEEE, pp. 616–625.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging.*, Springer, Place of publication not identified.

Stoyan, D. & Stoyan, H. (1994), *Fractals, Random Shapes, and Point Fields: Methods of Geometrical Statistics*, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Strang, G. & Fix, G. J. (2008), *An Analysis of the Finite Element Method*, 2. ed edn, Wellesley-Cambridge Press, Wellesley, Mass.

Stuart, A. M. (2010), 'Inverse problems: A Bayesian perspective', *Acta Numerica* **19**, 451–559.

Sun, S., Zhang, G., Shi, J. & Grosse, R. B. (2019), Functional variational Bayesian neural networks, *in* '7th International Conference on Learning Representations', New Orleans, LA, USA.

Taddy, M. A. (2010), 'Autoregressive mixture models for dynamic spatial poisson processes: Application to tracking intensity of violent crime', *Journal of the American Statistical Association* **105**(492), 1403–1417.

Tan, L. S. L. & Nott, D. J. (2018), 'Gaussian variational approximation with sparse precision matrices', *Statistics and Computing* **28**(2), 259–275.

Tarakanov, A. & Elsheikh, A. H. (2020), 'Optimal Bayesian experimental design for subsurface flow problems', *Computer Methods in Applied Mechanics and Engineering* **370**.

Tarantola, A. (2005), *Inverse Problem Theory and Methods for Model Paramenter Estimation*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Tikhonov, A. N. & Arsenin, V. Y. (1977), *Solutions of Ill-Posed Problems*, Scripta Series in Mathematics, Winston.

Titsias, M. (2008), Variational model selection for sparse Gaussian process regression, Technical report, School of Computer Science, University of Manchester.

Titsias, M. (2009), Variational learning of inducing variables in sparse Gaussian processes, *in* D. van Dyk & M. Welling, eds, 'Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics', Vol. 5 of *Proceedings of Machine Learning Research*, PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, pp. 567–574.

Tobler, W. R. (1970), 'A computer movie simulating urban growth in the Detroit region', *Economic Geography* **46**, 234–240.

Tompson, L., Johnson, S., Ashby, M., Perkins, C. & Edwards, P. (2015), 'UK open source crime data: Accuracy and possibilities for research', *Cartography and Geographic Information Science* **42**(2), 97–111.

Townsley, M., Birks, D., Bernasco, W., Ruiter, S., Johnson, S. D., White, G. & Baum, S. (2015), 'Burglar target selection: A cross-national comparison', *Journal of Research in Crime and Delinquency* **52**(1), 3–31.

Townsley, M., Birks, D., Ruiter, S., Bernasco, W. & White, G. (2016), 'Target selection models with preference variation between offenders', *Journal of Quantitative Criminology* **32**(2), 283–304.

Tran, D., Blei, D. & Airoldi, E. M. (2015), Copula variational inference, *in* C. Cortes, N. Lawrence, D. Lee, M. Sugiyama & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 28, Curran Associates, Inc.

Tsilifis, P., Bilionis, I., Katsounaros, I. & Zabaras, N. (2016), 'Computationally efficient variational approximations for Bayesian inverse problems', *Journal of Verification, Validation and Uncertainty Quantification* **1**(3).

Turner, R. E. & Sahani, M. (2011), Two problems with variational expectation maximisation for time series models, *in* A. T. Cemgil, D. Barber & S. Chiappa, eds, 'Bayesian Time Series Models', Cambridge University Press, Cambridge, pp. 104–124.

Uribe, F., Papaioannou, I., Betz, W. & Straub, D. (2020), 'Bayesian inference of random fields represented with the Karhunen–Loève expansion', *Computer Methods in Applied Mechanics and Engineering* **358**.

van der Bles, A. M., van der Linden, S., Freeman, A. L. J., Mitchell, J., Galvao, A. B., Zaval, L. & Spiegelhalter, D. J. (2019), 'Communicating uncertainty about facts, numbers and science', *Royal Society Open Science* **6**(5), 181870.

van Lieshout, M.-C. (2010), Spatial point process theory, *in* S. Frühwirth-Schnatter, G. Celeux & C. P. Robert, eds, 'Handbook of Spatial Statistics', CRC Press, Boca Raton, Florida.

Villa, U., Petra, N. & Ghattas, O. (2021), 'hIPPYlib: An extensible software framework for large-scale inverse problems governed by PDEs: Part I: Deterministic inversion and linearized Bayesian inference', *ACM Transactions on Mathematical Software* **47**(2), 1–34.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., Wünsch, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C. & Silver, D. (2019), 'Grandmaster level in StarCraft II using multi-agent reinforcement learning', *Nature* **575**(7782), 350–354.

Wang, B. & Titterington, D. M. (2005), Inadequacy of interval estimates corresponding to variational Bayesian approximations, *in* R. G. Cowell & Z. Ghahramani, eds, 'Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics', Vol. R5 of *Proceedings of Machine Learning Research*, PMLR, pp. 373–380.

Wang, H. F. & Anderson, M. P. (2014), *Introduction to Groundwater Modeling: Finite Difference and Finite Element Methods.*, Elsevier Science.

Wang, Y. & Blei, D. M. (2019), 'Frequentist consistency of variational Bayes', *Journal of the American Statistical Association* **114**(527), 1147–1161.

Williams, C. & Seeger, M. (2001), Using the Nyström method to speed up kernel machines, *in* T. Leen, T. Dietterich & V. Tresp, eds, 'Advances in Neural Information Processing Systems', Vol. 13, MIT Press.

Wu, L., Zulueta, K., Major, Z., Arriaga, A. & Noels, L. (2020), 'Bayesian inference of non-linear multiscale model parameters accelerated by a deep neural network', *Computer Methods in Applied Mechanics and Engineering* **360**.

Young, G. A. & Smith, R. L. (2005), *Essentials of Statistical Inference: G.A. Young, R.L. Smith.*, Cambridge University Press, Cambridge, UK; New York.

Zhang, C., Butepage, J., Kjellstrom, H. & Mandt, S. (2019), 'Advances in variational inference', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(8), 2008–2026.

Appendix A

Supplementary Material for Chapter 4

A.1 Poisson Regression Model: Excess of Zeros, Overdispersion

In this section we demonstrate that the standard Poisson regression (McCullagh & Nelder 1998) is not a suitable model for the London burglary point pattern.

Firstly, the dataset consists of areas with no buildings in it, e.g. parks, airports, which results in counts equal to zero due to structure rather than due to chance. This is further supported by the plot of the observed count and the corresponding histogram, both shown in Figure A.1. This phenomenon is often referred to as *excess of zeros*.

Secondly, we fit Poisson GLM with all four specifications of covariates to the 2015 burglary dataset, as described in the paper. Then we use the overdispersion test proposed in Cameron & Trivedi (1990), and implemented in the AER package (Kleiber & Zeileis 2008). For the standard Poisson GLM model, $\text{Var}(y_n) = \mu_n$. The overdispersion test uses it as the null hypothesis, where the alternative is $\text{Var}(y_n) = \mu_n + c \times g(\mu_n)$, where $g(\cdot)$ must be specified. For our test, we choose $g(\cdot) = 1$. Table A.1 shows the estimated c values and the p-values for each estimate, given that null hypothesis is $c = 0$. The data clearly show the presence of overdispersion in all four models.

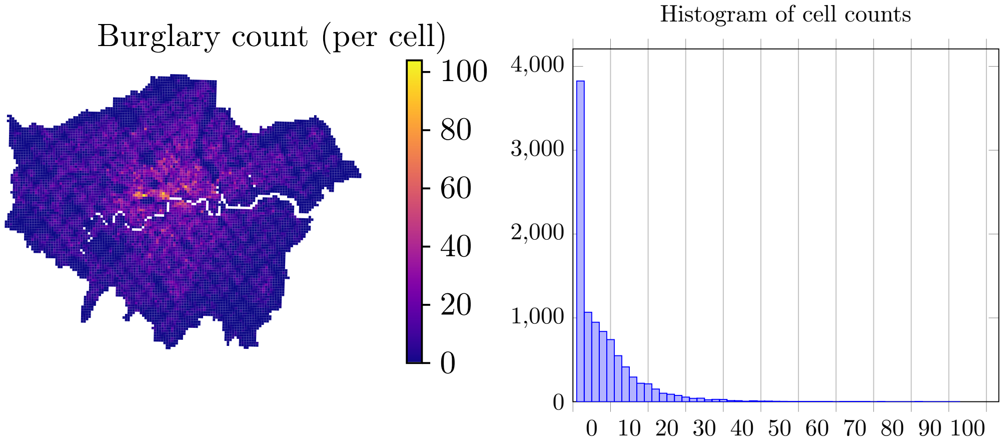


FIGURE A.1: Observed count on the map (left) and the corresponding histogram (right) for the point pattern of burglary aggregated over the grid for the time period 1/2015-12/2015.

TABLE A.1: Overdispersion test for Poisson GLM model of burglary counts.

Specification	c	p-value
1	1.905	2.2e-16
2	1.897	2.2e-16
3	1.910	2.2e-16
4	1.911	2.2e-16

A.1.1 Poisson Regression vs SAM-GLM

Figure A.2 shows the scatter plot of expected vs observed counts for the Poisson regression model (SAM-GLM with $K = 1$) and the proposed model with $K = 3$. It is evident from the plot that adding extra components to the standard Poisson regression reduces the overdispersion issue.

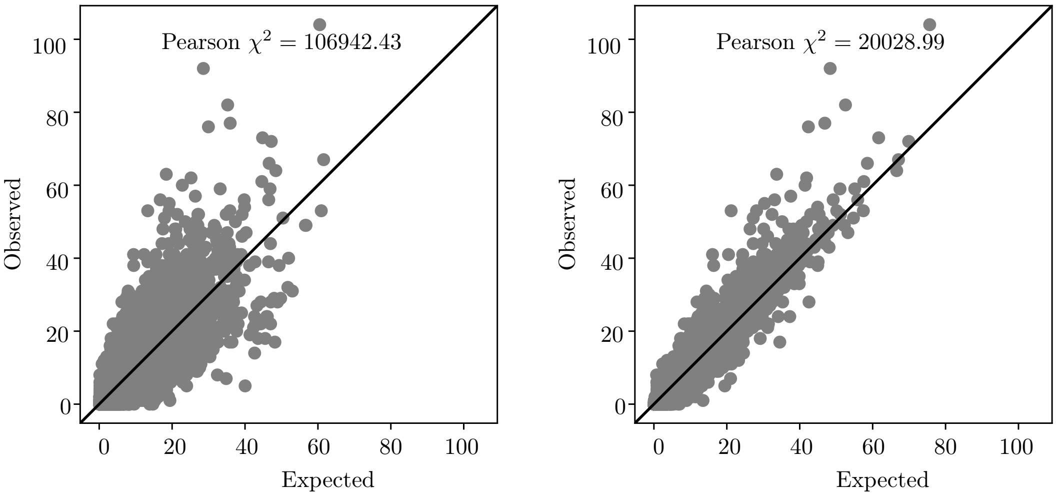


FIGURE A.2: Scatter plot of predicted counts vs observed counts (training data) for the Poisson GLM model (left), and SAM-GLM K=3 (right). Blocking: MSOA, training data: 2015, using specification 4.

A.2 Log-Gaussian Cox Process

Dicretising the spatial domain to a regular grid, the full Bayesian formulation of the model is given as follows:

$$y_n|\boldsymbol{\beta}, \mathbf{f}, \boldsymbol{X} \sim \text{Poisson}\left(\exp(\boldsymbol{X}_n^\top \boldsymbol{\beta} + f_n)\right) \tag{A.1}$$

$$f(\cdot)|\boldsymbol{\theta} \sim \mathcal{GP}\left(0, k_{\boldsymbol{\theta}}(\cdot,\cdot)\right) \tag{A.2}$$

$$\beta_j \sim \mathcal{N}(0, \sigma_j^2) \tag{A.3}$$

$$\sigma_j^2 \sim \text{InvGamma}(1, 0.01) \tag{A.4}$$

$$\boldsymbol{\theta} \sim \text{weakly-informative log-normal prior,} \tag{A.5}$$

where $n = 1, \ldots, N$ is the index over the cells on the map, $j = 1, \ldots, J$ is the index over the covariates, $f()$ is a zero-mean Gaussian process with covariance function $k_{\boldsymbol{\theta}}(\cdot,\cdot)$, and hyperparameters $\boldsymbol{\theta}$, f_n is the value of $f(\cdot)$ in the centre of cell n, $\boldsymbol{X}_n$ is the vector of the covariates at cell n, and $\boldsymbol{\beta}_j$ is the jth regression coefficient with a scale hyperparameter σ_{kj}^2. A plain Poisson generalised linear model (GLM) formulation assumes no spatial correlation, i.e. $f_n = 0$ for all n. Compared to the Poisson GLM model, LGCP allows for modelling the variation in the intensity that cannot be explained by the covariates $\boldsymbol{X}$.

In order to allow for Kronecker product factorisation of the covariance matrix of the Gaussian process, we specify $k_{\boldsymbol{\theta}}(\cdot,\cdot)$ as a product of two Matérn covariance functions, one for the easting (E) coordinate, the other for the northing (N) coordinate. If the GP is trained using n points of a regular grid, the computational cost of inverting the covariance matrix goes down from $\mathcal{O}\big(n^3\big)$ to $\mathcal{O}\big(2n^{3/2}\big)$.

Matérn covariance function is a standard choice in spatial statistics as it allows specifying smoothness of the function (Stein 1999). It is given as follows

$$k_{\mathrm{Matern}}(\boldsymbol{x},\boldsymbol{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}|\boldsymbol{x}-\boldsymbol{x}'|}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}|\boldsymbol{x}-\boldsymbol{x}'|}{\ell}\right), \tag{A.6}$$

where ℓ is the characteristic lengthscale, ν is the smoothness parameter, and K_ν is a modified Bessel function (Rasmussen & Williams 2006). It can be shown that that the Gaussian processes with Matérn covariance functions are k-times mean-square differentiable if and only if $\nu > k$. Abramowitz & Stegun (1965) show that if ν is a half-integer, i.e. for an integer p, $\nu = p + \frac{1}{2}$, the covariance function becomes especially simple, giving

$$k_{\ell,\nu=p+1/2}(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\frac{\sqrt{2\nu}|\boldsymbol{x}-\boldsymbol{x}'|}{\ell}\right)\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\left(\frac{\sqrt{8\nu}|\boldsymbol{x}-\boldsymbol{x}'|}{\ell}\right)^{p-i}. \tag{A.7}$$

For this reason, we set $\nu = 3/2$. The final covariance function, including the σ^2 parameter to control the range of $f()$ therefore becomes

$$k_{\boldsymbol{\theta}}((x_{\mathrm{E}},x_{\mathrm{N}}),(y_{\mathrm{E}},y_{\mathrm{N}})) \quad = \quad \sigma^2 k_{\ell,\nu=3/2}(x_{\mathrm{E}},y_{\mathrm{E}}) \times k_{\ell,\nu=3/2}(x_{\mathrm{N}},y_{\mathrm{N}}), \tag{A.8}$$

where $\boldsymbol{\theta} = [\sigma^2,\ell]^\top$.

A.2.1 Inference

To infer posterior distribution of the regression coefficients, $\boldsymbol{\beta}$, latent field $\mathbf{f}$, and its hyperparameters $\boldsymbol{\theta}$, we use a Hamiltonian Monte Carlo sampler. The scale parameters $\sigma_1^2,\ldots,\sigma_J^2$ are analytically integrated out (see (A.14) in the appendix). Due to positivity constraint of the hyperparameters, we sample from $\boldsymbol{\phi} = \log\boldsymbol{\theta}$ (applied component-wise).

The density function of the joint posterior distribution we are sampling from is proportional to the product of likelihood and the priors, i.e.

$$p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\phi}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta})p(\mathbf{f}|\exp(\boldsymbol{\phi}))p(\boldsymbol{\beta})p_{\boldsymbol{\theta}}(\exp(\boldsymbol{\phi})) \prod_i \left| \frac{d}{d\phi_i} \exp(\phi_i) \right| . \tag{A.9}$$

To effectively use HMC sampler, log-likelihood of the posterior and its gradient need to be tractable. Thanks to the grid structure of our study region, we utilise Kronecker product structure that is present in the covariate matrix in $p(\mathbf{f}|\boldsymbol{\theta})$ if the covariance function $k_{\boldsymbol{\theta}}(\cdot, \cdot)$ is assumed to be a product of covariance functions, one per each dimension (For more details, see Saatçi (2012)). After expansion, the unnormalised log-density becomes

$$\begin{aligned} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\phi}|\mathbf{y}) &= \log p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) + \log p(\boldsymbol{\beta}) + \log p(\mathbf{f}|\exp(\boldsymbol{\phi})) + \log p_{\boldsymbol{\theta}}(\exp(\boldsymbol{\phi})) + \sum_i \phi_i + \text{const}_1 \\ &= \left(\mathbf{y}^\top \boldsymbol{X\beta} + \mathbf{y}^\top \mathbf{f} - \exp(\boldsymbol{X\beta} + \mathbf{f}) \right) + \log p(\boldsymbol{\beta}) \\ &\quad + \left(-\frac{1}{2} \log |\boldsymbol{K_\theta}| - \frac{1}{2} \mathbf{f}^\top \boldsymbol{K_\theta}^{-1} \mathbf{f} \right) + \log p_{\boldsymbol{\theta}}(\exp(\boldsymbol{\phi})) + \sum_i \phi_i + \text{const}_1, \end{aligned} \tag{A.10}$$

The gradients of the log posterior density w.r.t. quantities of interest are

$$\nabla_{\mathbf{f}} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\phi}|\mathbf{y}) = (\mathbf{y} - \exp(\boldsymbol{X\beta} + \mathbf{f})) + \left(-\boldsymbol{K_\theta}^{-1} \mathbf{f} \right) \tag{A.11}$$

$$\nabla_{\boldsymbol{\beta}} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\phi}|\mathbf{y}) = \left(\boldsymbol{X}^\top \mathbf{y} - \boldsymbol{X}^\top \exp(\boldsymbol{X\beta} + \mathbf{f}) \right) + \nabla_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}) \tag{A.12}$$

$$\begin{aligned} \nabla_{\phi_i} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\phi}|\mathbf{y}) &= \frac{1}{2} \mathbf{f}^\top \boldsymbol{K_\theta}^{-1} \frac{\partial \boldsymbol{K_\theta}}{\partial \theta_i} \boldsymbol{K_\theta}^{-1} \mathbf{f} - \frac{1}{2} \operatorname{tr} \left(\boldsymbol{K_\theta}^{-1} \frac{\partial \boldsymbol{K_\theta}}{\partial \theta_i} \right) \\ &\quad + \nabla_{\phi_i} \log p_{\boldsymbol{\theta}}(\exp(\boldsymbol{\phi})) + 1. \end{aligned} \tag{A.13}$$

The expansion of un-normalised log-density of $\boldsymbol{\beta}$ and the gradients are derived in (A.15) and (A.16) below.

All operations involving $\boldsymbol{K_\theta}$ can be sped up using Kronecker product factorisation. Given n^2 is the number of elements in the full matrix $\boldsymbol{K_\theta}$, operations in (A.11) and (A.13) can be computed in $\mathcal{O}\left(n^{\frac{3}{2}}\right)$ time by utilising the Kronecker structure in matrix inversion and matrix-vector multiplication. For full details, see Saatçi (2012).

Lastly, we note that the restriction of performing inference on a grid-based domain has been relaxed by non-grid approaches based on SPDEs proposed (Simpson et al. 2016, Lindgren et al. 2011).

A.3 Model Derivations

A.3.1 Beta Prior

Given a vector of J independent random variables $\boldsymbol{\beta}$, of which each component is distributed as follows

$$\begin{aligned}\beta_j &\sim \mathcal{N}(0, \sigma_j^2),\\ \sigma_j^2 &\sim \text{InvGamma}(a, b).\end{aligned}$$

Let $\Psi = \left(\sigma_1^2, \ldots, \sigma_J^2\right)^\top$, then the prior for the coefficients is given by integrating out the nuisance parameter Ψ

$$\begin{aligned}p(\boldsymbol{\beta}) &= \prod_j p(\beta_j)\\ &= \prod_j \int p(\beta_j|\Psi_j)p(\Psi_j)d\Psi_j\\ &= \prod_j \int \frac{1}{\sqrt{2\pi}}\Psi_j^{-1/2}\exp\left(-\frac{1}{2\Psi_j}\beta_j^2\right)\frac{b^a}{\Gamma(a)}\Psi_j^{-a-1}\exp\left(-\frac{b}{\Psi_j}\right)d\Psi_j\\ &= \prod_j \frac{b^a}{\sqrt{2\pi}\Gamma(a)}\int \Psi_j^{-a-\frac{1}{2}-1}\exp\left(-\frac{\frac{1}{2}\beta_j^2+b}{\Psi_j}\right)d\Psi_j\\ &= \prod_j \frac{b^a}{\sqrt{2\pi}\Gamma(a)}\frac{\Gamma\left(\frac{1}{2}+a\right)}{\left(\frac{1}{2}\beta_j^2+b\right)^{\frac{1}{2}+a}}\end{aligned} \tag{A.14}$$

For the purposes of HMC, we derive both log-density and the gradient of log-density w.r.t. the each individual components. Log-density is given as

$$\log p(\boldsymbol{\beta}) = \sum_i -\left(\frac{1}{2}+a\right)\log\left(\frac{1}{2}\beta_i^2+b\right), \tag{A.15}$$

from which the gradient is equal to

$$\frac{\partial \log p(\boldsymbol{\beta})}{\partial \beta_i} = \frac{(-\frac{1}{2}-a)\beta_i}{\frac{1}{2}\beta_i^2+b}. \tag{A.16}$$

A.3.2 Conditional Densities for SAM-GLM Inference

The derivations below use the properties of the density function of the Dirichlet distribution and the following property of the Gamma function, $\Gamma(a+1) = a\Gamma(a)$.

A.3.2.1 Regression Coefficients Update

$$\begin{aligned} p(\boldsymbol{\beta}|\alpha, \boldsymbol{X}, \mathbf{y}, \mathbf{z}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{X}, \mathbf{z})p(\boldsymbol{\beta}) \\ &\propto \left\{\prod_{k=1}^{K}\prod_{j=1}^{J} p(\beta_{k,j})\right\}\left\{\prod_{n=1}^{N} p(y_n|\boldsymbol{\beta}, \boldsymbol{X}, z_n)\right\} \\ &\propto \left\{\prod_{k=1}^{K}\prod_{j=1}^{J} p(\beta_{k,j})\right\}\left\{\prod_{n=1}^{N}\prod_{k=1}^{K} p(y_n|\boldsymbol{\beta}_k, \boldsymbol{X})^{I(z_n=k)}\right\} \\ &\propto \left\{\prod_{k=1}^{K}\prod_{j=1}^{J} p(\beta_{k,j})\right\}\left\{\prod_{n=1}^{N}\prod_{k=1}^{K} \left(\frac{\exp(\boldsymbol{X}_n^\top\boldsymbol{\beta}_k)^{y_n}\mathrm{e}^{-\exp(\boldsymbol{X}_n^\top\boldsymbol{\beta}_k)}}{y_n!}\right)^{I(z_n=k)}\right\}, \end{aligned} \tag{A.17}$$

where $p(\boldsymbol{\beta})$ is expanded according to (A.14). For the purposes of Hamiltonian Monte Carlo, the gradient of the posterior distribution is analytically available.

A.3.2.2 GPs Updates

The unnormalised joint posterior density of the K GPs and their hyperparameters is given as

$$\begin{aligned} p(\boldsymbol{F}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) &\propto p(\mathbf{z}|\boldsymbol{F})p(\boldsymbol{F}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto \prod_{n=1}^{N} p(z_n|\boldsymbol{F}) \prod_{k=1}^{K} p(f_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k) \\ &\propto \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\frac{\exp(f_{k,b[n]})}{\sum_{l=1}^{K}\exp(f_{l,b[n]})}\right)^{I(z_n=k)} \prod_{k=1}^{K} p(f_k|\boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k), \end{aligned}$$

where $p(f_k|\boldsymbol{\theta}_k)$ is the density function of the zero-mean multivariate Gaussian distribution with covariance matrix parameterised by $\boldsymbol{\theta}$, and $p(\boldsymbol{\theta}_k)$ is a suitable prior for the hyperparamers. The gradient of the joint posterior with respect to $\boldsymbol{F}$ and $\boldsymbol{\theta}$ are analytically available.

A.3.2.3 Mixture Allocation Update for Spatially-dependent Blocks

$$
\begin{aligned}
p(z_n = k|\mathbf{z}^{\bar{n}}, \boldsymbol{X}_n, \boldsymbol{\beta}, \mathbf{y}, \boldsymbol{F}) &= p(y_n|z_n = k, \boldsymbol{X}_n, \boldsymbol{\beta}_k)p(z_n|\boldsymbol{F}) \\
&\propto p(y_n|z_n = k, \boldsymbol{X}_n, \boldsymbol{\beta}_k)\frac{\exp(f_{k,b[n]})}{\sum_{l=1}^{K}\exp(f_{l,b[n]})} \\
&= \prod_{k=1}^{K}\left(\frac{\exp(\boldsymbol{X}_n^\top \boldsymbol{\beta}_k)^{y_n}\mathrm{e}^{-\exp(\boldsymbol{X}_n^\top \boldsymbol{\beta}_k)}}{y_n!}\right)^{I(z_n=k)} \frac{\exp(f_{k,b[n]})}{\sum_{l=1}^{K}\exp(f_{l,b[n]})}
\end{aligned}
$$

A.3.2.4 Mixture Allocation Update for Independent Blocks

$$
\begin{aligned}
p(z_n = k|\mathbf{z}^{\bar{n}}, \alpha, \boldsymbol{X}_n, \boldsymbol{\beta}, \mathbf{y}) &\propto p(y_n|z_n = k, \boldsymbol{X}_n, \boldsymbol{\beta})\int p(z_n|\boldsymbol{\pi}_{b[n]})p(\boldsymbol{\pi}_{b[n]}|\alpha, \mathbf{z}^{\bar{n}})d\boldsymbol{\pi}_{b[n]} \\
&\propto p(y_n|z_n = k, \boldsymbol{X}_n, \boldsymbol{\beta})\int \prod_{k}\pi_{b[n],k}^{I(z_n=k)}\frac{\Gamma(\sum_{j=1}^{K}B_{b[n],j})}{\prod_{j=1}^{K}\Gamma(B_{b[n],j})}\prod_{j=1}^{K}\pi_{b[n],j}^{B_{b[n],j}-1}d\boldsymbol{\pi}_{b[n]} \\
&\propto p(y_n|z_n = k, \boldsymbol{X}_n, \boldsymbol{\beta})\frac{\Gamma(\sum_{j=1}^{K}B_{b[n],j})}{\prod_{j=1}^{K}\Gamma(B_{b[n],j})}\frac{\prod_{j=1}^{K}\Gamma(B_{b[n],j}+I(j=k))}{\Gamma(\sum_{j=1}^{K}B_{b[n],j}+I(j=k))} \\
&\propto p(y_n|z_n = k, \boldsymbol{X}_n, \boldsymbol{\beta})\frac{B_{b[n],k}}{\sum_{j=1}^{K}B_{b[n],j}} \\
&\propto \prod_{k=1}^{K}\left(\frac{\exp(\boldsymbol{X}_n^\top \boldsymbol{\beta}_k)^{y_n}\mathrm{e}^{-\exp(\boldsymbol{X}_n^\top \boldsymbol{\beta}_k)}}{y_n!}\right)^{I(z_n=k)} \frac{c_{b[n]k}^{\bar{n}}+\alpha}{K\alpha+\sum_{j=1}^{K}c_{b[n]j}^{\bar{n}}},
\end{aligned}
\tag{A.18}
$$

where $B_{b,k} = c_{b,k}^{\bar{n}} + \alpha$, and $c_{b,k}^{\bar{n}}$ is the number of cells in block b other than cell n that are assigned to component k.

A.4 Dependence of Blocks – Extra Plots

This section includes two plots related to the discussion of dependence of blocks in the paper. We compare the independent blocks version of our model with the variant that addresses the dependence via Gaussian random fields. The plots below show that considering dependence between the blocks can improve model predictions in some cases but it requires sampling from a high-dimensional distribution ($K \times B$), resulting in slow mixing.

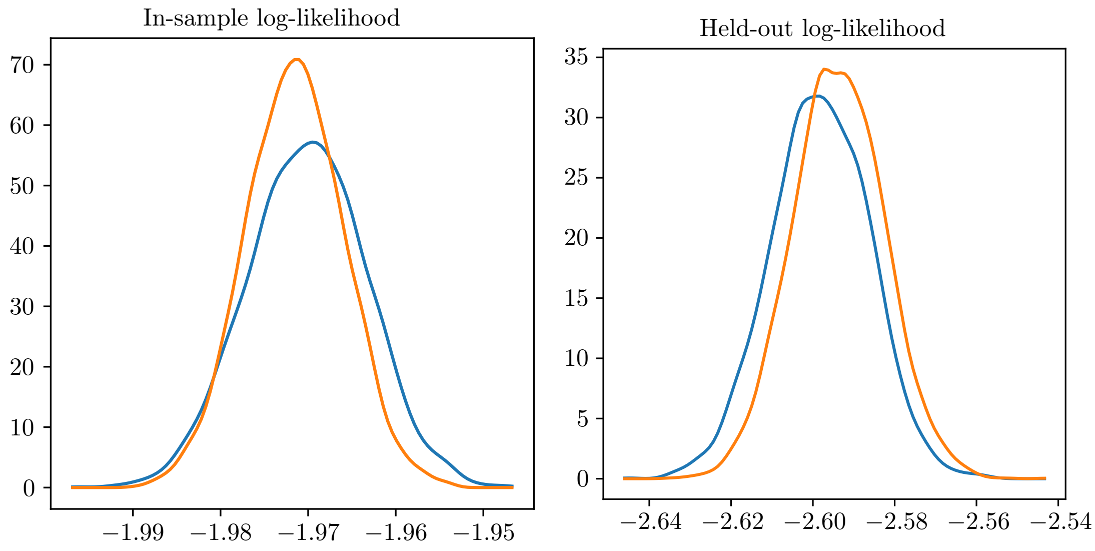


FIGURE A.3: Smoothed histograms of log likelihood computed on in-sample counts (left), and out-of-sample counts (right) using the proposed model with dependent blocks (—), and independent blocks (—) when $K = 3$. Blocking: MSOA, training data: 2015, test data: 2016, model specification 4.

Figure A.3 compares smoothed histograms for samples of in-sample log-likelihood $p(\mathbf{y}|\boldsymbol{\phi})$ for both variants of the model when $K = 3$, with their out-of-sample counterpart using samples from $p(\tilde{\mathbf{y}}|\boldsymbol{\phi})$. While independent-blocks model performs better in-sample, the dependent-blocks model generalises better to out-of-sample data. However, for $K = 2$ and $K = 4$, the model with independent blocks has lower RMSE on out-of-sample data as reported in the paper.

Figure A.4 shows the autocorrelation plot for the in-sample log-likelihood obtained from 50 000 samples that were thinned to 5000 for both variants to assess mixing performance. It is clear that successive samples obtained from the complex dependent-blocks model are more correlated to each other than for the case of independent blocks indicating slower mixing. Further, the inferences made using a Markov chain with high autocorrelation may lead to biased results.

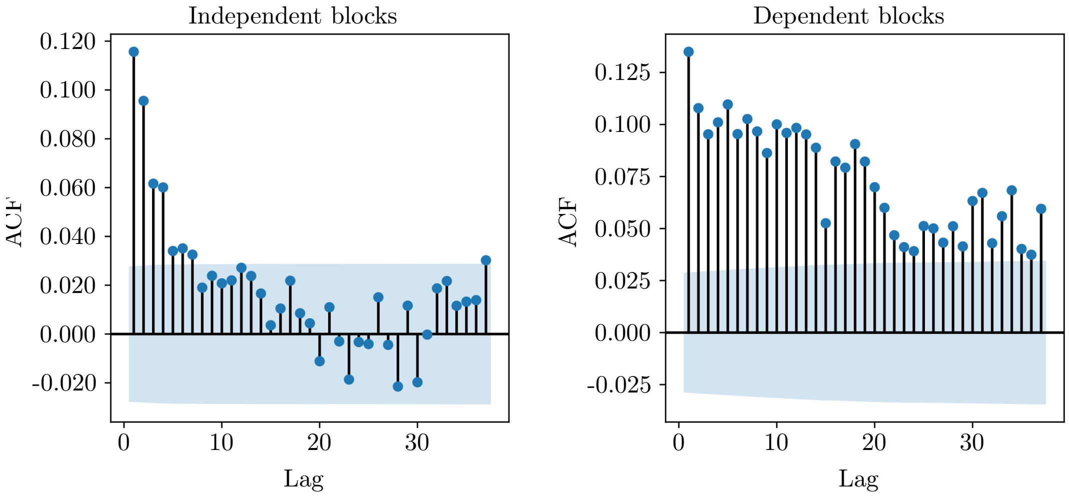


Figure A.4: Autocorrelation plots for the samples of in-sample log-likelihood when $K = 3$. Blocking: MSOA, training data: 2015, model specification 4.

Appendix B

Supplementary Material for Chapter 6

B.1 Short Length-scale Results

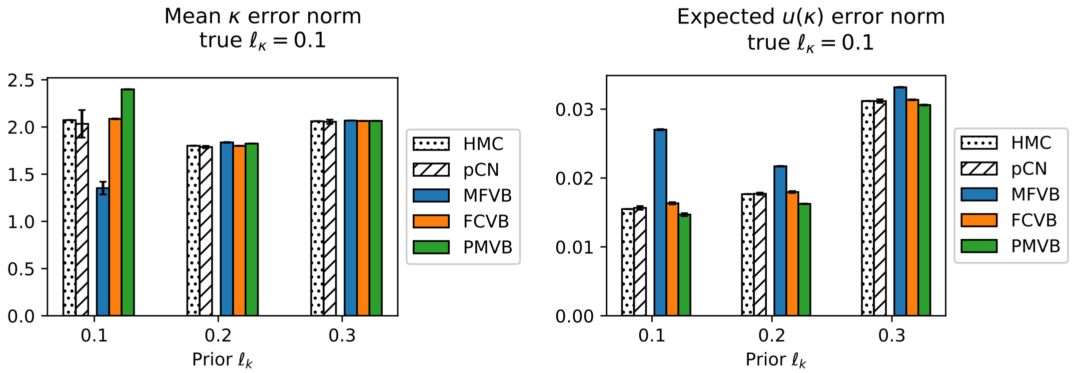


FIGURE B.1: Mean $\boldsymbol{\kappa}$ error norm for the Poisson 1D problem (left), as defined in (6.31), and expected solution error norm (right), as defined in (6.32). Both quantities are estimated using 10,000 samples from the inferred posterior distribution of $\boldsymbol{\kappa}$. Quantitatively, the sampling methods (HMC and pCN) and VB produce comparable results in both metrics, except MFVB parametrisation which captures the mean of $\boldsymbol{\kappa}$ very well, but fails to account for the uncertainty as manifested in high error norm in the solution space. For a qualitative comparison, see Fig. 6.4 where each row of results corresponds to a different value of the true prior length-scale ℓ_κ.

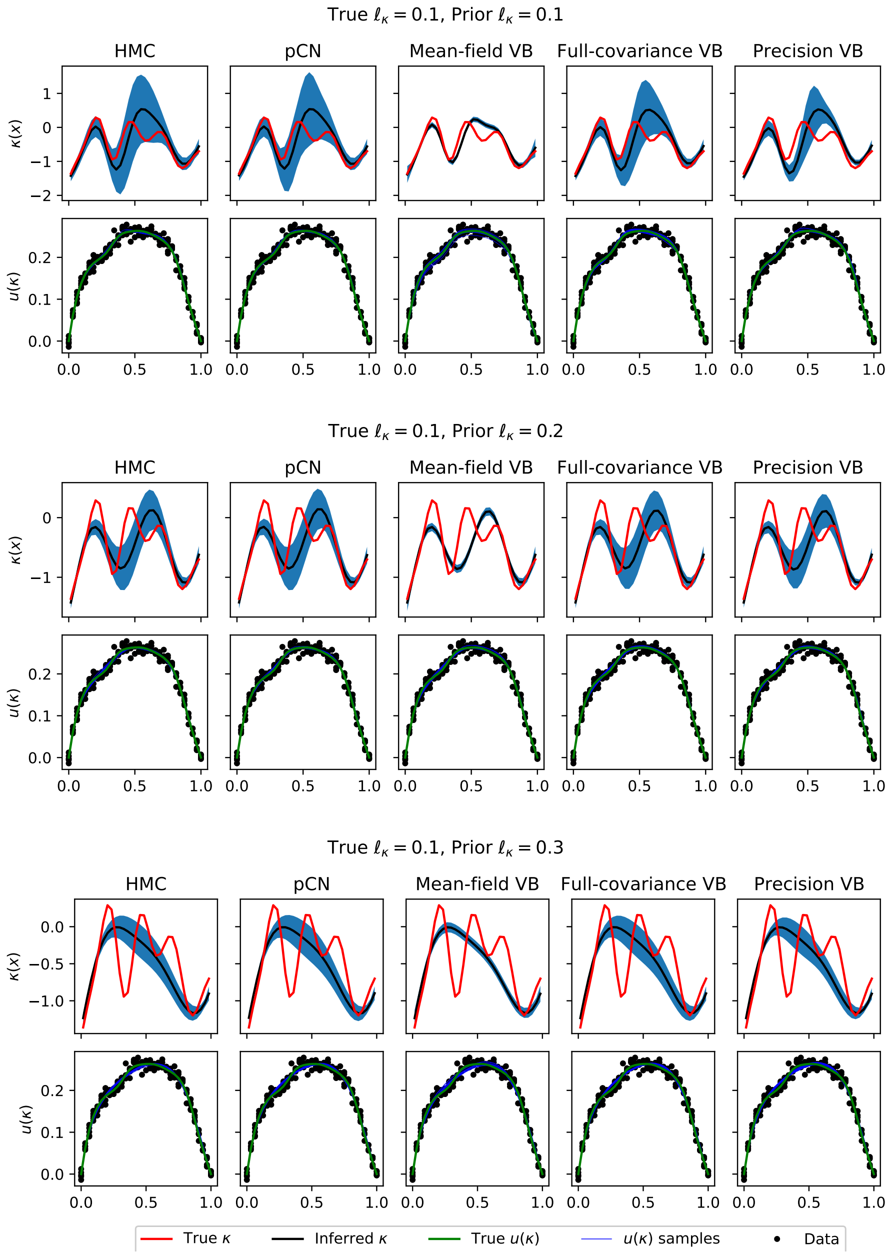


FIGURE B.2: Top row in each of the three panels show true values of $\kappa(x)$ (red), posterior means (black) and ±2 standard deviations (blue shaded regions) for HMC and VB variants for different values of prior length-scales ℓ_κ. The bottom rows show the data (black), true solution $\mathbf{u}$ (green), solutions for different samples of κ (blue). For the PMVB estimate, the bandwidth is set to 10.

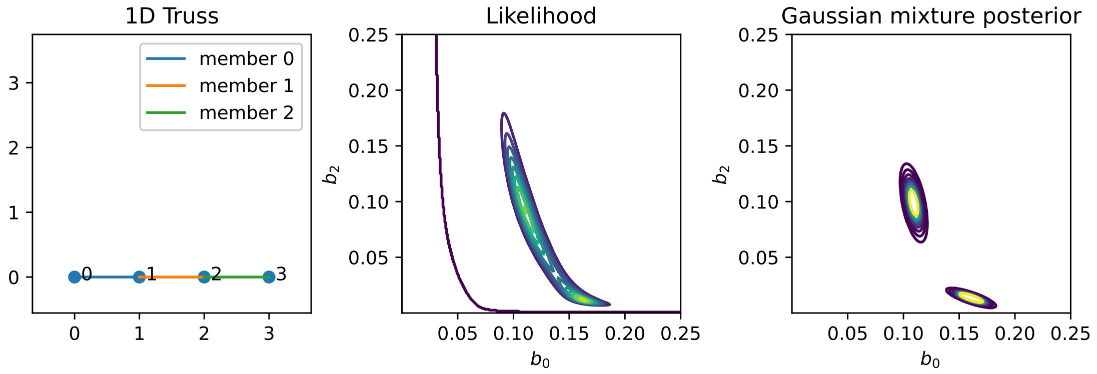


Figure B.3: One-dimensional truss discretised using three elements (shown in the left panel). The middle panel shows the likelihood surface for varying stiffness of the first and the third element. The right panel shows the posterior inference of the stiffness after displacements have been observed.

B.2 Bimodal Example

One of the advantages of VB over Laplace approximation is the flexibility of the approximating distribution. To illustrate this, we consider an example with a one-dimensional truss which is fixed at one node and contains three degrees of freedom that correspond to the horizontal motion of the three nodes as shown in the left panel of Fig. B.3. We assume that the stiffness b_i of each member i is constant within each member and, furthermore, the stiffness of the member 1 is the average of the stiffness of the members at the ends, *i.e.* $b_1 = (b_0 + b_2)/2$. The inverse problem is then defined as follows. Given the displacement vector $\mathbf{d}$ and boundary conditions, find the unknown stiffness parameters b_0 and b_1. To prevent negative or small stiffness, constraints are imposed on the stiffness of each member, $b_i > 0.1$, and Neumann boundary conditions are set to $\mathbf{f} = (0, 1, 0.05)^T$. Due to the constraint on the stiffness of member 1, the image of the forward problem is a manifold with dimension 2 embedded in displacement space $\mathbb{R}^3$. Due to the symmetry in this problem, the likelihood function, shown in centre panel of Fig. B.3, is bimodal.

We place a multivariate Gaussian as prior on the stiffness parameters, and use a bimodal trial distribution to infer the posterior distribution of the parameters b_0 and b_2 given observed displacement $\mathbf{d} = (0.1, 0.17, 0.23)^T$. Specifically, we consider a mixture of two multivariate Gaussians with equal fixed mixture weights. As there is no closed form expression for the KL divergence between a mixture of Gaussians and a single Gaussian, we estimate the KL divergence term in the ELBO using Monte Carlo sampling. As shown in the right panel of Fig. B.3, the resulting posterior distribution is bimodal and recovers the two modes present in the likelihood function. This illustrative example shows that when a proposed model exhibits multi-modality, the flexibility of variational Bayes methodology allows for specifying a family of trial distributions that can capture that property of the model.