University College London

Master Thesis

Understanding London Crime with Log-Gaussian Cox Processes

Author:
Jan Povala

Supervisor:
Louis Ellam
Prof. Mark Girolami

A thesis submitted in fulfillment of the requirements for the degree of Master by Research

August 16, 2018

UNIVERSITY COLLEGE LONDON

Abstract

Faculty of Engineering Sciences
Department of Computer Science

Master by Research

Understanding London Crime with Log-Gaussian Cox Processes

by Jan Povala

It is well understood that crime is clustered in both space and time. This is recognized by the emergence of so-called hot-spots, which are often short-lived. There are a number of theories in criminology that provide an explanation for this, although crime is a complex phenomenon and matching theories with empirical evidence is an ongoing task. Whilst standard regression techniques can be used, the assumptions are typically violated because of spatio-temporal autocorrelation, resulting in poor predictive performance. Recent work has advocated the use of Gaussian process models for their flexibility and tractability, although a Gaussian assumption is only an appropriate approximation in regions of high crime. Instead, a log-Gaussian Cox process (LGCP) is proposed as a model of criminal activity in space and time. We adopt a Bayesian approach to provide a full quantification of uncertainty. The resulting posterior distribution is intractable and poses significant computational challenges that we overcome by the use of Laplace approximation or Markov Chain Monte Carlo methods. In either case, by assuming a grid structure, we use Kronecker methods to accelerate the linear algebra procedures involved. We demonstrate the proposed methodology with London crime data for 2016, using socio-economic covariates derived from census data.

Acknowledgements

I would like to thank my supervisors Louis Ellam and Prof Mark Girolami for all their advice and help throughout the course of this project. I would also like to thank Dr Seppo Virtanen for his helpful suggestions and comments.

Contents

1 Introduction

The endeavour of understanding crime is not new. Firstly, the institutions that administer justice have already existed in ancient civilisations. Secondly, the field of criminology as an academic discipline has long been established. Criminologists recognised early that a systematic approach to understanding crime is necessary for designing effective responses to crime, and for implementing policy. The modern criminology community started systematically developing their hypotheses as early as the 1700s (Jewkes and Letherby, 2002). The understanding of criminal activity and its sources has changed over time. The answers to very fundamental questions such as 'Why some people commit crime and others do not?' have been changing as a result of secular trends and structural shifts in society such as industrial revolution (Jewkes and Letherby, 2002). The discipline of criminology involves both laying out the theoretical underpinnings as well as providing supporting evidence in order to present a theory that explains some aspect of crime. Empirical evidence involves asking the right questions and designing appropriate experiments that can answer them. In recent decades, improvement in the availability of data - improved census collection, digitalisation - have enabled researchers to ask questions which had not been possible before. By representing theories as mathematical models, the collected data can be compared with the predictions obtained from the models and thus provide evidence for, or against, the theory. To build models that are in line with the theory, it is necessary to review the main ideas in the field of criminology.

1.1 Overview of modern criminology

It is common to split criminological theories into three groups. Firstly, those concerned with the individual, secondly, the theories which pay particular attention to factors external to the offending individual, and lastly those that pay close attention to the role of the state with regards to criminology (Jewkes and Letherby, 2002).

1.1.1 Theories focused on the individual

In the first category which focuses on the individual offenders, **classical criminology** which was pioneered by an Italian philosopher Cesare Beccaria and a British philosopher Jeremy Bentham in the second half of the 18th century had a huge impact on legal and penal systems across Europe and North America (Carpenter, 2010). At the centre of this are intellectual principles and political ideas of the Enlightenment - utilitarianism and power of rationality. The theory is built on the presumption that criminals go through a rational calculative decision making process before committing a crime. As the fundamental cause of criminal behaviour of the criminals are the competing concepts of *free will* and *hedonism*, *maximisation of pleasure*, and *minimisation of pain*, that all people (both offender and non-offenders) need to manage (Jewkes and Letherby, 2002, ch.2). The theory strongly advocates punishing the crime rather than punishing individual's social or physical characteristics, presumption of innocence, abolition of capital punishment, and condemnation of the use of torture. Despite its

popularity at the time, this theory was not able to account for children criminality, rising crime rates in the aftermath of industrial revolution. This limitation led to alternative theories.

In the late 18th century, Italian physician Cesare Lombroso put forward what has become known as **biological theories**. Contrary to the classical theory, the theory suggests, criminals do not act out of free will but out of the innate urge to commit crime (Beccalossi, 2010). Lombroso has been largely influenced by Darwinism which he used to provide supporting evidence for this theory. The key concepts include *degeneration* - an arrest in the development of an individual, and *atavism* - reappearance of ancestral characteristics which have been lost through evolutionary change. Lombroso's theories and methodology were criticised for their anecdotal nature, inconsistent methodology, and conceptual flaws, and therefore criminologists of the early 20th century focused on sociological theories instead (Beccalossi, 2010).

However, interest in the theories based on biology has re-emerged in the middle of the 20th century, with focus on examining biological and psychological factors together with the importance of social context. These methods, known as **biosocial theories**, step away from unethical eugenic measures advocated by the early biological theories and focus on improving the environment (Rocque, Welsh, and Raine, 2012). In the context of criminology, advances in research in genetics and neuroscience have helped pinpointing correlations of particular genetic polymorphisms and brain functioning with increased criminal behaviour. An example of such correlation is cognitive deficit among offenders. Discovering the causes of the brain deficit is still an active area of research but the current hypotheses with substantial evidence include: substance abuse during in foetal stages, traumatic experiences in early childhood, child abuse. Other examples of brain functioning with a strong link to criminal activity are impulsivity, negative emotionality, aggression. The main point lies in recognising that factors such as genes or brain functioning interact with the environment. Biosocial criminologists focus on identifying the former, but intervening in the latter. The intervention measures suggested by biosocial criminologists focus on developmental prevention. These include family-centred programs, preschool and school-based programs among others. For a more extensive treatment, please refer to Rocque, Welsh, and Raine (2012).

1.1.2 Theories focused on the external social factors

The second category of theories focuses on the factors external to the individual - the neighbourhood, the peer group, and the family. Inspired by the positivism philosophy of scientifically identifying the facts, these theories aim to explain the causes of criminal behaviour in order to subsequently design appropriate responses.

One of the most prominent schools of thought in this category is the **social disorganisation** theory proposed in the 1920-30s by the 'Chicago school' sociologists. Social disorganisation theory suggests that offenders are more likely to offend in areas with poor housing, poor health, socio-economic disadvantage, and high turnover of the population. This theory proposes that crime is function of neighbourhood dynamics, rather than the individuals (Jewkes and Letherby, 2002). In order words, social disorganisation considers how the social fabric of a community affects offender's decisions. The theory is still applicable in contemporary criminology research despite its age. For example, Johnson and Summers (2015) show that offenders tend to target areas with low social cohesion in the case of theft from vehicle crimes. In policy terms, the focus is on understanding how environment contributes to crime, and on reorganising socially disorganised communities (Jewkes and Letherby, 2002). In the same work, Johnson and Summers (2015) also provide evidence that offenders tend

to target areas that are within their awareness spaces - close to their home, place of work, or on places encountered during regular activities - as well as places which are easily accessible. The criminology theory which focuses on explaining individual's place of offending by considering how their routines activities influence their awareness of criminal opportunities is referred to as **crime pattern theory** (CPT).

Strain theory also focuses on factors external to the individual. It was proposed by Robert Merton in the 1938 (Agnew, 2008). Strain refers to events or conditions that an individual dislikes, for example inability to achieve goals (mainly monetary success), loss of positive stimuli, presentation of negative stimuli, and others. The strains are said to cause emotions such as anger, fear, depression, which majority of the people find legal ways of coping with. However, if an individual sees that they are unable to cope with the strain, the affected individual often seeks a form of escape, revenge, or alleviation from the strain, and thus the likelihood of committing crime increases. Factors that increase the risk of resorting to criminal behaviour are lack of social control, poor coping skills, and insufficient resources (Agnew, 2008).

Building upon strain theory, **subcultures theory** highlights status frustration and differential opportunity as the causes of crime (Blackman, 2014). Individuals unable to achieve cultural goals or expectations often reject them, and replace them with their own where they find justification for their criminal action. They often form groups, gangs, where acts such as crime are often treated with prestige and respect from the peers in the group. Mainstream moral authorities such as parents, teachers are replaced by the fellow members of the group. For an extensive discussion on subcultural theories please refer to Blackman (2014).

Another theory focusing on factors of the environment, but still focusing on the individual is **social control theory** which was introduced in Hirschi (1974). Rather than asking why criminals engage in criminal activity, Hirschi asks why non-criminals do not do so. The theory builds on the premise that all of us, beginning at birth, have the hedonistic drive to act in selfish and aggressive ways that lead to criminal behaviour. Hirschi posits that it is only through social bonds that we refrain from such behaviours. He categorises the bonds in four categories (Pratt, Gau, and Franklin, 2018):

1. *Attachment* - it refers to the level of psychological affection that one has for prosocial others and institutions. In practice, it means that thanks to the attachment to parents, family, school, etc., one tries not to disappoint those that he/she cares about.

2. *Commitment* - not wanting to look bad in front of those we are attached to prevents us from criminal and aggressive behaviours. For example, the risk of the break-up of marriage or of losing a job helps in refraining from any harmful activities.

3. *Involvement* - if people spend their time in a prosocial activity they simply has less time to pursue criminal activities. By being enganged in social, academic, or athletic activities, the individuals will have less time for engaging in antisocial activities. This is mainly related to youth population.

4. *Belief* - this refers to the degree to which one adheres to values associated with lawful behaviour. Hirschi suggests that there is a strong link between attitudes and behaviour.

According to Hirschi, these four types of social bonds come together to control our behaviour indirectly - not all rules have to be written in the law - i.e. many rules

are not written in the law, yet they are part of the social norms that we all observe (Pratt, Gau, and Franklin, 2018). Although this theory is still controversial among criminologists, it is still considered to be relevant. When the theory was presented it provided strong empirical evidence for its claims, and thus raised the bar of experiment quality in the criminology community (Pratt, Gau, and Franklin, 2018).

Lastly, **labelling theory** is a school of thought within criminology that deals with not only the individual and the environment around them, but also involves thinking about criminal justice system. We summarise the treatment on labelling theorists from Jewkes and Letherby (2002, ch. 2). Their main concerns can be summarised in two questions:

1. How is it that some behaviour is *labelled* deviant?
2. What are the consequences of that labelling process?

According to the proponents of this theory, reaction to the crime is more important than the crime itself. The research of this group of theorists often focuses on bias within agencies responsible for administration of justice, social impact of being labelled as deviant and others. In terms of policy, they advocate decriminalisation and diversionary policies. Critics attack this theory for being based on a strong assumption of the society being democratic and consensual. It seems that those who are in power get to shape the normative view what is deviant and what is not.

1.1.3 Theories considering importance of the state

Rather than labelling theory's view of society as consensual, this group of theories views society as *rooted in conflicting interests*.

Firstly, there is a branch of literature that stems from the Marxist presumption: "powerful in society use the various resources available to them (including the law) to secure and maintain their dominant position" (Jewkes and Letherby, 2002). Capitalism creates a desire to consume, but not all members of the society are able to participate to the same extent. As a result of capitalist process, there is more situations where those who have and those who do not are put in conflict with one another. Subsequently, criminal behaviour is a result of *class conflict* and if a an individual chooses to commit crime, it is viewed as a *political expression*. Within this wide class of theories, **radical criminology** connects the Marxist ideas with labelling theory, in which the the ruling class (those who have) are able to put labels on the lower class thanks to their control of legal and penal system (Jewkes and Letherby, 2002). **Critical criminology**, partially borrowing ideas from Marxist presumption explores the ways in which *class*, *gender*, and *race* play a role in the criminal justice system. Their hypothesis is that the state and its agencies serve to marginalise and consequently criminalise some groups and not others. Often relying on history, proponents of this theory question the objectivity of knowledge where they treat it as an ideology that serves state and its practices. These theories were however not put into practice (Jewkes and Letherby, 2002).

Lastly, one of the richest criminology theories is **realism**. We give a brief summary of on this topic based on the treatment in Young (1992). Realists argue that criminology should be faithful to the nature of crime. They define the crime square: *offender*, *victim*, *public*, and *police/state agencies*. Whilst traditional criminology theories focus mainly on the causes of offending, realism urges us to look at the other three components as well: victim, police, and the general public. They see *relative deprivation* as a major cause of criminal activity (similar to subculture theory). Stemming from

involving all four forces when understanding crime is the principle of *focusing on lived realities*. In practical terms it means avoiding generalisations by understanding social context, choosing local approaches, breaking analyses into fine social axes such as age, gender,class and race.

1.2 Statistical approaches to criminology

The two main approaches to crime modelling in the literature are *crime forecasting models* and *crime understanding models*. The former approach is concerned with accurately forecasting the crime rates. By interpolating the data, they are able to build models based purely on space and time without including any external factors such as socio-economic indicators. For example, Taddy (2010) treats criminal activity as an inhomogeneous Poisson process with intensity specified by an auto-regressive model, Flaxman et al. (2015) uses Gaussian Processes with a rich covariance function to specify prior over the intensity of a Poisson process, Mohler (2013) treat crime counts as realisations of self-exciting Hawkes process. These approaches tend to be focused on the time-varying macro trends in criminal activity. While these approaches are useful for law-enforcement agencies in deploying their limited resources, they do not provide insights into the criminal activity, thus are of limited use for tasks such as designing policies.

For the methods whose main concern is to reconcile the criminology-based theories with data, the literature is divided into approaches that use socio-economic data to help explain *aggregate* crime rates and the approaches which answer very *specific* questions within the local context of crime as advocated by the realists. In the first group, Marchant et al. (2018) apply a semi-parametric regression model to crime rates in Sydney, Australia, Osgood (2000) use Poisson regression to infer the drivers of crime, and Deadman (2003) specify an econometric linear model, estimated by Ordinary Least Squares method, to forecast residential burglaries in the United Kingdom. In the second group, for example Johnson and Summers (2015) use data on thefts from vehicles to corroborate crime pattern theory and social disorganisation theory, Kurland, Johnson, and Tilley (2014) test whether certain establishments such as soccer stadium attract criminal activity.

1.3 Aims, Objectives

Our choice of methodology is to a large extent driven by the study area and the data availability. Our study area is London, UK for which we are able to publicly access location, type, and the month of each crime. Additionally we have access to the UK census data which provides socio-economic indicators about the location in question (Office For National Statistics, National Records Of Scotland, and Northern Ireland Statistics And Research Agency, 2016). However, the data do not provide any information about the offender, and therefore we will not be able to give evidence for, or against, any theories which rely on understanding the offender. For this reason, our method will be focused on finding links between *aggregate* rates of criminal activity and external socio-economic factors.

The following list of objectives summarises what this work will attempt to achieve.

1. Specify models based on external socio-economic factors to create a representation of the true process driving crime.

2. The uncertainty in the models must be fully quantified.

3. Employ appropriate tools to check performance of different models.
4. Prefer models that are parsimonious and interpretable.
5. Discuss strengths and weaknesses of the chosen modelling approaches.
6. The modelling approach should be scalable to large problem sizes.

After the introduction to criminology theory, a quick overview of statistical approaches and a set of objectives in this chapter, chapter 2 gives the mathematical machinery used to build models. Chapter 3 gives the details of statistical inference, results of which we report in chapter 4, and we give conclusions, and directions for future work in chapter 5.

2 Mathematical Background

Having identified environmental factors as a significant driver of crime in chapter 1, we will treat criminal activity as an environmentally-driven stochastic process that varies in space and time. Cox process has been a natural choice for environmentally driven point processes in spatial statistics (Diggle et al., 2013). As described in the original paper by Cox (1955), it is a suitable model for "events occurring haphazardly in space or time". In this chapter, we first give a formal definition of the Cox process, then give details of log-Gaussian Cox process (LGCP), which is a variant of Cox process that allows full uncertainty quantification using the Bayesian framework. The model defines a latent function with a Gaussian Process prior to drive the intensity of the Cox process. The properties of the Gaussian prior are specified using a set of hyper-parameters. Subsequently, we describe two different approaches to statistical inference for LGCP. The first approach proceeds by integrating out the latent function using Laplace approximation to obtain an approximate marginal likelihood which allows obtaining a point estimate of the optimal hyper-parameters. The optimal hyper-parameters are then used to approximately infer the posterior distribution of the latent function. In the second approach, which is fully-Bayesian, we treat the hyper-parameters as random variables. Using a Markov Chain Monte Carlo sampling scheme posterior distribution of all the quantities of interest can be inferred with convergence guarantees. While the first approach is less demanding computationally, it does not fully capture the uncertainty in the hyper-parameters and does not provide convergence guarantees for the posterior distribution of the latent function. We conclude the chapter with a section focussed on computational issues.

2.1 Log-Gaussian Cox Process

Definition 2.1 (Cox Process)**.** Given domain $\boldsymbol{x}$, Cox process $\mathrm{Y}(\boldsymbol{x})$ is defined by two postulates:

1. $\Lambda(\boldsymbol{x})$ is a nonnegative-valued stochastic process;
2. conditional on the realisation $\lambda(\boldsymbol{x})$ of the process $\Lambda(\boldsymbol{x})$, the point process $\mathrm{Y}(\boldsymbol{x})$ is an inhomogeneous Poisson process with intensity $\lambda(\boldsymbol{x})$.

The suitability of Cox process is justified by the properties of Poisson distribution that is at the core of the process. The Poisson distribution describes the probability of a number of events occurring in a fixed interval given the rate at which these events occur (Grimmett and Stirzaker, 2001). The events occur in an unstructured manner, often referred to as the *lack of memory property*.

A popular choice of the intensity process, $\Lambda(\boldsymbol{x})$, that satisfies definition 2.1 and with extensively studied mathematical properties is

$$\Lambda(\boldsymbol{x}) = \exp(f(\boldsymbol{x})), \tag{2.1}$$

where $f(\cdot)$ is the latent function with a Gaussian process prior:

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k_\theta(\cdot,\cdot)), \tag{2.2}$$

where $m(\boldsymbol{x})$ is the mean function, and $k_\theta(\cdot,\cdot)$ is the covariance function parameterised by $\boldsymbol{\theta}$ (Diggle et al., 2013). This version of Cox process was formally introduced by Moller, Syversveen, and Waagepetersen (1998) and is referred to as log-Gaussian Cox process. In the next section, we give overview of Gaussian processes.

2.1.1 Gaussian Processes

This treatment closely follows Rasmussen and Williams (2006), in both notation and content.

Definition 2.2 (Gaussian Process)**.** Gaussian Process (GP) is a stochastic process whose finite-dimensional distributions are jointly Gaussian.

Gaussian Processes can be used to express distribution over functions. In the Bayesian setting, it allows us to express our prior beliefs about a class of functions before we observe any data.

Let f be the true function that we are interested in. Before observing any data, we can express our prior beliefs using a GP. A GP is defined by the mean function $\mu(\boldsymbol{x})$ and the covariance function $k_{\boldsymbol{\theta}}(\cdot,\cdot)$. The mean function specifies the expected value at any given location of the domain, and the covariance function gives a measure of similarity/distance between function values at any two locations. The covariance function is parameterised by hyperparameters $\boldsymbol{\theta}$. There are conditions that need to be met in order for a function to be a valid covariance function. We discuss details of covariance functions, including common examples, in section 2.1.2. For the purposes of illustration in this section we assume a zero-mean GP prior with squared exponential covariance function with lengthscale hyperparameter $\ell = 1$. For example, four draws from the GP prior are shown in Figure 2.1a. Next, we illustrate how observing data changes the prior beliefs about the function f.

After observing function values $\mathbf{f}$ at locations X, our belief about the true function f changes. By denoting any other locations than the observed ones as X_*, we can express the posterior distribution of the function values $\mathbf{f}_*$ at locations X_*, given that we observed $\mathbf{f}$ at X. Given that the prior is a GP, the function values at observed location and any other locations are jointly Gaussian (see equation A.2):

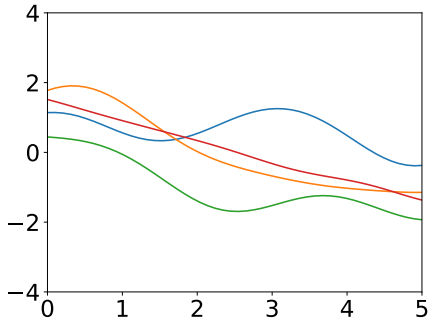
$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \boldsymbol{K}(X,X) & \boldsymbol{K}(X,X_*) \\ \boldsymbol{K}(X_*,X) & \boldsymbol{K}(X_*,X_*) \end{bmatrix}\right),$$

where $\boldsymbol{K}(X,X')$ is the covariance matrix with $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}'_j)$. Using the properties of normal distribution (see equation A.3) it follows that $\mathbf{f}_*|X_*, X, \mathbf{f} \sim \mathcal{N}(\mu_*, \boldsymbol{K}_{\mathbf{f}_*})$, where

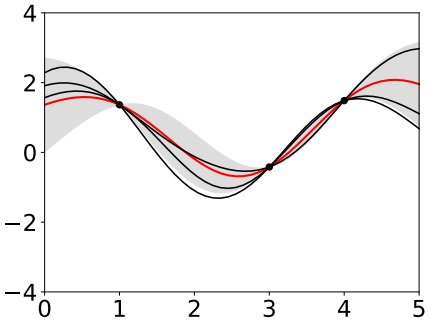
$$\begin{aligned} \mu_* &= \boldsymbol{K}(X_*,X)\boldsymbol{K}(X,X)^{-1}\mathbf{f} \\ \boldsymbol{K}_{\mathbf{f}_*} &= \boldsymbol{K}(X_*,X_*) - \boldsymbol{K}(X_*,X)\boldsymbol{K}(X,X)^{-1}\boldsymbol{K}(X,X_*). \end{aligned}$$

An example of posterior distribution over the functions with the GP prior after observing three function values is shown in Figure 2.1b

In a realistic scenario, the function values are not observable and we only observe noisy versions of it, possibly with a transformation. For example, in the log-Gaussian Cox process above, we can only observe a realisation of the random Poisson process



(A) Four draws from the zero-mean GP prior with squared exponential kernel with lengthscale $\ell = 1.2$.



(B) Posterior distribution over functions after three function values were observed. Red line represents the mean, and the shaded area around it is the 95% credible interval.

whose intensity is driven by $\exp(f)$. We denote the vector of observations at a set of locations X as $\mathbf{y}$. The connection between the latent (unobserved) function f and the observations $\mathbf{y}$ is often referred to as *observation model* and is denoted as $\mathbf{y}|\mathbf{f}$. A common observation model is assuming that the each observation is a result of perturbing the function value with independent identically distributed Gaussian noise. This assumption results in a posterior distribution which is tractable. We do not give more detail about this case due to our focus on log-Gaussian Cox process model, but more details can be found in Rasmussen and Williams (2006, p.16). The details of statistical inference for LGCP processes is dealt with in 2.2.

2.1.2 Covariance Functions

As already mentioned in 2.1.1, the covariance function is the part of the GP that encodes nearness or similarity of two function values as a function of their locations. In general, locations that are close to each other tend to have similar function value and thus higher degree of covariance than the locations further apart. However, there are use cases when more complicated stucture needs to be incorporated within a covariance function. For example, there is a class of covariance functions which encode periodicity.

Covariance functions are a special case of kernels. Kernel is a function k of two arguments that maps the arguments $\boldsymbol{x} \in \mathcal{X}, \boldsymbol{x}' \in \mathcal{X}$ to $\mathbf{R}$. In order for a kernel to be valid covariance function it must be positive semidefinite. A kernel is positive semidefinite if

$$\int k(\boldsymbol{x}, \boldsymbol{x}') f(\boldsymbol{x}) f(\boldsymbol{x}') d\mu(\boldsymbol{x}) d\mu(\boldsymbol{x}') \geq 0, \tag{2.3}$$

for all $f \in L_2(\mathcal{X}, \mu)$ (Rasmussen and Williams, 2006). As a consequence, the gram matrix for the covariance function k and a given set of inputs must be positive definite. *Gram matrix* is a matrix with elements $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}'_j)$.

Choice of the covariance function for a particular problem is often determined by the properties it satisfies. The summary of the most relevant properties is based on Rasmussen and Williams (2006, sec. 4.1). Firstly, *smoothness* of the covariance function determines whether the resulting GP is continuous and differentiable. Depending on the mean function of the GP, if the covariance function is continuous and differentiable, the properties will carry over to the GP. Smoothness properties determine

whether the inference is analytically tractable, e.g. using derivatives to guide the search for optimal parameters. Another important property of covariance functions is *stationarity*. It refers to invariance to translations in the input space. Consequently, the resulting process has the same moments at any location, and the covariance function is only a function of of $\boldsymbol{x} - \boldsymbol{x}'$. If in addition, it holds that the covariance function is a function of only $|\boldsymbol{x} - \boldsymbol{x}'|$, the covariance function is said to be *isotropic*.

There are cases in which it is desirable to combine existing covariance functions. One such scenario is having separate kernel per dimension and then combining them into a single kernel. For example, in spatio-temporal modelling the kernels for the spatial dimension and the temporal dimension are specified individually and then combined. Combining operations must however ensure that the resulting function is a covariance function, i.e. satisfy equation 2.3. We state without a proof that the sum of two covariance functions is a valid covariance function. Similarly, the product of two covariance functions is a covariance function. For other operations that produce valid kernel, and for proofs of sum and product, refer to Rasmussen and Williams (2006). Next, we give examples of common covariance functions.

Squared Exponential

Squared exponential kernel, defined as

$$k_{\text{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(\frac{|\boldsymbol{x} - \boldsymbol{x}'|^2}{2\ell^2}\right), \tag{2.4}$$

where ℓ is the characteristic lengthscale, is one of the most popular covariance functions. The lengthscale parameter can be loosely interpreted as how fast the target function oscillates - the larger the lengthscale the slower the oscillation. A more rigorous interpretation of the lengthscale is calculating the expected number of function value upcrossings of a particular level as a function of ℓ. For details, please refer to Rasmussen and Williams (2006, sec 4.1-4.2) and Adler and Taylor (2007, sec 11.1).

Square exponential kernel is infinitely differentiable, hence the GPs with this covariance function are smooth. This kernel is also stationary and isotropic. Stein (1999) argues that it is more prudent to specify the smoothness of the process using data rather than apriori. The next class of kernels which we introduce allows for controlling both the characteristic lengthscale and the smoothness.

Matérn

Matérn family of covariance functions is defined as

$$k_{\text{Matern}}(\boldsymbol{x}, \boldsymbol{x}') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}|\boldsymbol{x} - \boldsymbol{x}'|}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}|\boldsymbol{x} - \boldsymbol{x}'|}{\ell}\right), \tag{2.5}$$

where ℓ is the characteristic lengthscale, ν is the smoothness parameter, and K_ν is a modified Bessel function (Rasmussen and Williams, 2006). It can be shown that that the GPs with Matérn covariance functions are k-times mean-square differentiable if and only if $\nu > k$. This is important especially if inference methods based on likelihood optimisation are chosen. Additionally, Abramowitz and Stegun (2013) show that if ν is a half-integer, i.e. $\nu = p + \frac{1}{2}$, the covariance function becomes especially simple,

giving

$$k_{\nu=p+1/2}(\boldsymbol{x},\boldsymbol{x}') = \exp\left(-\frac{\sqrt{2\nu}|\boldsymbol{x}-\boldsymbol{x}'|}{\ell}\right)\frac{\Gamma(p+1)}{\Gamma(2p+1)}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\left(\frac{\sqrt{8\nu}|\boldsymbol{x}-\boldsymbol{x}'|}{\ell}\right)^{p-i}. \tag{2.6}$$

For this reason, the most popular choices in machine learning have become $\nu = 3/2$ and $\nu = 5/2$. Left panel of Figure 2.2 shows covariance functions from the Matérn family with fixed lenghtscale $\ell = 1$ and varying smoothness ν. The right panel shows a sample from the corresponding GPs. One can distinctly see the smoothness properties of the GPs which correspond to different values of ν.

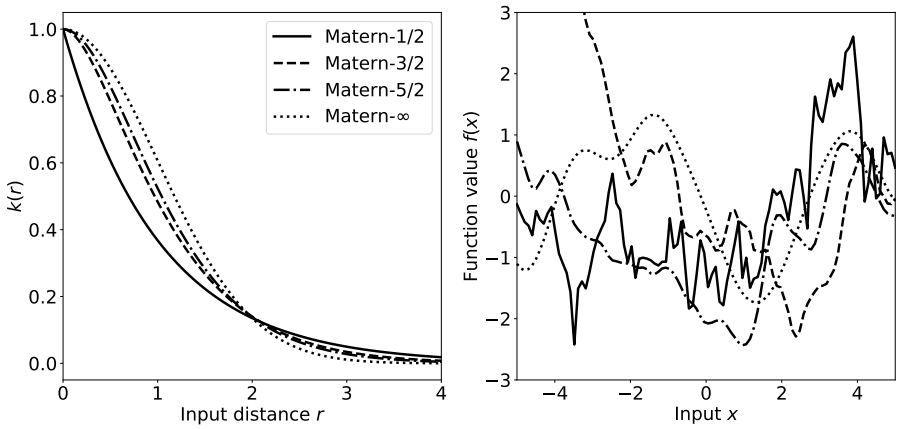


FIGURE 2.2

Other kernels

Given that the necessary and sufficient requirement for a valid kernel (equation 2.3) gives a lot of freedom, there are many possibilities to encode various properties such as periodicity. One such example is exp-sine squared kernel defined as

$$k\left(\boldsymbol{x},\boldsymbol{x}'\right) = \exp\left(-\frac{2\sin^2\left(\frac{\boldsymbol{x}-\boldsymbol{x}'}{2}\right)}{\ell^2}\right), \tag{2.7}$$

where ℓ represents the periodicity. This kernel is formed by first projecting the one-dimensional input x into two-dimensional $\boldsymbol{u}(x) = (\cos(x), \sin(x))$ and then applying the square exponential kernel. This process is referred to as warping (Rasmussen and Williams, 2006). Figure 2.3 illustrates the exp-sine squared covariance function with period one (left), and two draws from such a process (right). For other examples of kernels, please refer to Rasmussen and Williams (2006).

2.2 Inference

The goal of inference is to make conclusions about the true process or mechanism giving rise to the data (Young and Smith, 2005). For our purposes, where we assume an instance of log-Gaussian Cox process as the true random process (see sec 2.1), we would like to understand the properties of the latent function f as well as its

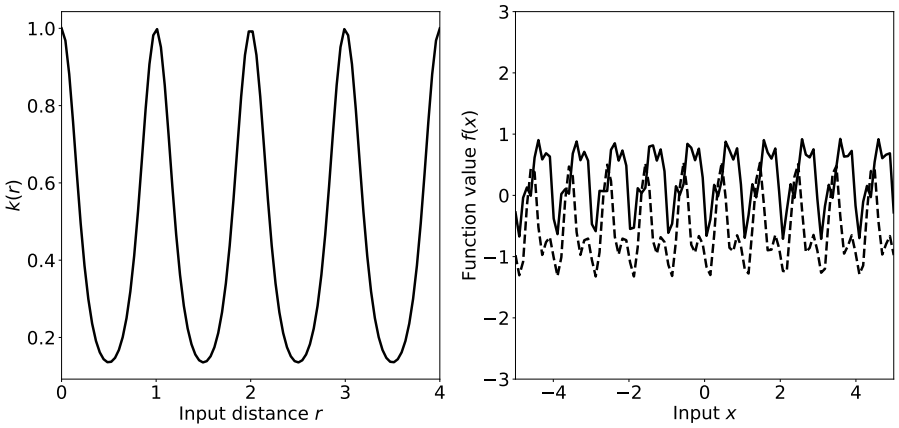


FIGURE 2.3

parameters $\boldsymbol{\theta}$. The two main approaches to statistical inference are the frequentist approach in which the parameter values of the data generating process are treated as fixed values, and the Bayesian approach that treats the parameters as random variables. We consider it important to quantify the uncertainty in all the quantities, including parameters, and therefore we choose to work in the Bayesian framework. It allows inferring full posterior distribution of all the quantities of interest.

In the Bayesian framework, inference of the distribution of the unobserved quantities, which in our case are the latent function values $\mathbf{f}$ and the hyper-parameters $\boldsymbol{\theta}$, relies on Bayes's theorem. Bayes's theorem states that for a pair of *random* variables X and Y,

$$p(X = x|Y = y) = \frac{p(Y = y|X = x)p(X = x)}{p(Y = y)}. \tag{2.8}$$

Now, we apply Bayes's theorem in the context of log-Gaussian Cox process. Given the vector of observed values $\mathbf{y}$ of $Y(\boldsymbol{x}_i)$ at a set of locations $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots\}$ the joint model of the data $\mathbf{y}$, corresponding latent function values $\mathbf{f}$, and the hyper-parameters $\boldsymbol{\theta}$ has density

$$p(\mathbf{y}, \mathbf{f}, \boldsymbol{\theta}) = \underbrace{p(\mathbf{y}|\mathbf{f})}_{\text{Poisson}} \underbrace{p(\mathbf{f}|\boldsymbol{\theta})}_{\mathcal{GP}} \underbrace{p(\boldsymbol{\theta})}_{\boldsymbol{\theta}-\text{prior}}.$$

Our aim is to use data, $\mathbf{y}$, to inform us about the posterior distribution of the latent field values, as well as, the hyper-parameters. We apply the Bayes's theorem to get

$$p(\mathbf{f}, \boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}, \tag{2.9}$$

where

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}d\mathbf{f}, \tag{2.10}$$

which is intractable.

There are different options to overcome the intractability. Two very popular approaches that we will focus on are *Laplace Approximation*(LA) which we describe in section 2.2.1, and *Markov chain Monte Carlo sampling*(MCMC) for which the background is given in 2.2.2.

2.2.1 Laplace Approximation

For an approximate Bayesian inference, where the hyper-parameters $\boldsymbol{\theta}$ are treated as fixed values, we want to obtain the posterior distribution of $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$. This, however, requires knowing the true value of $\boldsymbol{\theta}$. A common approach is to approximately integrate out the latent variables $\mathbf{f}$ to obtain $p(\mathbf{y}|\boldsymbol{\theta})$, and maximise this quantity. Having obtained the estimate of $\boldsymbol{\theta}$, $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$ is then inferred using an approximation method. One of the few possible approaches is Laplace approximation.

This exposition of Laplace approximation closely follows Bishop (2006, p. 213-216). The main goal is to find an approximation to the intractable probability density $p(\boldsymbol{z})$ of which we have access to only an unnormalised version $f(\boldsymbol{z})$, such that

$$p(\boldsymbol{z}) = \frac{1}{Z} f(\boldsymbol{z}).$$

Laplace approximation method finds a Gaussian density $q(\boldsymbol{z})$ centered around the mode of $p(\boldsymbol{z})$, a point $\boldsymbol{z}_0$ such that

$$\nabla_{\boldsymbol{z}} f(\boldsymbol{z})|_{\boldsymbol{z}=\boldsymbol{z}_0} = 0.$$

The mode of $p(\boldsymbol{z})$ is the same as that of $f(\boldsymbol{z})$. The mode is usually found by an optimisation procedure which exploits the gradients such as Newton method.

Using the property of multivariate normal distribution, log of the density function is a quadratic function of the vector of variables. For this reason, we perform Taylor expansion of $\log f(\boldsymbol{z})$ around $\boldsymbol{z}_0$ up to and including second order which gives

$$\log f(\boldsymbol{z}) \simeq \log f(\boldsymbol{z}_0) + (\boldsymbol{z} - \boldsymbol{z}_0)^\top \underbrace{\nabla_{\boldsymbol{z}} \log f(\boldsymbol{z})|_{\boldsymbol{z}=\boldsymbol{z}_0}}_{=0} - \frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)^\top \boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0),$$

where $\boldsymbol{A} = -\nabla\nabla_{\boldsymbol{z}} \log f(\boldsymbol{z})|_{\boldsymbol{z}_0}$. After exponentiating the Taylor expansion of $\log f(\boldsymbol{z})$ we obtain

$$f(\boldsymbol{z}) \simeq f(\boldsymbol{z}_0) \exp\left(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)^\top \boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0)\right),$$

which is in the form of an unormalised Gaussian density. Using the property of Gaussian distribution, we obtain normalised density

$$q(\boldsymbol{z}) = \det\left(\frac{1}{2\pi}\boldsymbol{A}\right)^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{z}_0)^\top \boldsymbol{A}(\boldsymbol{z} - \boldsymbol{z}_0)\right) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{z}_0, \boldsymbol{A}^{-1}). \tag{2.11}$$

Given the definition of this method, it is clear that the approximation will not be good for random variables whose densities are multi-modal, or the density is different from Gaussian density which means that the approximation will not be able to capture the global properties. Also, it is difficult to quantify the approximation error in practice. MCMC, which we describe next, comes with central limit theorem guarantees for convergence to the target distribution.

2.2.2 Markov Chain Monte Carlo sampling

The fully-Bayesian inference approach finds the posterior distribution of both of the unobserved quantities, $\mathbf{f}$ and $\boldsymbol{\theta}$, given the observations $\mathbf{y}$. However, for the purposes of this section we assume a simple, non-hierarchical observation model for our data, parametrised by vector $\boldsymbol{\theta}$, and with known density $p(\mathbf{y}|\boldsymbol{\theta})$. We treat $\boldsymbol{\theta}$ as a random vector over which we place a prior $p(\boldsymbol{\theta})$.

This overview follows closely Gelman (2014, ch. 11-12) and Rogers and Girolami (2017, ch. 4, 9). The main idea lies in drawing correlated samples from an approximate distribution of the posterior and then correcting those draws to better approximate the target distribution. The MCMC methods create an ergodic *Markov Process* whose stationary distribution is the quantity of interest $p(\boldsymbol{\theta}|\mathbf{y})$ of which we have access to only an unnormalised version, i.e. $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.

The samples are drawn sequentially such that the sequence forms a *Markov Chain*. In other words, the distribution of the next sample $\boldsymbol{\theta}^t$ given all the previous samples depends only on the last value, $\boldsymbol{\theta}^{t-1}$. Stating without a proof, if the desired stationary distribution and a particular set of transition rules (transition matrix if the state space is discrete, transition kernel otherwise) satisfy the *detailed balance condition*, the particular set of transition rules converges to the desired stationary distribution. The detailed balance condition is given by

$$p(\boldsymbol{\theta}^i|\mathbf{y})p(\boldsymbol{\theta}^j|\boldsymbol{\theta}^i) = p(\boldsymbol{\theta}^j|\mathbf{y})p(\boldsymbol{\theta}^i|\boldsymbol{\theta}^j),$$

where $p(\boldsymbol{\theta}^i|\boldsymbol{\theta}^j)$ is the probability of moving from state j to state i. Detailed balance condition states that being in state j and moving to state i is equally likely as being in state i and moving to state j (Rogers and Girolami, 2017). One of the most important algorithms that simulates an ergodic Markov Chain and preserved the detailed balance condition is *Metropolis-Hastings* algorithm. We also consider its variant *Hamiltonian Monte Carlo*.

Metropolis-Hastings Algorithm

Metropolis-Hastings algorithms generates a set of samples $\boldsymbol{\theta}^1, \boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^{N_s}$, where N_s is the number of samples/iterations. Each iteration involves two steps:

1. *Proposal:* at iteration t, the algorithm proposes a candidate sample $\boldsymbol{\theta}^*$ from the previous sample, $\boldsymbol{\theta}^{t-1}$. The proposal can be anything as long as we define the density
$$p(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1}).$$
In practice, the choice of proposal determines convergence and mixing properties of the chain. The most important requirement is the ability to efficiently sample from that density.

2. *Acceptance/Rejection:* the candidate sample, $\boldsymbol{\theta}$, is then tested whether it should be accepted or not. It is this step that ensures that the Markov chain coverges to the target distribution. If accepted, we set $\boldsymbol{\theta}^t = \boldsymbol{\theta}^*$, otherwise $\boldsymbol{\theta}^t = \boldsymbol{\theta}^{t-1}$. The proposed sample is accepted with probability r, such that
$$r = \frac{p(\boldsymbol{\theta}^*|\mathbf{y})/p(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})}{p(\boldsymbol{\theta}^{t-1}|\mathbf{y})/p(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*)}.$$
Note that if the proposal density is symmetric, $p(\boldsymbol{\theta}^{t-1}|\boldsymbol{\theta}^*) = p(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{t-1})$, we recover the simplified version of the algorithm called *Metropolis algorithm*.

The details of why Metropolis-Hastings scheme preserves the detailed balance conditions can be found in Rogers and Girolami (2017, sec. 9.3).

From the theoretical perspective, the algorithm is guaranteed to converge to the target distribution if we sample for long enough. In practice, it is not trivial to claim that the chain has converged. Another issue is within-sequence correlation of the draws as the new sample is proposed based on the previous sample, hence the

correlation. Inferences using correlated samples are usually less precise than the ones from the same number of independent samples. We give possible solutions to these problems after we introduce Hamiltonian Monte Carlo.

Tuning of the proposal distribution can hugely alleviate the two problems above. The ideal jumping distribution would closely approximate the target posterior distribution. *Adaptive proposals* is a popular and simple approach to generate draws that are less likely to get rejected (Gelman, 2014). In this approach, existing samples are used to approximate the target distribution which is then used as the proposal density. In the case of Gaussian random walk proposals it amounts to updating the covariance matrix of the proposal distribution with the empirical covariance. Additionally, the size of the proposal jumps can be scaled appropriately if the acceptance is either too high, or too low. Gelman (2014) state that the optimal acceptance rate is around 25%. For a detailed treatment of adaptive MCMC algorithms see Roberts and Rosenthal (2009).

Hamiltonian Monte Carlo

Random walk Metropolis-Hastings algorithm may lead to inefficient explorations of the sample space, especially in the case of oddly-shaped densities and in higher dimensions. Improved proposal densities can alleviate this issue, but as the number of dimensions increases these measures become less effective. *Hamiltonian Monte Carlo* is a variant of Metropolis-Hastings which takes advantage of the gradients of the target distribution in the proposal allows more rapid exploration of the sample space, even in a high-dimensional target space. For each component $\boldsymbol{\theta}_i$ of the target space, the scheme adds a 'momentum' variable $\boldsymbol{\phi}_j$. Subsequently, $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are updated jointly in a series of updates in order to propose a new sample $(\boldsymbol{\theta}^*, \boldsymbol{\phi}^*)$ that is then accepted/rejected.

The proposal is largely driven by the momentum variable. The proposal step starts with drawing a new value of $\boldsymbol{\phi}$ from $p(\boldsymbol{\phi})$ which needs to be specified. Then in a series of user-specified steps, L, the momentum variable $\boldsymbol{\phi}$ is updated based on the gradient of the log of the target density, and $\boldsymbol{\theta}$ is moved based on the momentum. Usually, the distribution of the momentum variable is $\mathcal{N}(0, M)$, where M is the so called 'mass' matrix. A diagonal matrix is often chosen in order to be able efficiently sample from the momentum distribution. The full steps of the procedure are given in algorithm 1.

In contrast to Random Walk Metropolis-Hastings algorithm, the proposed sample at each iteration is independent from the sample in the previous iteration. This is thanks to an independent draw from the momentum variable at the beginning of each iteration. The reason why HMC is suitable for high-dimensional problems is that in order to efficiently explore the target space, the algorithm exploits gradients of log of the target distribution. The gradients need to be available analytically, otherwise numerical differentiation would require too many target density evaluations as the number of dimensions increases.

The same considerations with regards to the convergence need to be taken. The performance of the algorithm can be tuned in three ways: (i) choice of the momentum distribution, which in the version above requires specifying the mass matrix, (ii) adjusting the scaling factor of the leapfrog step, ϵ, (iii) the number of leapfrog steps, L. Gelman (2014) 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

Algorithm 1: HAMILTONIAN MONTE CARLO as given in Gelman (2014).

Input: $p(\boldsymbol{\theta}|\mathbf{y})$: unnormalised target density , $p(\boldsymbol{\phi})$: momentum density and its mass matrix $\boldsymbol{M}$, L: leapfrog steps, ϵ: scaling factor

Output: A list of samples from $p(\boldsymbol{\theta}|\mathbf{y})$.

for $t \leftarrow 1, 2, \ldots$ **do**
  Sample $\boldsymbol{\phi}$ from $p(\boldsymbol{\phi})$
  for $i \leftarrow 1$ ***to*** L **do**
    $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{1}{2}\epsilon \frac{d \log p(\boldsymbol{\theta}|\mathbf{y})}{d\boldsymbol{\theta}}$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \epsilon \boldsymbol{M}^{-1}\boldsymbol{\phi}$
    $\boldsymbol{\phi} \leftarrow \boldsymbol{\phi} + \frac{1}{2}\epsilon \frac{d \log p(\boldsymbol{\theta}|\mathbf{y})}{d\boldsymbol{\theta}}$
    $r = \frac{p(\boldsymbol{\theta}^*|\mathbf{y})p(\boldsymbol{\phi}^*)}{p(\boldsymbol{\theta}^{t-1}|\mathbf{y})p(\boldsymbol{\phi}^{t-1})}$
    $\boldsymbol{\theta}^t = \begin{cases} \boldsymbol{\theta}^* & \text{with probability } \min(r, 1) \\ \boldsymbol{\theta}^{t-1} & \text{otherwise} \end{cases}$
return $[\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots]$

distribution, $(\mathrm{Cov}(\boldsymbol{\theta}|\mathbf{y}))^{-1}$. This can be achieved by a pre-run from which empirical covariance matrix can be computed.

Convergence diagnostics and correlation issues

Firstly, we would like to have a well-defined process which would tell us if our chain of samples has converged to the target distribution. The simplest and first thing to do is to get rid of the first samples, so-called *burn-in* samples. After the burn-in samples have been discarded we would like to determine if the samples indeed converged to the target distribution. A popular heuristic, developed by Gelman and Rubin (1992), uses multiple chains started at different starting points, splitting the chains in half, and then assessing how within-chain variance compares to between-chains variance. We give an updated version of the original heuristic, as presented in Gelman (2014).Define ϕ to be a single component of $\boldsymbol{\theta}$. We give details of the procedure to assess if the samples of ϕ converged to the correct posterior distribution. This procedure is repeated for each quantity of interest. Let m be the number of chains after splitting, and n be the number of samples in each chain. Then, we label the samples ϕ_{ij} $(i = 1, \ldots, n; j = 1, \ldots, m)$, and compute within-sequence variance, W, and between-sequence variance, B:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} \left(\overline{\psi}_{.j} - \overline{\psi}_{..}\right)^2, \text{ where } \quad \overline{\psi}_{.j} = \frac{1}{n}\sum_{i=1}^{n} \psi_{ij}, \quad \overline{\psi}_{..} = \frac{1}{m}\sum_{j=1}^{m} \overline{\psi}_{.j} \tag{2.12}$$

$$W = \frac{1}{m}\sum_{j=1}^{m} s_j^2, \quad \text{where} \quad s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(\psi_{ij} - \overline{\psi}_{.j}\right)^2. \tag{2.13}$$

The between-sequence variance estimate contains the factor n as it is based on the variance of within-sequence means. Marginal posterior variance $\mathrm{Var}(\phi|\mathbf{y})$ can the be estimated by a weighted average of B and W

$$\widehat{\mathrm{var}}^{+}(\psi|\mathbf{y}) = \frac{n-1}{n}W + \frac{1}{n}B.$$

This quantity overestimates the true marginal variance, but is unbiased under stationarity or in the limit $n \to \infty$. Then we compute the ratio $\widehat{R}$

$$\widehat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}(\psi|\mathbf{y})}{W}}.$$

This ration should decrease to 1 as $n \to \infty$. In practice, if the ratio is still high, we should keep sampling further.

Secondly, the issue of within-sequence correlation which can result in less precise inferences. For this reason, samplers need to be run for a longer period of time to achieve a desired level of accuracy. From the practical standpoint, if computer memory is insufficient, the samples can be *thinned*. The thinned sample will usually result in a lower within-sequence correlation of the samples but the information from those samples is inevitably lost. One of the most common approaches to monitor the 'quality' of the drawn sequence is called *effective sample size*. It is the size k of independent samples from the quantity of interest whose average has the same variance as the average of the drawn sequence for the quantity of interest.

$$k = \frac{mn}{1 + 2\sum_{t=1}^{\infty} \rho_t},$$

where ρ_t is the autocorrelation of the sequence at lag t, m and n denote the same quantities as in convergence discussion. In practice, ρ_t is estimated using spectral methods from the time series modelling. For more details, see Gelman (2014).

2.2.3 Fast linear algebra methods

Inference of the properties of log-Gaussian Cox processes involves expensive linear algebra operations due to the presence of the Gaussian process prior $p(\mathbf{f}|\boldsymbol{\theta})$. Scalability of models to large grids demands efficient linear algebra operations. This will become more clear in Chapter 3, but for now it suffices to say that the inference will require the following operations involving the covariance matrix $\boldsymbol{K}$ of the GP prior: matrix inverse ($\boldsymbol{K}^{-1}$), matrix-vector products ($\boldsymbol{K}\boldsymbol{x}$), determinant ($\det(\boldsymbol{K})$), gradient of $\boldsymbol{K}$ with respect to the hyper-parameters $\boldsymbol{\theta}$ ($\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}$), and eigendecomposition ($\boldsymbol{K} = \boldsymbol{Q}\Lambda\boldsymbol{Q}^{\top}$). Additionally, we will also require $\log\det(\boldsymbol{K}+\boldsymbol{W})$, where $\boldsymbol{W}$ is a diagonal matrix. Since the size of the matrix scales with the number of observed input points, it is vital that we can perform the operations efficiently.

One approach to make linear algebra operations efficient on the covariance matrix $\boldsymbol{K}$ is to assume special structure of the matrix. Motivated by the works of Saatçi (2012) and Flaxman et al. (2015), we will assume that the covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$ is a product of D covariance functions, one per each dimension. For example, in two dimensions we can write

$$k((x_1, x_2), (x_1', x_2')) = k_1(x_1, x_1')k_2(x_2, x_2').$$

Using the result from section 2.1.2 that the product of two valid kernels is a valid kernel, together with the assumption that all the input locations are on a Cartesian grid, leads to the covariance matrix that has a special form - it is a Kronecker product of D covariance matrices, each of which is computed by projecting the input data onto the respective dimension. Some of the linear algebra operations are able to exploit Kronecker structure of matrices in order to perform computations faster. Next, we

define Kronecker product and summarise results which follow from the Kronecker structure of a matrix.

Kronecker methods

Firstly, we give a definition of Kronecker product as given in Golub and Van Loan (2013, p. 27).

Definition 2.3 (Kronecker product)**.** If $\boldsymbol{A}$ is an m_1-by-n_1 matrix and $\boldsymbol{B}$ is an m_2-by-n_2 matrix, then their Kronecker product is an m_1-by-n_1 block matrix whose (i, j) block is the m_2-by-n_2 matrix $a_{ij}B$. For example, if

$$C = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

then

$$C = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{11}b_{13} & a_{12}b_{11} & a_{12}b_{12} & a_{12}b_{13} \\ a_{11}b_{21} & a_{11}b_{22} & a_{11}b_{23} & a_{12}b_{21} & a_{12}b_{22} & a_{12}b_{23} \\ a_{11}b_{31} & a_{11}b_{32} & a_{11}b_{33} & a_{12}b_{31} & a_{12}b_{32} & a_{12}b_{33} \\ a_{21}b_{11} & a_{21}b_{12} & a_{21}b_{13} & a_{22}b_{11} & a_{22}b_{12} & a_{22}b_{13} \\ a_{21}b_{21} & a_{21}b_{22} & a_{21}b_{23} & a_{22}b_{21} & a_{22}b_{22} & a_{22}b_{23} \\ a_{21}b_{31} & a_{21}b_{32} & a_{21}b_{33} & a_{22}b_{31} & a_{22}b_{32} & a_{22}b_{33} \end{bmatrix}.$$

We state basic properties of Kronecker products that are relevant for the LGCP inference. The properties are taken from Golub and Van Loan (2013, sec. 12.3) and are stated without proof. Let $\boldsymbol{A}$ be an m-by-m matrix and $\boldsymbol{B}$ be an n-by-n matrix, then

$$\text{Transpose:} \quad (\boldsymbol{A} \otimes \boldsymbol{B})^\top = \boldsymbol{A}^\top \otimes \boldsymbol{B}^\top, \tag{2.14}$$

$$\text{Inverse:} \quad (\boldsymbol{A} \otimes \boldsymbol{B})^{-1} = \boldsymbol{A}^{-1} \otimes \boldsymbol{B}^{-1}, \tag{2.15}$$

$$\text{Determinant:} \quad \det(\boldsymbol{A} \otimes \boldsymbol{B}) = \det(\boldsymbol{A})^n \det(\boldsymbol{B})^m, \tag{2.16}$$

$$\text{Trace:} \quad \operatorname{tr}(\boldsymbol{A} \otimes \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{A}) \operatorname{tr}(\boldsymbol{B}). \tag{2.17}$$

Other properties which Kronecker products exhibit and are relevant for LGCP inference are eigendecomposition, gradients, and matrix-vector product. In this exposition, we follow Saatçi (2012). Throughout, let the n-by-n covariance matrix $\boldsymbol{K}$ be a Kronecker product of D matrices such that $\boldsymbol{K} = \otimes_d \boldsymbol{K}_d$:

Eigendecomposition Let $\boldsymbol{K}_d = \boldsymbol{Q}_d \boldsymbol{\Lambda}_d \boldsymbol{Q}_d^\top$ be the eigendecomposition of $\boldsymbol{K}_d$. Then, the eigendecomposition of $\boldsymbol{K}$ is given by $\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$, where $\boldsymbol{Q} = \otimes_d \boldsymbol{Q}_d$, and $\boldsymbol{\Lambda} = \otimes_d \boldsymbol{\Lambda}_d$.

Gradient The gradient of the covariance matrix $\boldsymbol{K}$ with respect to the hyperparameters $\boldsymbol{\theta}_i$ can be decomposed as follows:

$$\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i} = \sum_{d=1}^{D} \frac{\partial \boldsymbol{K}_d}{\partial \boldsymbol{\theta}_i} \otimes \left(\bigotimes_{j \neq d} \boldsymbol{K}_j \right).$$

Matrix-vector multiplication Saatçi (2012) show that $\boldsymbol{b} = (\otimes_d \mathbf{K}_d)\, \boldsymbol{x}$ can be computed efficiently in $\mathcal{O}(n)$ time and space. By representing $\boldsymbol{x}$ as a tensor $\mathbf{T}^{\boldsymbol{x}}_{j_D \ldots j_1}$,

the vector $\boldsymbol{b}$ can be represented as a series of matrix-tensor products and tensor transpose operations:

$$\boldsymbol{b} = \text{vec}\left(\left(\boldsymbol{K}_1 \ldots \left(\boldsymbol{K}_{D-1}\left(\boldsymbol{K}_D \mathsf{T}^{\boldsymbol{x}}\right)^\top\right)^\top\right)^\top\right),$$

where matrix-tensor products of the form $\mathsf{Z} = \boldsymbol{A}\mathsf{T}$ are defined as:

$$\mathsf{Z}_{i_1 \ldots i_D} = \sum_{k=1}^{\text{size}(\mathsf{T},1)} \boldsymbol{A}_{i_1 k} \mathsf{T}_{k i_2 \ldots i_d}.$$

The operator $\top$ performs a cyclic permutation of the indices of the tensor:

$$\mathsf{Y}^\top_{i_D i_1 \ldots i_{D-1}} = \mathsf{Y}_{i_1 \ldots i_D}.$$

Log-determinant approximations

Another computation that we require is the calculation of $\log\det(\boldsymbol{K} + \boldsymbol{W})$, where $\boldsymbol{W}$ is a diagonal matrix.Again, let the n-by-n covariance matrix $\boldsymbol{K}$ be a Kronecker product of D matrices such that $\boldsymbol{K} = \otimes_d \boldsymbol{K}_d$. There are methods that exploit the Kronecker structure and thus reduce the computational cost, which, before optimisation, is $\mathcal{O}(n^3)$.

Fiedler bound method Fiedler (1971) showed that for Hermitian positive semi-definite matrices $\boldsymbol{U}$ and $\boldsymbol{V}$:

$$\prod_i (u_i + v_i) \leq \det(\boldsymbol{U} + \boldsymbol{V}) \leq \prod_i (u_i + v_{n-i+1}),$$

where u_i and v_j are sorted eigenvalues of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Applying the result to $\log\det(\boldsymbol{K} + \boldsymbol{W})$:

$$\log\det(\boldsymbol{K} + \boldsymbol{W}) \geq \log \prod_i (e_i + w_i) \tag{2.18}$$

$$= \sum_i \log(e_i + w_i), \tag{2.19}$$

where $e_1 \leq e_2 \leq \ldots\ e_n$ are the eigenvalues of $\boldsymbol{K}$, and $w_1 \leq w_2 \leq \ldots\ w_n$ are the eigenvalues of $\boldsymbol{W}$. Thanks to the Kronecker structure of $\boldsymbol{K}$, the eigenvalues can be obtained using $\mathcal{O}(Dn^{3/D})$ operations. Obtaining eigenvalues of $\boldsymbol{W}$ is trivial because the matrix is diagonal. It is important to point out that bound is not always a good estimate as it is biased - it is acceptable for certain tasks such as optimisation, but for example cannot bed used for Monte Carlo methods.

Lanczos decomposition method This method follows closely Dong et al. (2017). Let $\boldsymbol{A} = \boldsymbol{K} + \boldsymbol{W}$. It can be shown that $\log\det(\boldsymbol{A}) = \text{tr}(\log(\boldsymbol{A}))$, where log is the matrix logarithm. For more details on matrix functions, see Higham (2008). Dong et al. (2017) then use *stochastic trace estimation*. The approach relies on

$$\text{tr}(\log(\boldsymbol{A})) = \mathbb{E}(\boldsymbol{z}^\top \log(\boldsymbol{A})\boldsymbol{z}),$$

where $\boldsymbol{z}$ is a random probe vector whose each component has mean 0 and variance 1. without going into too much detail, $\mathbb{E}(\boldsymbol{z}^\top \log(\boldsymbol{A})\boldsymbol{z})$ is computed by sampling n_z probe vectors and taking the mean. Each single computation of

$\boldsymbol{z}^\top \log(\boldsymbol{A})\boldsymbol{z}$ is approximated using Lanczos-based computation which relies on efficient computations of matrix-vector products. For a vector $\boldsymbol{x}$, $\boldsymbol{A}\boldsymbol{x}$ can be computed very efficiently due to Kronecker structure of $\boldsymbol{K}$ and $\boldsymbol{W}$ being a diagonal matrix. It turns out the $\log\det(\boldsymbol{K}+\boldsymbol{W})$ can be computed in $\mathcal{O}(n)$ computational steps. Dong et al. (2017) provides more details as well as error analysis of the approximation.

This chapter has introduced log-Gaussian Cox process as a flexible framework for modelling spatial point processes, including the inference procedures commonly used for LGCPs. By treating crime spatial pattern as a realisation of a point process, chapter 3 develops the methodology required to accurately fit the model to the data, and make inferences from the model.

3 Methodology

This chapter gives the details of methodology we use in order to achieve the objectives formulated in section 1.3. Firstly, we describe the dataset that we will use for the experiments, then we specify the discretised version of LGCP model, after which we proceed with the details of the two inference methods: the approximate Bayesian method using Laplace approximation, and the fully-Bayesian method using Markov Chain Monte Carlo sampling.

3.1 Dataset

Our study area in this work is London, United Kingdom. For the purposes of building models of criminal activity that are based on external socio-economic factors, we need to have access to both the crime data, and the socio-economic data.

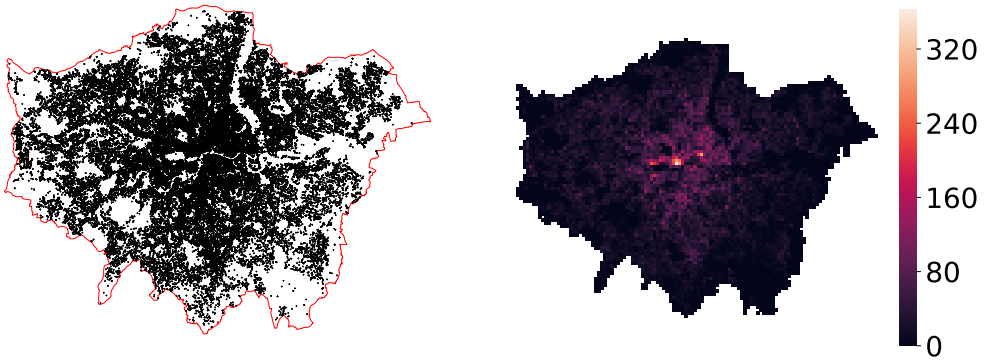
3.1.1 Crime data

The crime dataset is obtained from the UK Police website.[1] Each entry refers to an instance of a crime with the following fields recorded:

Location Only approximate location is reported due to privacy reasons. The location is published as latitude and longitude.

Month Again, due to privacy concerns only the month of an incident is published.

Crime type Police use 15 crime type categories according to which each crime is classified. The full list of the categories can be found in table B.1.



(A) (B)

FIGURE 3.1: Burglaries, London 2016. *(Left)* Point pattern of crime over the map of London. *(Right)* Grid discretisation of the point pattern. Each cell represents an area of 500m by 500m. and the colour intensity indicates the count.

[1] https://data.police.uk/data/

For example, figure 3.1a shows the point pattern of burglaries in London throughout the year 2016. In order to be able to induce the Kronecker structure for the inference (see section 2.2.3), we will discretise the LGCP process onto a Cartesian grid. A grid discretisation where one cell corresponds to an area of 500m-by-500m is shown in figure 3.1b. The discretisation is done by summing the number of incidents that fall in to a cell. We treat the point observations as a realisation of the underlying point process which is a random counting measure defined on the domain.

3.1.2 Socio-economic covariates

The modelling framework for Log-Gaussian Cox process which we introduced in section 2.1 allows specifying a deterministic part of the intensity process, such that the intensity function (equation 2.1) is replaced with

$$\Lambda(\boldsymbol{x}) = \exp\left(\boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\beta} + f(\boldsymbol{x})\right), \tag{3.1}$$

where $\boldsymbol{h}(\boldsymbol{x})$ is a set of spatially indexed covariates, $\boldsymbol{\beta}$ is the vector of their coefficients, and we set the mean function of the GP prior, $f(\boldsymbol{x})$ to be $\mathbf{0}$. Another way of thinking about this formulation is that of Poisson regression that accounts for spatial correlation through the term $f(\boldsymbol{x})$.

The dataset with socio-economic variables is sourced from the UK 2011 Census (Office For National Statistics, National Records Of Scotland, and Northern Ireland Statistics And Research Agency, 2016) and the portal of Mayor of London which publishes statitics more frequently.[2] Most of the statistics are collected at the spatial granularity of an area unit called *Output area* (OA), but some of them are only available at coarser levels: *Lower layer super output area* (LSOA), and *Middle layer super output area* (MSOA). The OAs are built from clusters of adjacent postcodes such that the OAs have similar population sizes and are as socially homogeneous as possible. Each OA contains at least 100 persons and 40 households. The OAs roll up into LSOAs and MSOAs. In order to align the data collected at OA, LSOA, and MSOA levels, we overlay the geometry of those output areas on top of a grid. For a specific cell of the grid, the value of the output area with the largest intersection with the cell will be used. Figure 3.2 shows the median household income data collected at the LSOA level *(Left)* and the projection onto a grid *(Right)*.

Fourteen covariates were obtained from the sources above. The covariates along with four summary statistic are shown in table 3.1.

3.2 Model

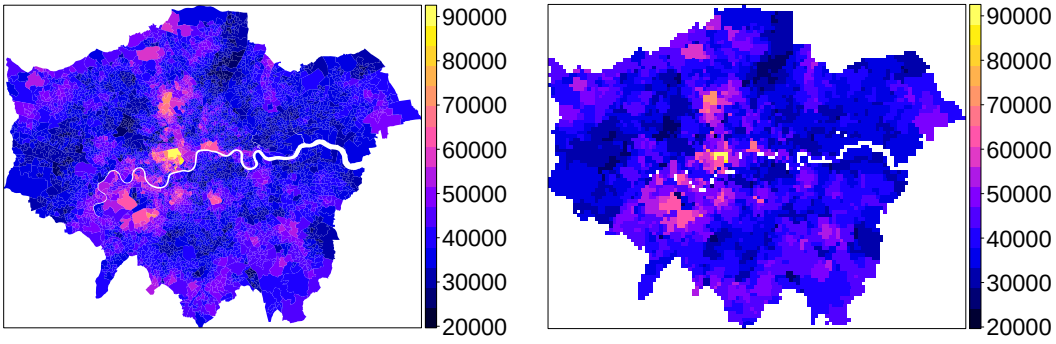
By combining the definition of Log-Gaussian Cox Process (definition 2.1) with the deterministic component incorporating socio-economic covariates introduced in section 3.1.2, the count of events in a region A, $\mathbf{y}_A$, is modelled as follows:

$$\mathbf{y}_A \sim \text{Poisson}\left(\int_{\boldsymbol{x}\in A} \lambda(\boldsymbol{x})\right) \tag{3.2}$$

$$\lambda(\boldsymbol{x}) = \exp\left(\boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\beta} + f(\boldsymbol{x})\right) \tag{3.3}$$

$$f(\boldsymbol{x}) \sim \mathcal{GP}(\mathbf{0}, k_{\boldsymbol{\theta}}(\cdot,\cdot)) \tag{3.4}$$

[2] https://data.london.gov.uk/



(A) (B)

FIGURE 3.2: Median household income (in £) plotted over the map of London. *(Left)* Heatmap of the data at the LSOA level as it was originally collected. *(Right)* Heatmap of the data obtained by projecting the LSOA values onto a grid.

where $\boldsymbol{x}$ refers to a point in the spatial domain. Given an observation of crime counts at n locations that are a subset of a grid, the discretised version of the model is

$$\mathbf{y}_i|\mathbf{f}_i, \boldsymbol{\beta} \sim \text{Poisson}\left(\exp\left(\boldsymbol{Z}_{i\cdot}^\top \boldsymbol{\beta} + \mathbf{f}_i)\right)\right), \tag{3.5}$$

where $\boldsymbol{Z}$ is a matrix whose i-th row is the set of covariates at location i.

To complete the specification of the model, we set the covariance function, $k_{\boldsymbol{\theta}}(\cdot,\cdot)$, to be the Matérn kernel with the smoothness parameter $\nu = 2.5$. Matérn class is a recommended choice in spatial statistics because of its flexibility to specify both smoothness (ν) and lengthscale (ℓ). (Stein, 1999). In practice it is hard to jointly estimate ℓ and ν due to identifiability issues. Often, ν is set a priori to one of the values that make it computationally efficient (see section 2.1.2). We expect the count observations to vary smoothly, so we fix $\nu = 5/2$. Additionally, we specify variance of the GP prior, σ^2. This quantity can be thought of as 'amplitude' of the prior. Afterwards, the final form of the covariance function is

$$k_{\nu=5/2}(r) = \sigma^2\left(1 + \frac{\sqrt{5}r}{\ell} + \frac{5r^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}r}{\ell}\right). \tag{3.6}$$

As a consequence of the formulation above, the likelihood of the n observations $\mathbf{y}$ given the corresponding latent variables $\mathbf{f}$, parameters $\boldsymbol{\beta}$, and hyper-paramters $\boldsymbol{\theta}$ is

$$p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}) = p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta})p(\mathbf{f}|\boldsymbol{\theta})p(\boldsymbol{\beta})p(\boldsymbol{\theta}), \tag{3.7}$$

where $p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta})$ factorises as $\prod_i p(\mathbf{y}_i|\mathbf{f}_i, \boldsymbol{\beta})$, with each $p(\mathbf{y}_i|\mathbf{f}_i, \boldsymbol{\beta})$ being the Poisson density, $p(\mathbf{f}|\boldsymbol{\theta})$ is a multivariate Gaussian density with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{K}$, $p(\boldsymbol{\beta})$ is the prior for the coefficients of the covariates, and $p(\boldsymbol{\theta})$ is the prior for the hyper-parameters.

3.2.1 Model performance assessment

Our ultimate goal is to create a model that is as truthful as possible to the true process that generated the observations.

Covariate	min	max	mean	SD
Median House Price (£)	125238.00	4650000.00	513997.56	327316.94
Percentage of Houses	0.00	1.00	0.69	0.26
Population Density	2.90	246.70	52.25	40.53
Percentage of L4 Educated	7.98	69.69	29.17	11.57
Percentage of Immigrants	0.37	0.82	0.64	0.09
Percentage of Unemployed	0.01	0.14	0.05	0.02
Median Household Income	25690.00	88090.00	39899.78	7413.65
Percentage of Households with No Car	0.03	0.84	0.30	0.18
Percentage of Non-religious	0.01	0.45	0.20	0.08
Median Age	20.00	55.00	36.69	5.71
Percentage of Age under 16	0.03	0.40	0.20	0.05
Percentage of Age 16 to.24	0.05	0.66	0.12	0.04
Percentage of Age 25 to 65	0.24	0.77	0.55	0.06
Percentage of Age over 65	0.01	0.36	0.14	0.06

TABLE 3.1: List of covariates considere for fitting the model and summary statistics

Root mean square error (RMSE)

As an absolute measure of predictive performance of a model, we first consider *root mean square error* (RMSE). It is a metric that measures the difference between the observed values and the values predicted by the model. By letting $\hat{\mathbf{y}}$ be the estimate obtained by the model, and $\mathbf{y}$ be the observed sample, RMSE is obtained by computing

$$\text{RMSE} = \sqrt{\mathbf{E}\left((\hat{\mathbf{y}} - \mathbf{y})^2\right)} \tag{3.8}$$

This metric unfortunately does not take into account the complexity of the model and relying solely on this metric can lead to models that overfit the training data and perform poorly on out-of-sample data.

Watanabe-Akaike information criterion (WAIC)

There have been a number of approaches proposed in the literature to measure predictive performance of a model while taking into account complexity of the model. Given that we operate in the Bayesian setting, *Watanabe-Akaike Information Criterion* (WAIC) is the most applicable approach as it uses the entire posterior. The criterion computes the log pointwise predictive density and corrects for the effective number of parameters. The definition we present is adapted from Gelman (2014). Given a set of observations $\mathbf{y}_i$, where $i \in \{1, \ldots, n\}$, and the vector of parameters or unobserved quantities $\boldsymbol{\theta}$, of which we obtain S posterior samples $(\boldsymbol{\theta}^1, \ldots \boldsymbol{\theta}^S)$, the log pointwise predictive density is computed as

$$\text{lppd} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{i=1}^{S} p(\mathbf{y}_i | \boldsymbol{\theta}^s) \right). \tag{3.9}$$

The WAIC correction term that estimates the number of effective parameters is defined as

$$p_{\text{WAIC}} = \sum_{i=1}^{n} V_{s=1}^{S}(\log p(\mathbf{y}_i|\boldsymbol{\theta}^s)), \tag{3.10}$$

where $V_{s=1}^{S}$ computes the sample variance. The final value of the WAIC criterion is then given by

$$\text{WAIC} = -2\text{lppd} + 2p_{\text{WAIC}}. \tag{3.11}$$

The WAIC does not have units and is not normalised. Therefore, it is mainly used as a relative measure of predictive performance. Lower values of WAIC indicate better predictive performance.

3.3 Approach 1: Laplace approximation for approximate Bayesian inference

In our first approach to inference, we treat the hyper-parameters $\boldsymbol{\theta}$ as fixed values and exclude the covariates, i.e. $\boldsymbol{Z} = \mathbf{0}$. The inference will proceed by inferring the posterior distribution of the latent field $\mathbf{f}$ given the observed data and the hyper-parameters, $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$, using Laplace approximation. Accurate inferences and predictions require knowing the true value of $\boldsymbol{\theta}$. The optimal value of $\boldsymbol{\theta}$ is found by maximising the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$. This approach is inspired from and closely follows Flaxman et al., 2015.

First, using Bayes's theorem, we obtain an unnormalised version of $p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta})$:

$$p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})}{p(\mathbf{y})} \tag{3.12}$$

$$\propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta}). \tag{3.13}$$

By defining $\Psi(\mathbf{f}) \coloneqq \log p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}) \overset{\text{const}}{=} \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|\boldsymbol{\theta})$, the Laplace approximation (see equation 2.11 in section 2.2.1) of the posterior density is given by

$$p(\mathbf{f}|\mathbf{y}, \boldsymbol{\theta}) \approx \mathcal{N}\left(\mathbf{f}|\hat{\mathbf{f}}, -(\nabla\nabla_{\mathbf{f}}\Psi(\hat{\mathbf{f}}))^{-1}\right). \tag{3.14}$$

The posterior mode, $\hat{\mathbf{f}}$, is found using Newton method because solving $\nabla\Psi(\mathbf{f}) = \mathbf{0}$ is not analytically tractable. The update step of the Newton method is:

$$\mathbf{f}^{\text{new}} = \mathbf{f} - (\nabla\nabla_{\mathbf{f}}\Psi(\mathbf{f}))^{-1}\nabla_{\mathbf{f}}\Psi(\mathbf{f}). \tag{3.15}$$

The optimal value of $\boldsymbol{\theta}$ is found by maximising the marginal likelihood

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\theta}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f} \\ &= \int \exp\left(\Psi(\mathbf{f})\right) d\mathbf{f}. \end{aligned} \tag{3.16}$$

Using a Taylor expansion of $\Psi(\mathbf{f})$ around the mode $\hat{\mathbf{f}}$, $\Psi(\mathbf{f}) \simeq \Psi(\hat{\mathbf{f}}) - \frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^\top \boldsymbol{A}(\mathbf{f}-\hat{\mathbf{f}})$, where $\boldsymbol{A} = -\nabla\nabla\Psi(\mathbf{f})|_{\hat{\mathbf{f}}}$. Substituting the 2nd order Taylor expansion estimate of $\Psi(\mathbf{f})$

into equation 3.16, we obtain

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\theta}) \simeq q(\mathbf{y}|\boldsymbol{\theta}) &= \int \exp\left(\Psi(\hat{\mathbf{f}}) - \frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^\top \boldsymbol{A}(\mathbf{f}-\hat{\mathbf{f}})\right) d\mathbf{f} \\ &= \exp\left(\Psi(\hat{\mathbf{f}})\right) \int \exp\left(-\frac{1}{2}(\mathbf{f}-\hat{\mathbf{f}})^\top \boldsymbol{A}(\mathbf{f}-\hat{\mathbf{f}})\right) d\mathbf{f} \\ &= \exp\left(\Psi(\hat{\mathbf{f}})\right) \det\left(2\pi \boldsymbol{A}^{-1}\right)^{1/2}. \end{aligned} \tag{3.17}$$

3.3.1 Computations

In this section, we give the full algorithm that infers the distribution of the latent field $\mathbf{f}$ given the observed counts $\mathbf{y}$ and the fixed hyperparameters $\boldsymbol{\theta}$. The algorithm also computes the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ in order to find the optimal value of $\boldsymbol{\theta}$. Before we present the full algorithm, we give the full details of the computations required.

In order to write the full expression for the Newton step and the marginal likelihood (equation 3.15), we need

$$\begin{aligned} \nabla\Psi(\mathbf{f}) &= \nabla \log p(\mathbf{y}|\mathbf{f}) + \nabla \log p(\mathbf{f}|\boldsymbol{\theta}) \\ &= \nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{K}^{-1}\mathbf{f} \end{aligned} \tag{3.18}$$

$$\begin{aligned} \nabla\nabla\Psi(\mathbf{f}) &= \nabla\nabla \log p(\mathbf{y}|\mathbf{f}) + \nabla\nabla \log p(\mathbf{f}|\boldsymbol{\theta}) \\ &= -\boldsymbol{W} - \boldsymbol{K}^{-1}, \end{aligned} \tag{3.19}$$

where $\boldsymbol{W} \coloneqq \nabla\nabla \log p(\mathbf{y}|\mathbf{f})$ is an $n \times n$ diagonal matrix. For the Poisson observation model, $\boldsymbol{W}_{ii} = \exp(\mathbf{f}_i)$, and $\nabla \log p(\mathbf{y}|\mathbf{f}) = \mathbf{y} - \exp(\mathbf{f})$.

Rasmussen and Williams (2006) show that most of the computations can be expressed using the symmetric positive definite matrix

$$\boldsymbol{B} = \boldsymbol{I} + \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{W}^{\frac{1}{2}}, \tag{3.20}$$

whose eigenvalues are bounded below by 1, and bounded above by $1 + n \max_{ij}(\boldsymbol{K}_{ij})/4$. The bounds guarantee that the matrix is well-conditioned for many covariance matrices which results in numerically stable matrix computations.

The Newton step

Using the expressions above, the Newton step is given as

$$\begin{aligned} \mathbf{f}^{\text{new}} &= \mathbf{f} + \left(\boldsymbol{K}^{-1} + \boldsymbol{W}\right)^{-1} \left[\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{K}^{-1}\mathbf{f}\right] \\ &= \boldsymbol{K}\left(\boldsymbol{b} - \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{B}^{-1} \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{b}\right), \end{aligned} \tag{3.21}$$

where $\boldsymbol{b} = \nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{W}\mathbf{f}$. For the full derivation, see section C.1.1. In order to avoid Cholesky factorisation of $\boldsymbol{B}$ which requires $\mathcal{O}(n^3)$ operations, we perform the updates on $\boldsymbol{a} = \boldsymbol{K}^{-1}\mathbf{f}$. As a result, the update step can be expressed as

$$\boldsymbol{B}\boldsymbol{z} = \boldsymbol{W}^{\frac{1}{2}} \boldsymbol{K} \boldsymbol{b}, \tag{3.22}$$

where $\boldsymbol{z} = \boldsymbol{W}^{-\frac{1}{2}}(\boldsymbol{b} - \boldsymbol{a})$. Equation 3.22 can be efficiently solved for $\boldsymbol{z}$ using *conjugate gradient* method utilising the Kronecker structure of $\boldsymbol{K}$ and the diagonal structure of $\boldsymbol{W}$. We adapt conjugate gradient method presented in (Shewchuk, 1994) to account

for the special structure of $\boldsymbol{W}$ and $\boldsymbol{K}$. After solving for $\boldsymbol{z}$, the new value of $\boldsymbol{a}$ is given as

$$\boldsymbol{a} = \boldsymbol{b} - \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{z}.$$

Marginal likelihood computation

For finding the optimal hyper-parameters $\boldsymbol{\theta}$, we optimise the log of the marginal likelihood. For a given $\boldsymbol{\theta}$, using equation 3.17, we obtain

$$\log p(\mathbf{y}|\boldsymbol{\theta}) \overset{\text{const}}{=} \log p(\hat{\mathbf{f}}|\mathbf{y}) + \frac{1}{2}\log\left|(2\pi(\boldsymbol{W} + \boldsymbol{K}^{-1})^{-1}\right| \tag{3.23}$$

$$= \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\hat{\mathbf{f}}^{\top}\boldsymbol{K}^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log|\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}|\,, \tag{3.24}$$

where $\hat{\mathbf{f}}$ is obtained using the Newton method described above. The full details of the derivation are shown in section C.1.2.

The only operation which is not able to exploit the Kronecker structure of $\boldsymbol{K}$ and the diagonal structure of $\boldsymbol{W}$ is $\log|\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}|$. By expressing the log determinant as $\log\left|\boldsymbol{K} + \boldsymbol{W}^{-1}\right||\boldsymbol{W}| = \log\left|\boldsymbol{K} + \boldsymbol{W}^{-1}\right| + \log|\boldsymbol{W}|$, approximation methods introduced in section 2.2.3 are applicable. In order to choose between *Fiedler bound* and *stochastic Lanczos approximation*, we designed an experiment which simulated an LGCP process. After maximising the marginal likelihood, while using the respective method for the log determinant calculation, we obtained inferred paramters $\boldsymbol{\theta}$. In our experiment, the Fiedler bound estimate was more biased. Figure 3.3 shows the results of the experiment and compares it with the calculation obtained using the true value of the determinant. Due to larger bias of the Fiedler bound method, we did not pursue the Fiedler bound further. In our implementation, the code for Lanczos approximation is provided by Gardner et al. (2018).

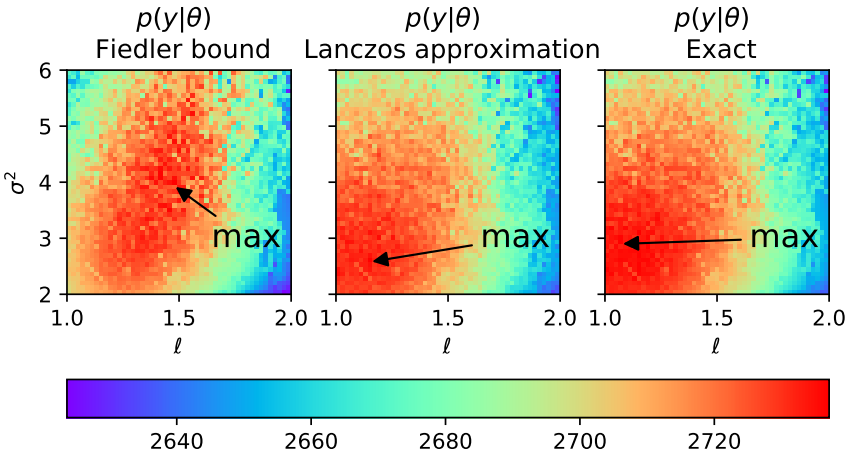


FIGURE 3.3: Log marginal likelihood of the data given the hyperparameters obtained via Laplace approximation. The method requires computation of $\log|\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}|$. We compare two approximations of log determinant against the true value and how it affects the inference. Fiedler bound approximation leads to a very different value from the true value ($\ell = 1.2$, $\sigma^2 = 3$).

Full algorithm

Combining the Newton step and the log-marginal likelihood calculation, the full algorithm for inference is given in algorithm 2.

Algorithm 2: LAPLACE APPROXIMATION of the posterior latent field.

Input: $\boldsymbol{\theta}$, $\boldsymbol{K}$ as a list of $\boldsymbol{K}_d$ matrices, $p(\mathbf{y}|\mathbf{f}, \boldsymbol{\theta})$: observation model, $\mathbf{y}$ observations

Output: $\mathbf{f}$: mode of the posterior, Z: the marginal likelihood

$\boldsymbol{a} \leftarrow \mathbf{0}$
repeat
  $\boldsymbol{f} \leftarrow \boldsymbol{K}\boldsymbol{a}$
  $\boldsymbol{W} \leftarrow \nabla\nabla \log p(\mathbf{y}|\mathbf{f})$
  $\boldsymbol{b} \leftarrow \nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{W}\mathbf{f}$
  Solve $\boldsymbol{B}\boldsymbol{z} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{b}$ using CG
  $\boldsymbol{a} \leftarrow \boldsymbol{b} - \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{z}$
until *convergence of* $\boldsymbol{a}$
$Z \leftarrow$ Lanczos-based algorithm to solve $\log|\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}|$
return $\boldsymbol{f}$, Z

3.4 Approach 2: MCMC for fully-Bayesian inference

The Laplace approximation method above has a limitation of not being able to quantify uncertainty in the hyper-parameters $\boldsymbol{\theta}$. In this section, we treat $\boldsymbol{\theta}$ as a random variable, and we consider the semi-parametric formulation with covariates included in the model. As before, we operate on a computational grid.

In order to complete the model specification we specify prior distributions for $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$. Due to the constraint of $\boldsymbol{\theta} > \mathbf{0}$ and practical reasons which will become apparent later, we transform the hyper-parameters into log-space: $\boldsymbol{\phi} = \log \boldsymbol{\theta}$. Since, we do not have any strong beliefs about the true value of $\boldsymbol{\theta}$, an uninformative prior is used:

$$\begin{aligned} \log(\boldsymbol{\theta}) &\sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\phi}}) \\ \Sigma_{\boldsymbol{\phi}} &= 10^3 \boldsymbol{I} \end{aligned} \tag{3.25}$$

In line with our objectives to quantify the explanatory variables for crime, we a priori assume that a specific covariate is not significant in explaining the crime rates. We would like the data inform us, whether this assumption is true or otherwise. Therefore, we set the prior of $\boldsymbol{\beta}$ as

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}(\mathbf{0}, \Sigma_{\boldsymbol{\beta}}) \\ \Sigma_{\boldsymbol{\beta}} &= 10\boldsymbol{I} \end{aligned} \tag{3.26}$$

Our goal is to design a scheme to obtain the joint posterior distribution of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, as well as $\mathbf{f}$ which is necessary for making predictions. It is well known that efficiently sampling from the join distribution of the latent variables and hyper-parameters is difficult. As explained in Filippone and Girolami (2014), it is extremely unlikely that a set of latent variables, parameters, and hyper-parameters that are consistent with each other and the data is proposed. One possible remedy, introduced in the same

paper for a model without covariates, is to define a Gibbs scheme where the hyper-parameters and the latent variables are sampled in turn, such that $\boldsymbol{\theta}$ is sampled from independently of $\mathbf{f}$. This is achieved by sampling from an approximation of $p(\boldsymbol{\theta}|\mathbf{y})$. We briefly considered this approach, but it turned out that, we were not fully able to utilise the Kronecker structure of our model. Because of our objective to provide a highly-scalable framework, we did not pursue this further.

The difficulty of jointly sampling from the joint posterior of the latent variables, the parameters, and the hyper-parameters can be overcome to some extent by informing the sampling proposal about the properties of the posterior distribution. Hamiltonian Monte Carlo sampling algorithm, introduced in section 2.2.1, takes advantage of the geometry of the posterior in proposing a new sample. The only requirement is that the gradient of the posterior distribution can be expressed analytically. Our model has analytic gradients, and the Kronecker structure of the covariance matrix $\boldsymbol{K}$ can be exploited. We give the full details of the derivation of the HMC scheme below.

The target distribution we want to sample from is $p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y})$ of which we have access to only an unnormalised version

$$p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) p(\mathbf{f}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\boldsymbol{\beta}). \tag{3.27}$$

Next, we require the gradient of the log posterior with respect to $[\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}]^\top$:

$$\begin{aligned} \nabla \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) &= \begin{bmatrix} \nabla_{\mathbf{f}} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) \\ \nabla_{\boldsymbol{\beta}} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) \\ \nabla_{\boldsymbol{\phi}} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) \end{bmatrix} \\ &= \begin{bmatrix} \nabla_{\mathbf{f}} \log p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) + \nabla_{\mathbf{f}} \log p(\mathbf{f}|\boldsymbol{\theta}) \\ \nabla_{\boldsymbol{\beta}} \log p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) + \nabla_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}) \\ \nabla_{\boldsymbol{\phi}} \log p(\mathbf{f}|\boldsymbol{\theta}) + \nabla_{\boldsymbol{\phi}} \log p(\boldsymbol{\theta}) \end{bmatrix} \end{aligned} \tag{3.28}$$

Note that we sample from the $\boldsymbol{\theta}$ in the log-space, hence the gradient with respect to $\boldsymbol{\phi} = \log \boldsymbol{\theta}$.

Without showing the detail (see section C.2), the components of the gradient are

$$\nabla_{\mathbf{f}} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) = [\mathbf{y} - \exp(\boldsymbol{Z}\boldsymbol{\beta} + \mathbf{f})] + \left[-\boldsymbol{K}^{-1}\mathbf{f}\right] \tag{3.29}$$

$$\nabla_{\boldsymbol{\beta}} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) = \left[\boldsymbol{Z}^\top \mathbf{y} - \boldsymbol{Z}^\top \exp(\boldsymbol{Z}\boldsymbol{\beta} + \mathbf{f})\right] + \left[-\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right] \tag{3.30}$$

$$\nabla_{\boldsymbol{\phi}_i} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) = \left[\frac{1}{2}\mathbf{f}^\top \boldsymbol{K}^{-1} \frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\phi}_i} \boldsymbol{K}^{-1}\mathbf{f} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\phi}_i}\right)\right] + \left[-(\Sigma_{\boldsymbol{\phi}}^{-1}\boldsymbol{\phi})_i\right] \tag{3.31}$$

It is clear that equation 3.29 can utilise the Kronecker structure of $\boldsymbol{K}$, resulting in $\mathcal{O}\big(Dn^{3/D}\big)$ operations required for the inverse of $\boldsymbol{K}$. Equation 3.30 is also computationally efficient as all the operations require only $\mathcal{O}(n)$ operations. For equation 3.31,

as Saatçi (2012) shows, the computation can be broken down into operations only involving the smaller matrices $\boldsymbol{K}_D$ such that $\boldsymbol{K} = \bigotimes_{d=1}^{D} \boldsymbol{K}_d$:

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\phi}_i}\right) &= \operatorname{tr}\left(\left(\bigotimes_{d=1}^{D}\boldsymbol{K}_d^{-1}\right)\left(\sum_{d=1}^{D}\frac{\partial \boldsymbol{K}_d}{\partial \boldsymbol{\phi}_i}\otimes\left(\bigotimes_{j\neq d}\boldsymbol{K}_j\right)\right)\right)\\ &= \sum_{d=1}^{D}\operatorname{tr}\left(\boldsymbol{K}_d^{-1}\frac{\partial \boldsymbol{K}_d}{\partial \boldsymbol{\phi}_i}\right)\prod_{j\neq d}\operatorname{tr}\left(\boldsymbol{K}_j^{-1}\boldsymbol{K}_j\right),\end{aligned} \tag{3.32}$$

where $\frac{\partial \boldsymbol{K}_d}{\partial \boldsymbol{\phi}_i}$ is an element-wise gradient. For example, for the Matérn-5/2 covariance matrix (see equation 3.6), the gradient with respect to the log-lengthscale, $\boldsymbol{\phi}_\ell = \log \ell$, is given by

$$\frac{\partial}{\partial \boldsymbol{\phi}_\ell}k_{\nu=5/2}(r) = \sigma^2 \exp\left(-\frac{\sqrt{5}r}{\ell}\right)\left(\frac{5r^2}{3\ell^2} + \frac{5\sqrt{5}r^3}{3\ell^3}\right). \tag{3.33}$$

The full derivation of this result is shown in section C.2.

3.4.1 HMC algorithm tuning

As already discussed in section 2.2, the performance of an HMC scheme is tuned using three parameters. Firstly, the mass matrix, $\boldsymbol{M}$, should roughly approximate the inverse of the posterior covariance of the quantities we sample from. Next, one can adjust the number of leapfrog steps, L, and the scaling factor, ϵ. Gelman (2014) recommends to set L and ϵ such that $L\epsilon = 1$. A high-level procedure of the tuning of the HMC scheme for our problem is described below:

1. Firstly, we initialise the mass matrix of the momentum distribution, $\boldsymbol{M}$, to the inverse of the approximation of the posterior covariance of $(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta})$. For computational reasons we force $\boldsymbol{M}$ to be diagonal, i.e. consider only the inverse of the variance. The variance of $\mathbf{f}$ can for example be estimated from the Laplace approximation above. For the parameter $\boldsymbol{\theta}$, we take a trial and error approach. We learnt that in practice it is sufficient to specify the values with only the order of magnitude accuracy. We estimate the posterior variance of the coefficients $\boldsymbol{\beta}$ by fitting a Poisson regression which assumes i.i.d. noise. This was sufficient to get the algorithm started. We also initialise $L = 10$ and $\epsilon = 0.1$.

2. After the initialisation, we keep sampling and adjusting the values of L and ϵ, until the acceptance rate is about 65% as suggested by Gelman (2014). After each adjustment, the previous samples are thrown away to keep the chain ergodic.

3. At iteration $N_{\text{burn-in}}$ all the previous samples are thrown away. This value is set to an iteration number at which the chain is 'stabilised' according to the trace plots, but could be quantified by the Gelman-Rubin diagnostic defined in 2.2.2.

4. After $N_{\text{calibration}}$ samples, we compute the empirical covariance from the $N_{\text{calibration}} - N_{\text{burn-in}}$ samples. We set the diagonal of $\boldsymbol{M}$ to the inverse of the diagonal of the empirical covariance matrix.

5. Afterwards, we keep sampling and adjusting the values of L and ϵ, until the acceptance rate is about 65%. After each adjustment, the previous samples are thrown away to keep the chain ergodic.

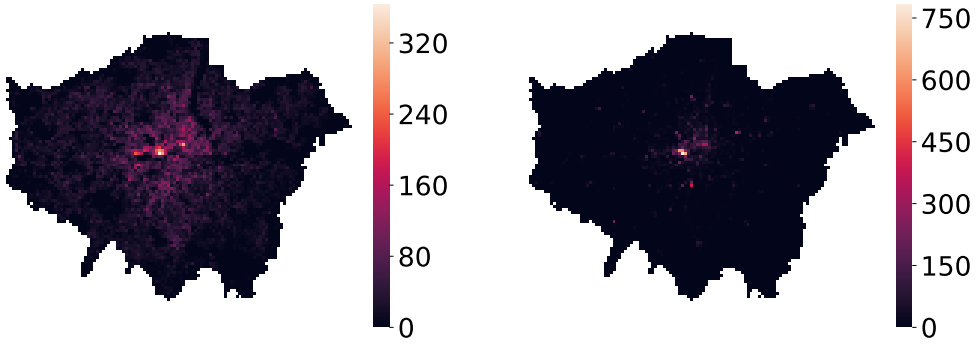
6. Once the chain has stabilised and the acceptance rate is around 65%, we start collecting the samples to be used for inferences later.

The dataset, the family of models, and the details of the two inference methods that we introduced in this chapter are applied in chapter 4 where we report the results.

4 Results and discussion

This chapter presents results obtained by using the inference procedures described in chapter 3. Throughout this chapter, there are three dimensions along which a specific model can be evaluated. Firstly, the **model specification** which is either *semi-parametric* or *non-parametric*, secondly, it is the **inference method** which is either *Laplace approximation* or *Markov Chain Monte Carlo sampling*, and lastly, it is the **crime type**. We start with the evaluation of non-parametric models in section 4.1, and then move to the semi-parametric models in section 4.2 where we also discuss the variable selection problem and its impact on the performance of a model.

Throughout this chapter, we narrow down our focus to two crime types: burglary, and theft from the person. Although both crime types involve stealing, they are very different. While for burglary the main target is a property, for theft from the person it is a person themselves. Burglary mainly occurs at locations with either residential or commercial properties. Theft from the person tends to cluster around the locations with high pedestrian traffic. Figure 4.1 shows the heat-map of both crime types using the data from 2016.



(A) Burglary (B) Theft from the person

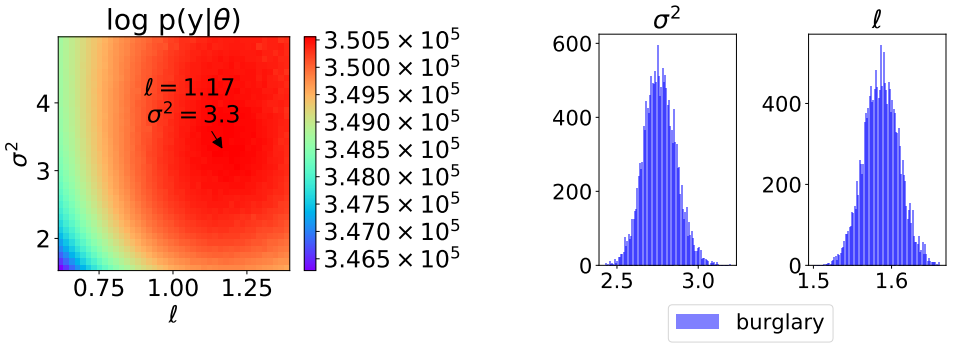
FIGURE 4.1: Heat-map of crime counts from 2016 plotted over the map of London.

4.1 Non-parametric model

In this section, we focus on non-parametric formulations of the LGCP as given in equation 3.5. Due to its simple formulation, we expect it to have limited explanatory power in terms of dynamics of crime. While the lengthscale hyper-parameter explains the degree of "similarity" of any two neighbouring locations it does not capture the underlying mechanisms of crime such as socio-economic predictors. For this reason, the focus of this section is on comparing Laplace approximation (LA) to MCMC sampling. The reported results are obtained from a model whose domain is discretised into a grid of cells, each of which corresponds to an area of 500m by 500m in real life.

Firstly, we compare the inferred hyper-parameters: lengthscale (ℓ) and variance (σ^2). Although the estimates are different in nature, the LA estimate is a point

estimate and MCMC estimate is a set of samples from the posterior distribution, one can compare the mode of the MCMC estimate to the LA estimate. Using the burglary dataset as an example, figure 4.2a shows the heatmap of marginal likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$, and its maximum value which corresponds to the 'optimal' set of parameters under the LA scheme. Figure 4.2b shows the posterior distribution of $p(\boldsymbol{\theta}|\mathbf{y})$ obtained using MCMC sampling.



(A) Marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$. The optimal set of parameters is found by an optimisation procedure.

(B) Histogram of samples from $p(\boldsymbol{\theta}|\mathbf{y})$ obtained using MCMC.

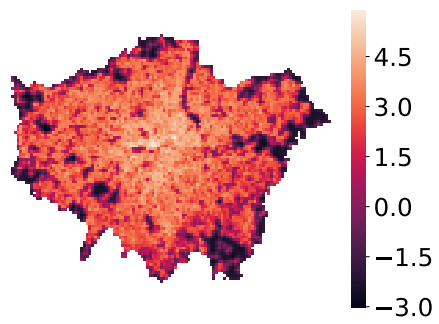
FIGURE 4.2: Hyper-parameter estimation using London burglary 2016 dataset - comparison of LA and MCMC estimates of the hyperparameters. Analogous plots for theft from the person dataset can be found in figure D.1.

It is interesting to see that the LA point estimate lies in the tails of the posterior distribution obtained using MCMC. The estimate from MCMC suggests that the latent function that drives the intensity of the Poisson process is smoother than the LA inference suggests. This is visualised in figure 4.3, where the mean and standard deviation of the posterior of the latent function is plotted over the map of London for both methods of inference. The 'smoothness' of the MCMC estimate is more visible in the standard deviation plots. Also note that the posterior distribution of the latent surface is more dispersed for location with low observed count of criminal activity. This can be seen by higher values of posterior standard deviations (for both LA and MCMC methods) in figure 4.3.

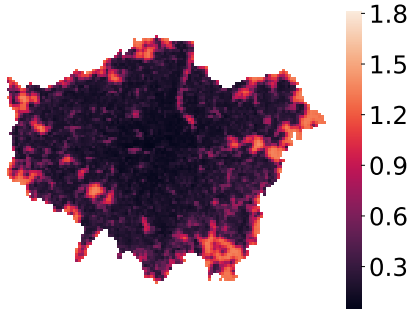
Although it intuitively makes sense to prefer a fully-Bayesian MCMC method due to its convergence properties (see section 2.2.2) and the ability to express full uncertainty about both the latent function and the hyper-parameters, we also provide supporting evidence by computing RMSE and WAIC criterion. Table 4.1 shows both of the metrics for London 2016 burglary and theft from the person datasets. As expected, the MCMC inference provides significantly better fit as can be seen mainly in the difference between the WAIC values for the two approaches.

The poor performance of the LA method compared to the MCMC method on the same model specification and on the same data could be attributed to two factors:

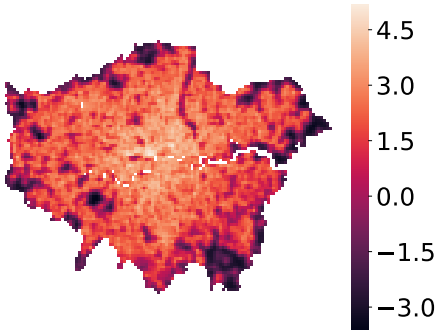
1. Marginal likelihood that is used to optimise the hyperparameters is a *stochastic estimate* as discussed in section 3.3. The noise in the estimate, which to a large extent depends on the condition number of the covariance matrix $\boldsymbol{K}$, could cause that the optimal value is not found.



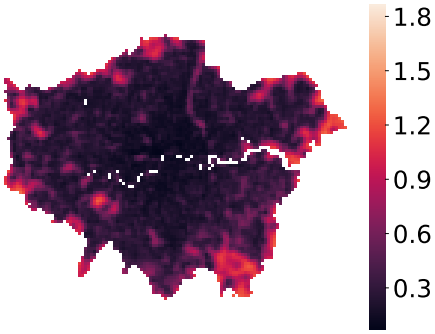
(A) Posterior mean (LA)



(B) Posterior standard deviation (LA)



(C) Posterior mean (MCMC)



(D) Posterior standard deviation (MCMC)

FIGURE 4.3: Inference of the posterior of the latent function, $p(\mathbf{f}|\mathbf{y})$. Dataset: London burglary 2016. For analagous plots on theft from the person see figure D.2.

Method	Burglary		Theft from the person	
	RMSE	WAIC	RMSE	WAIC
LA	4.8387	37039.242	2.6385	20942.765
MCMC	3.6302	32914.493	2.4407	17706.148

TABLE 4.1: Comparison of the fit of MCMC and LA.

2. The assumption that the posterior $p(\mathbf{f}|\mathbf{y})$ can be approximated by a Gaussian density might be too strong, especially for locations with little data.

Having demonstrated the superiority of the fully-Bayesian approach, we will not consider the LA method in subsequent evaluations. It is important to note that the MCMC approach is more computationally demanding and the inference generally takes longer than it does for the LA method.

4.2 Semi-parametric models

The non-parametric model acts as a smoothing kernel for the latent surface $\mathbf{f}$. Although a very succinct formulation, figure 4.3 shows that the posterior variance of the latent surface $\mathbf{f}$ is too large at some locations. Visual inspection of the population

density and the posterior variance, shown alongside each other in figure 4.4, suggests that the variance is higher at locations with low population density. This strongly suggests that the variance of the model predictions can be reduced by including a deterministic component into the intensity function as described in section 3.1.2. Motivated by this finding, we proceed with formulating a set of richer models that include socio-economic predictors such as population density.

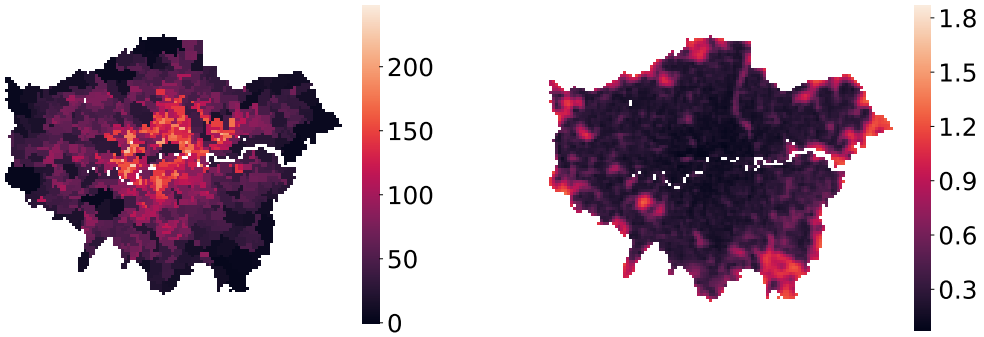


FIGURE 4.4: *(Left)* Population density expressed as the number of people per hectare plotted over the map of London. *(Right)* Posterior standard deviation of the latent surface $\mathbf{f}$ as estimated using MCMC for London burglary 2016 dataset.

Following the preliminary exploratory analysis of the relationship between the logarithm of crime counts and the individual covariates using QQ plots, we have decided to regress the log intensity of LGCP on the log of the covariates as it results in a more linear relationship for most of the covariates. Additionally, this choice results in more interpretable coefficients $\boldsymbol{\beta}$. The coefficients can then be interpreted as elasticities. By applying the log to the covariates $\boldsymbol{h}(\boldsymbol{x})$ in the original intensity function of the LGPCP process (equation 3.3) we obtain

$$\lambda(\boldsymbol{x}) = \exp\left((\log \boldsymbol{h}(\boldsymbol{x}))^\top \boldsymbol{\beta} + f(\boldsymbol{x})\right).$$

Then the elasticity of this intensity function with respect to a covariate $\boldsymbol{h}_i(\boldsymbol{x})$ is given as

$$\begin{aligned}
\text{elasticity} &= \frac{\dfrac{\partial \lambda(\boldsymbol{x})}{\partial \boldsymbol{h}_i(\boldsymbol{x})}}{\dfrac{\lambda(\boldsymbol{x})}{\boldsymbol{h}_i(\boldsymbol{x})}} \\
&= \frac{\lambda(\boldsymbol{x}) \dfrac{1}{\boldsymbol{h}_i(\boldsymbol{x})} \boldsymbol{\beta}_i}{\dfrac{\lambda(\boldsymbol{x})}{\boldsymbol{h}_i(\boldsymbol{x})}} \\
&= \boldsymbol{\beta}_i.
\end{aligned}$$

4.2.1 Model 1

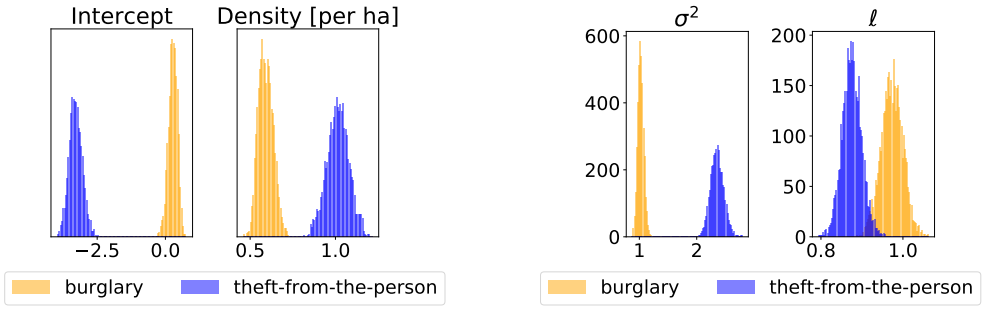
Drawing on the discussion above, the baseline model for this section includes only population density as the covariate. In other words,

$$\boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\beta} = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2 \log \texttt{population\_density}(\boldsymbol{x}), \tag{4.1}$$

where $\boldsymbol{\beta}_1$ corresponds to the intercept. We fit the model to two datasets: London 2016 burglary crimes and London 2016 theft from the person crimes. The domain of the model is a grid with a cell size corresponding to an area of 750m by 750m. Figure 4.5 shows the posterior of the inferred quantities: the coefficients and the hyperparameters. Firstly, we can see that thanks to smaller spatial variations of the burglary counts, the histograms for the posterior quantities for the burglary model are more sharply peaked.

The intercept for burglary just above 0 and the population density elasticity of burglary crime count of around 0.6 corresponds to a steady state of crime counts across the domain with expected increases as a result of higher population density. This is in contrast with theft from the person model that identifies lower crime counts across the domain (due to the negative intercept), but sharply increased crime counts for areas with high population density. This is what one would expect as theft from the person is a crime whose subject is a person rather than a property. This argument is supported by visual inspection of crime counts of theft form the person (see figure 4.1b) where the intensity of hotspots is more pronounced than it is for burglary (see figure 4.1a).

The unexplained residuals that correspond to the latent surface $\mathbf{f}$, also provide evidence that spatial variation of the crime count intensity is higher for theft from the person that it is for burglary, suggested by the distribution of the lengthscale parameter, ℓ, in figure 4.5b.

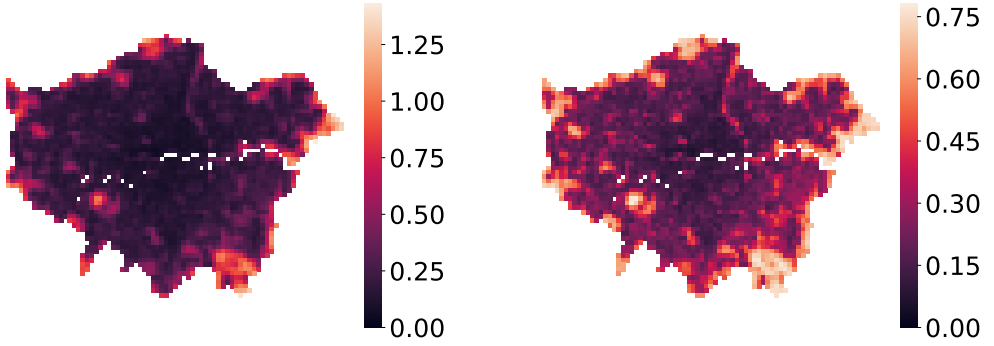


(A) $p(\boldsymbol{\beta}|\mathbf{y})$ (B) $p(\boldsymbol{\theta}|\mathbf{y})$

FIGURE 4.5: Posterior distributions of the coefficients of the covariates *(Left)*, and the hyper-parameters of the GP prior of the latent function *(Right)*.

To show that including population density as a covariate improves predictive performance of the model, we compute RMSE, WAIC which have both improved for both burglary and theft from the person. The values are reported in table 4.2.

Another way of looking at predictive performance is to look at the variance of predictions. To this end, we compute standard deviation of the posterior log intensity of the LGCP model, $\sqrt{\text{Var}_{\mathbf{f},\boldsymbol{\beta},\boldsymbol{\theta}|\mathbf{y}}\Lambda(\boldsymbol{x})}$. Figure 4.6 shows this quantity for both the non-parametric model and the semi-parametric model that includes population density as



(A) Non-parametric LGCP model fitted using MCMC.

(B) Semi-parametric LGCP model with population density as a covariate. Fitted using MCMC.

FIGURE 4.6: Standard deviation of the posterior log intensity of LGCP. Dataset: London burglary 2016.

an explanatory variable. It shows that including the information about population density shrinks the variance of model predictions.

4.2.2 Model 2

Building upon the population density model above, we add more socio-economic covariates. In the choice of predictors, we are mostly inspired by *social disorganisation theory* which suggests that offenders choose neighbourhoods with low social cohesion. Additionally, out of all the theories we outlined in section 1.1, social disorganisation theory focuses on indicators that we have access to via census data. On top of the covariates in the previous model, we believe that the following covariates will provide explanatory power for varying crime rates: median age, median household income, unemployment rate, median house price, percentage of immigrants, percentage of people with level 4 education and above. In other words, the deterministic component $\boldsymbol{h}(\boldsymbol{x})$ of equation 3.1 is equal to

$$\begin{aligned}\boldsymbol{h}(\boldsymbol{x})^\top \boldsymbol{\beta} = \boldsymbol{\beta}_1 &+ \boldsymbol{\beta}_2 \log \texttt{population\_density}(\boldsymbol{x}) \\ &+ \boldsymbol{\beta}_3 \log \texttt{median\_age}(\boldsymbol{x}) \\ &+ \boldsymbol{\beta}_4 \log \texttt{median\_household\_income}(\boldsymbol{x}) \\ &+ \boldsymbol{\beta}_5 \log \texttt{median\_house\_price}(\boldsymbol{x}) \\ &+ \boldsymbol{\beta}_6 \log \texttt{percentage\_of\_unemployed}(\boldsymbol{x}) \\ &+ \boldsymbol{\beta}_7 \log \texttt{percentage\_of\_immigrants}(\boldsymbol{x}) \\ &+ \boldsymbol{\beta}_8 \log \texttt{percentage\_of\_L4\_education}(\boldsymbol{x})\end{aligned} \tag{4.2}$$

The model in equation 4.2 is fitted using the MCMC scheme described in section 3.4. For traceplots of the MCMC inference see figure D.3. Figure 4.7 reports the posterior distributions for the coefficients of the covariates, $p(\boldsymbol{\beta}|\mathbf{y})$.

The *intercept*, whose value for both crime types is centred around -1, indicates relatively low baseline rate of crime counts across the map.

Population density can be interpreted in the same way as in the simple model. The more populated areas attract more crime, and this is more so for theft from the person.

Interestingly, the *median* age statistic has a strong decreasing effect on the crime count rate. This could be related to the fact that people tend to offend at younger ages. This would suggest that they offend in their own neighbourhood. The youth offending is a strong theme within strain theory and subcultures theory, where lack of ability to achieve cultural goals and differential opportunity leads youth population to resort to illegal ways of achieving those goals. The decreasing effect is stronger for theft from the person.

Positive values of $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$ suggests that more affluent neighbourhoods are more likely to become a target of offending than the poorer ones. Rather surprisingly, we would expect *median house price* to be a stronger attractor for burglary. In fact, it has stronger effect on theft from the person than on burglary.

Percentage of *unemployed* seems to as expected to have increasing effect on the crime count rates. This finding is consistent with social disorganisation theory, as neighbourhoods with higher unemployment rates are expected to be less socially cohesive.

Social disorganisation theory suggests that offenders often prefer areas with high level of immigration. Our model demonstrates that increased proportion of immigrants in the area has the opposite effect, for both burglary and theft from the person. Without identifying the reason for this, we conjecture that a large proportion of immigrants in London are highly-skilled workers who do not tend to engage in criminal activity. This observation invites a deeper investigation, which involves interactions between different covariates.

Proportion of *highly-educated* population in the area is positively correlated with the crime count rates. Given that this effect is stronger for theft from the person which peaks in the city centre and the areas of commercial activity, this could suggest that areas that are closer to the city centre have higher proportions of educated people. However, this hypothesis needs to be further verified.

The 'residuals' that are unexplained by the covariates seems to exhibit the same characteristics as for the simple model - spatial variation of the latent field $\mathbf{f}$ is higher for theft from the person than it is for burglary (see figure 4.8).

From this discussion it is clear that the task of choosing the right predictors is a challenging problem. Firstly, our assumption that the log intensity of the crime counts is a linear combination of the logged predictors might be flawed. Secondly, our model does not cater for interactions between the predictors, which we believe is necessary as demonstrated by the example of immigrants and the level of education above. In our modelling approach we almost certainly omitted predictors that are significant. And lastly, due to the possibility of almost perfect linear relationship between two predictors, often referred to as collinearity in the linear models literature, the model might infer coefficients with large variance. As a result, there is a set of heuristics that have been proposed in the literature to deal with some of these issues.

4.2.3 Model selection, evaluation

It is not surprising that adding more covariates improves the fit of the model. A good fit does not imply that the model identified the correct predictors. The model could simply overfit the training data and does not generalise well on out-of-sample data. A common approach to testing generality of the model by assessing out-of-sample predictive accuracy is cross-validation. However, in some cases this approach is not feasible as it requires many model fits which is an issue for models estimated using MCMC.

FIGURE 4.7: Posterior of the coefficient values, $p(\boldsymbol{\beta}|\mathbf{y})$.

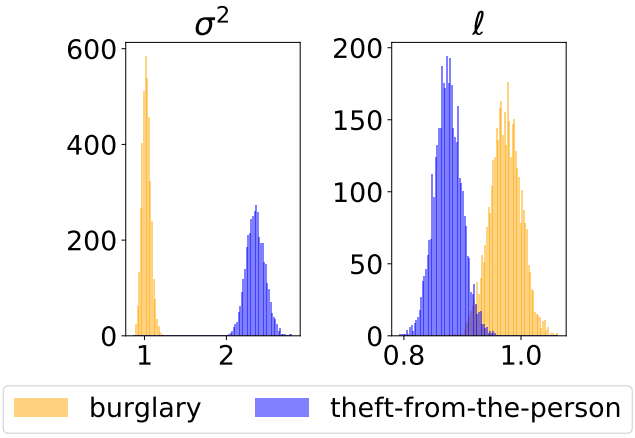


FIGURE 4.8: Posterior of the hyper-parameters, $p(\boldsymbol{\theta}|\mathbf{y})$.

Method	Burglary		Theft from the person	
	RMSE	WAIC	RMSE	WAIC
Non-parametric	5.312	17626.425	3.581	10470.633
Model 1	4.979	16906.597	3.564	10374.831
Model 2	4.963	16864.143	3.553	10296.501

TABLE 4.2: Evaluation of the models fit using MCMC.

A common technique that assesses within-sample predictive accuracy but discounts the complexity of the model is WAIC, which is described in section 3.2.1. In contrast with cross-validation, this approach only requires one fit per model. Table 4.2 shows both the RMSE and the WAIC information criterion for all the models that we fitted using MCMC: non-parametric model, population density-based model (Model 1), and a richer model that includes other socio-economic covariates (Model 2). As expected, adding population density into the model improved the fit which is measured by RMSE. Predictive performance for which WAIC is an estimate suggests a 4.1% improvement for burglary, and only a 1% improvement for theft from the person. The addition of other covariates has resulted in an improvement to both RMSE and WAIC metrics, but the improvements were only marginal: 0.2% for burglary and 0.7% for theft from the person.

Although these results seem disappointing, the two semi-parametric models we presented were just an example of how this framework can be used to test different models and their predictive power as a measure of how well the model represents the true process that generates crime occurrences in space. More expert knowledge on criminology can be used to propose models that can be fitted and tested using this framework.

Another approach to variable selection in regression problems is *lasso* (Friedman, Hastie, and Tibshirani, 2001). Without going into details, the method can shrink the coefficients of predictors towards zero by imposing penalty on their size. An equivalent of this approach in the Bayesian setting is placing strong priors centred around zero, such as Laplace prior, on the coefficients (Gelman, 2014, ch. 14). We suggest this as an area of future work.

5 Conclusion and further work

5.1 Evaluation

Modelling social phenomena such as crime is a challenging problem. Our treatment of crime as an environmentally-driven process was motivated by the criminology theories that stress external social factors as the main source of criminal behaviour. Such processes naturally fit with the Cox process models. Our formulation of the Cox process includes a deterministic, linear component and a latent noise function with a Gaussian Process prior. While linear component enabled us to add explanatory power to the model, the latent noise function interpolates the unexplained residuals. By applying Bayesian framework, we compared Laplace approximation (LA) and Markov Chain Monte Carlo sampling (MCMC) methods to overcome the problem of intractability of the posterior distribution of the unknown quantities of interest. By assuming grid structure of the domain we were able to utilize Kronecker methods in making our inference method scalable.

Applying the methodology above to London crime data for burglary and theft from the person from 2016, we have shown that fully-Bayesian inference using MCMC of the posterior distribution for the quantities of interest resulted in better predictive performance than approximate Bayesian inference with fixed hyper-parameters using Laplace approximation. The predictive performance, as measured by Watanabe-Akaike information criterion (WAIC), has improved by 11% for the model of burglary and 15% for the model of theft from the person.

Subsequently, by adding socio-economic covariates to the LGCP model and using MCMC inference, we have managed to reduce the variance of model's predictions. While the increase in predictive performance of the two richer models we proposed was not as significant as for MCMC vs LA comparison, this framework allows for a systematic testing of models using WAIC criterion, which combined with expert knowledge can be used to build more sophisticated models with better explanatory power and predictive performance. We leave the question of model selection open and discuss other potential improvements in section 5.3.

5.2 Contribution

Putting this work into the context of current literature, we have improved upon the semi-parametric approach proposed by Marchant et al. (2018). Firstly, our approach uses the Poisson observation model as opposed to the Gaussian likelihood. We believe that the Poisson likelihood, although resulting in analytically intractable inference methods, is a better representation of reality, especially at locations with very low crime rate. By borrowing the idea of Kronecker methods from Flaxman et al. (2015) and assuming the grid structure of our domain, our methods scale to much finer spatial resolution which allows more accurate inferences to be made.

5.3 Limitations and further work

As we have shown in the results (chapter 4), our method in its current form has a number of limitations. Although we attempted to appeal to criminology to propose plausible models, the richer models did not bring improvements we expected. By specifying the same parametric form for both of the crime types we considered, and by not including any interactions between covariates, we ignored particular characteristics that each crime type exhibits. In future, with the help of expert knowledge, models which are more representative of reality could be specified. These suggestions together with employing shrinkage methods could be used to specify a set of explanatory models whose predictive performance can be tested by the already implemented framework that utilises WAIC.

This work only considers the spatial dimension of criminal activity and ignores time. As indicated by papers that focus on crime forecasting (Taddy, 2010; Flaxman et al., 2015), the temporal dimension could offer valuable insights into understanding the drivers of crime.

Lastly, due to limited scope, this work does not benchmark its performance against the state of the art frameworks such as INLA (Rue et al., 2017). As the first suggestion, we suggest applying the INLA framework to our dataset compare the two approaches.

A Linear algebra and probability results

A.1 Matrix identities

Matrix inversion lemma, also known as Woodbury identity states that

$$\left(\boldsymbol{Z}+\boldsymbol{U}\boldsymbol{W}\boldsymbol{V}^{\top}\right)^{-1}=\boldsymbol{Z}^{-1}-\boldsymbol{Z}^{-1}\boldsymbol{U}\left(\boldsymbol{W}^{-1}+\boldsymbol{V}^{\top}\boldsymbol{Z}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}^{\top}\boldsymbol{Z}^{-1}, \tag{A.1}$$

where $\boldsymbol{Z}$ is $n \times n$, $\boldsymbol{W}$ is $m \times m$, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are of size $n \times m$ (Rasmussen and Williams, 2006).

A.2 Gaussian distribution

A random vector $\boldsymbol{x}$ of size D that follows multivariate Gaussian distribution has a joint probability density given by

$$p(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)=(2\pi)^{-D/2}\det(\Sigma)^{-1/2}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})), \tag{A.2}$$

where $\boldsymbol{\mu}$ is the mean vector and Σ is the $D \times D$ covariance matrix.

Let $\mathbf{x}$ and $\mathbf{y}$ be jointly Gaussian random vectors

$$\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}\mu_{\mathbf{x}}\\ \mu_{\mathbf{y}}\end{bmatrix},\begin{bmatrix}A & C\\ C^{\top} & B\end{bmatrix}\right),$$

then the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ is

$$\mathbf{x}|\mathbf{y}\sim\mathcal{N}\left(\mu_{\mathbf{x}}+CB^{-1}(\mathbf{y}-\mu_{\mathbf{y}}),A-CB^{-1}C^{\top}\right). \tag{A.3}$$

See Rasmussen and Williams (2006, sec A.2) for more details.

B Dataset

B.1 Crime types

Crime type	Description
Anti-social behaviour	Includes personal, environmental and nuisance anti-social behaviour.
Bicycle theft	Includes the taking without consent or theft of a pedal cycle.
Burglary	Includes offences where a person enters a house or other building with the intention of stealing.
Criminal damage and arson	Includes damage to buildings and vehicles and deliberate damage by fire.
Drugs	Includes offences related to possession, supply and production.
Other crime	Includes forgery, perjury and other miscellaneous crime.
Other theft	Includes theft by an employee, blackmail and making off without payment.
Possession of weapons	Includes possession of a weapon, such as a firearm or knife.
Public disorder and weapons	Includes offences which cause fear, alarm, distress or a possession of a weapon such as a firearm.
Public order	Includes offences which cause fear, alarm or distress.
Robbery	Includes offences where a person uses force or threat of force to steal.
Shoplifting	Includes theft from shops or stalls.
Theft from the person	Includes crimes that involve theft directly from the victim (including handbag, wallet, cash, mobile phones) but without the use or threat of physical force.
Vehicle crime	Includes theft from or of a vehicle or interference with a vehicle.
Violence and sexual offences	Includes offences against the person such as common assaults, Grievous Bodily Harm and sexual offences.

Table B.1: Crime type categories defined by UK Police

C Derivations

C.1 Laplace approximation derivations

C.1.1 Newton step derivations

Full derivation of equation 3.21:

$$
\begin{aligned}
\mathbf{f}^{\text{new}} &= \mathbf{f} - (\nabla\nabla_{\mathbf{f}}\Psi(\mathbf{f}))^{-1}\nabla_{\mathbf{f}}\Psi(\mathbf{f}) \\
&= \mathbf{f} + \left(\boldsymbol{K}^{-1} + \boldsymbol{W}\right)^{-1}\left[\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{K}^{-1}\mathbf{f}\right] \\
&= \mathbf{f} + [\boldsymbol{K} - \boldsymbol{K}\boldsymbol{Q}\boldsymbol{K}]\left[\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{K}^{-1}\mathbf{f}\right] \text{ (using equation A.1)} \\
&= \boldsymbol{K}\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{K}\boldsymbol{Q}\boldsymbol{K}\nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{K}\boldsymbol{Q}\mathbf{f} \\
&= \boldsymbol{K}\left(\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{Q}\boldsymbol{K}\nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{Q}\boldsymbol{W}^{-1}\boldsymbol{W}\mathbf{f}\right) \\
&= \boldsymbol{K}\left(\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{Q}\boldsymbol{K}\nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{Q}(\boldsymbol{K} + \boldsymbol{W}^{-1} - \boldsymbol{K})\boldsymbol{W}\mathbf{f}\right) \\
&= \boldsymbol{K}\left(\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{Q}\boldsymbol{K}\nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{Q}(\boldsymbol{K} + \boldsymbol{W}^{-1} - \boldsymbol{K})\boldsymbol{W}\mathbf{f}\right) \\
&= \boldsymbol{K}\left(\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{Q}\boldsymbol{K}\nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{Q}(\boldsymbol{Q}^{-1} - \boldsymbol{K})\boldsymbol{W}\mathbf{f}\right) \\
&= \boldsymbol{K}\left(\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{Q}\boldsymbol{K}\nabla \log p(\mathbf{y}|\mathbf{f}) + (\boldsymbol{I} - \boldsymbol{Q}\boldsymbol{K})\boldsymbol{W}\mathbf{f}\right) \\
&= \boldsymbol{K}\left(\nabla \log p(\mathbf{y}|\mathbf{f}) - \boldsymbol{Q}\boldsymbol{K}\nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{W}(\mathbf{f} - \boldsymbol{\mu}) - \boldsymbol{Q}\boldsymbol{K}\boldsymbol{W}\mathbf{f}\right) \\
&= \boldsymbol{K}\left(\nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{W}\mathbf{f} - \boldsymbol{Q}\boldsymbol{K}\left(\nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{W}\mathbf{f}\right)\right) \\
&= \boldsymbol{K}\left(\boldsymbol{b} - \boldsymbol{Q}\boldsymbol{K}\boldsymbol{b}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{b} &= \nabla \log p(\mathbf{y}|\mathbf{f}) + \boldsymbol{W}\mathbf{f}, \\
\boldsymbol{B} &= \boldsymbol{I} + \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{K}\boldsymbol{W}^{\frac{1}{2}}, \\
\boldsymbol{Q} &= (\boldsymbol{K} + \boldsymbol{W}^{-1})^{-1} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{B}^{-1}\boldsymbol{W}^{\frac{1}{2}}.
\end{aligned}
$$

C.1.2 Marginal likelihood computation

Full derivation of equation 3.24:

$$
\begin{aligned}
\log p(\mathbf{y}|\boldsymbol{\theta}) &\overset{\text{const}}{=} \log p(\hat{\mathbf{f}}|\mathbf{y}) + \frac{1}{2}\log\left|(2\pi(\boldsymbol{W} + \boldsymbol{K}^{-1})^{-1}\right| \\
&= \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\log|2\pi\boldsymbol{K}| - \frac{1}{2}\hat{\mathbf{f}}^{\top}\boldsymbol{K}^{-1}\hat{\mathbf{f}} + \frac{1}{2}\log\left|2\pi(\boldsymbol{W} + \boldsymbol{K}^{-1})^{-1}\right| \\
&= \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\log|2\pi\boldsymbol{K}| - \frac{1}{2}\hat{\mathbf{f}}^{\top}\boldsymbol{K}^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log\left|2\pi\boldsymbol{K}^{-1}\right|\,|\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}| \\
&= \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\hat{\mathbf{f}}^{\top}\boldsymbol{K}^{-1}\hat{\mathbf{f}} - \frac{1}{2}\log|\boldsymbol{I} + \boldsymbol{K}\boldsymbol{W}|
\end{aligned}
\tag{C.1}
$$

C.2 Markov-Chain Monte Carlo derivations

Expression for the log-posterior distribution. Taking the posterior from equation 3.27 and applying log gives:

$$\begin{aligned}
\log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) &= \log p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) + \log p(\mathbf{f}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \log p(\boldsymbol{\beta}) + \text{const}_1 \\
&= \left[\sum_i \log p(\mathrm{y}_i|\mathrm{f}_i, \boldsymbol{\beta})\right] + \left[-\frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\mathbf{f}^\top \boldsymbol{K}^{-1}\mathbf{f}\right] \\
&\quad + \left[-\frac{1}{2}\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}\right] + \left[-\frac{1}{2}\boldsymbol{\beta}^\top \Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right] + \text{const}_2 \\
&= \left[\sum_i \log \frac{\exp(\boldsymbol{Z}_{i\cdot}^\top \boldsymbol{\beta} + \mathrm{f}_i)^{\mathrm{y}_i} e^{-\exp(\boldsymbol{Z}_{i\cdot}^\top \boldsymbol{\beta} + \mathrm{f}_i)}}{\mathrm{y}_i!}\right] + \left[-\frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\mathbf{f}^\top \boldsymbol{K}^{-1}\mathbf{f}\right] \\
&\quad + \left[-\frac{1}{2}\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}\right] + \left[-\frac{1}{2}\boldsymbol{\beta}^\top \Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right] + \text{const}_2 \\
&= \left[\sum_i \mathrm{y}_i(\boldsymbol{Z}_{i\cdot}^\top \boldsymbol{\beta} + \mathrm{f}_i) - \exp(\boldsymbol{Z}_{i\cdot}^\top \boldsymbol{\beta} + \mathrm{f}_i)\right] + \left[-\frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\mathbf{f}^\top \boldsymbol{K}^{-1}\mathbf{f}\right] \\
&\quad + \left[-\frac{1}{2}\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}\right] + \left[-\frac{1}{2}\boldsymbol{\beta}^\top \Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right] + \text{const}_3 \\
&= \left[\mathbf{y}^\top \boldsymbol{Z}\boldsymbol{\beta} + \mathbf{y}^\top \mathbf{f} - \exp(\boldsymbol{Z}\boldsymbol{\beta} + \mathbf{f})\right] + \left[-\frac{1}{2}\log|\boldsymbol{K}| - \frac{1}{2}\mathbf{f}^\top \boldsymbol{K}^{-1}\mathbf{f}\right] \\
&\quad + \left[-\frac{1}{2}\boldsymbol{\theta}^\top \Sigma_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta}\right] + \left[-\frac{1}{2}\boldsymbol{\beta}^\top \Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right] + \text{const}_3
\end{aligned}$$

C.2.1 Gradients

Taking the derivative of the log-posterior with respect to the quantities of interest gives:

$$\begin{aligned}
\nabla_{\mathbf{f}} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) &= \nabla_{\mathbf{f}} \log p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) + \nabla_{\mathbf{f}} \log p(\mathbf{f}|\boldsymbol{\theta}) \\
&= [\mathbf{y} - \exp(\boldsymbol{Z}\boldsymbol{\beta} + \mathbf{f})] + \left[-\boldsymbol{K}^{-1}\mathbf{f}\right]
\end{aligned} \tag{C.2}$$

$$\begin{aligned}
\nabla_{\boldsymbol{\beta}} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) &= \nabla_{\boldsymbol{\beta}} \log p(\mathbf{y}|\mathbf{f}, \boldsymbol{\beta}) + \nabla_{\boldsymbol{\beta}} \log p(\boldsymbol{\beta}) \\
&= \left[\boldsymbol{Z}^\top \mathbf{y} - \boldsymbol{Z}^\top \exp(\boldsymbol{Z}\boldsymbol{\beta} + \mathbf{f})\right] + \left[-\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\right]
\end{aligned} \tag{C.3}$$

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_i} \log p(\mathbf{f}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}) &= \nabla_{\boldsymbol{\theta}_i} \log p(\mathbf{f}|\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}_i} \log p(\boldsymbol{\theta}) \\
&= \left[\frac{1}{2}\mathbf{f}^\top \boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\boldsymbol{K}^{-1}\mathbf{f} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right)\right] + \left[-(\Sigma_{\boldsymbol{\theta}}^{-1}\boldsymbol{\theta})_i\right],
\end{aligned} \tag{C.4}$$

where

$$\operatorname{tr}\left(\boldsymbol{K}^{-1}\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i}\right) = \operatorname{tr}\left(\left(\bigotimes_{d=1}^{D}\boldsymbol{K}_d^{-1}\right)\left(\sum_{d=1}^{D}\frac{\partial \boldsymbol{K}_d}{\partial \boldsymbol{\theta}_i}\otimes\left(\bigotimes_{j\neq d}\boldsymbol{K}_j\right)\right)\right)$$
$$= \sum_{d=1}^{D}\operatorname{tr}\left(\boldsymbol{K}_d^{-1}\frac{\partial \boldsymbol{K}_d}{\partial \boldsymbol{\theta}_i}\right)\prod_{j\neq d}\operatorname{tr}\left(\boldsymbol{K}_j^{-1}\boldsymbol{K}_j\right) \tag{C.5}$$

using

$$\frac{\partial \boldsymbol{K}}{\partial \boldsymbol{\theta}_i} = \sum_{d=1}^{D}\frac{\partial \boldsymbol{K}_d}{\partial \boldsymbol{\theta}_i}\otimes\left(\bigotimes_{j\neq d}\boldsymbol{K}_j\right). \tag{C.6}$$

Matérn gradients

We start with the Matérn covariance function parameterised by lengthscale ℓ:

$$k_{\nu=5/2}(r) = \left(1+\frac{\sqrt{5}r}{\ell}+\frac{5r^2}{3\ell^2}\right)\exp\left(-\frac{\sqrt{5}r}{\ell}\right)$$

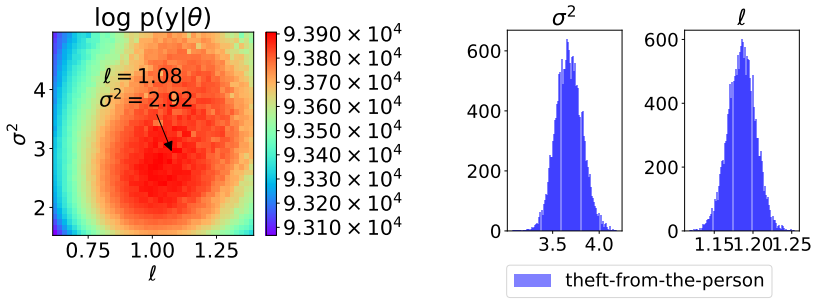
However, in order to avoid positivity constraint when doing inference, we let the $\ell = \exp(\phi_\ell)$ and the covariance function becomes

$$k_{\nu=5/2}(r) = \left(1+\frac{\sqrt{5}r}{\exp(\phi_\ell)}+\frac{5r^2}{3\exp(2\phi_\ell)}\right)\exp\left(-\frac{\sqrt{5}r}{\exp(\phi_\ell)}\right).$$

For the purposes of HMC, the derivative with respect to ϕ_ℓ is equal to:

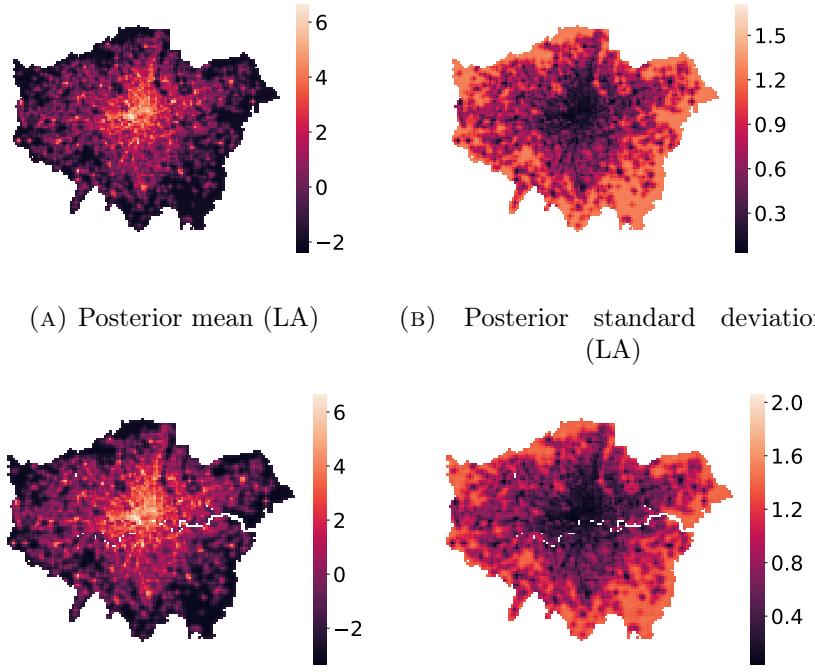
$$\begin{aligned}
\frac{\partial}{\partial\phi_\ell}k_{\nu=5/2}(r) &= \frac{\partial}{\partial\phi_\ell}\left(1+\frac{\sqrt{5}r}{\exp(\phi_\ell)}+\frac{5r^2}{3\exp(2\phi_\ell)}\right)\exp\left(-\frac{\sqrt{5}r}{\exp(\phi_\ell)}\right)\\
&\quad+\left(1+\frac{\sqrt{5}r}{\exp(\phi_\ell)}+\frac{5r^2}{3\exp(2\phi_\ell)}\right)\frac{\partial}{\partial\phi_\ell}\exp\left(-\frac{\sqrt{5}r}{\exp(\phi_\ell)}\right)\\
&= \left(-\frac{\sqrt{5}r}{\exp(\phi_\ell)}-\frac{10r^2}{3\exp(2\phi_\ell)}\right)\exp\left(-\frac{\sqrt{5}r}{\exp(\phi_\ell)}\right)\\
&\quad+\left(1+\frac{\sqrt{5}r}{\exp(\phi_\ell)}+\frac{5r^2}{3\exp(2\phi_\ell)}\right)\exp\left(-\frac{\sqrt{5}r}{\exp(\phi_\ell)}\right)\frac{\sqrt{5}r}{\exp(\phi_\ell)}\\
&= \exp\left(-\frac{\sqrt{5}r}{\exp(\phi_\ell)}\right)\left(-\frac{\sqrt{5}r}{\exp(\phi_\ell)}-\frac{10r^2}{3\exp(2\phi_\ell)}+\frac{\sqrt{5}r}{\exp(\phi_\ell)}+\frac{5r^2}{\exp(2\phi_\ell)}+\frac{5\sqrt{5}r^3}{3\exp(3\phi_\ell)}\right)\\
&= \exp\left(-\frac{\sqrt{5}r}{\exp(\phi_\ell)}\right)\left(\frac{5r^2}{3\exp(2\phi_\ell)}+\frac{5\sqrt{5}r^3}{3\exp(3\phi_\ell)}\right)\\
&= \exp\left(-\frac{\sqrt{5}r}{\ell}\right)\left(\frac{5r^2}{3\ell^2}+\frac{5\sqrt{5}r^3}{3\ell^3}\right).
\end{aligned}$$

D Additional plots



(A) Marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$. (B) Histogram of samples from $p(\boldsymbol{\theta}|\mathbf{y})$.

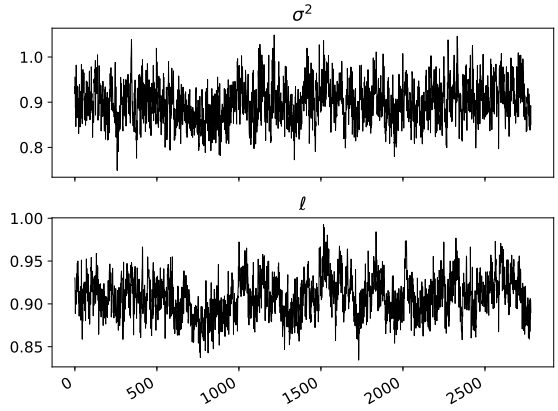
FIGURE D.1: Hyper-parameter estimation, *(Left)* LA, and *(Right)* MCMC. Dataset: London theft from the person 2016.



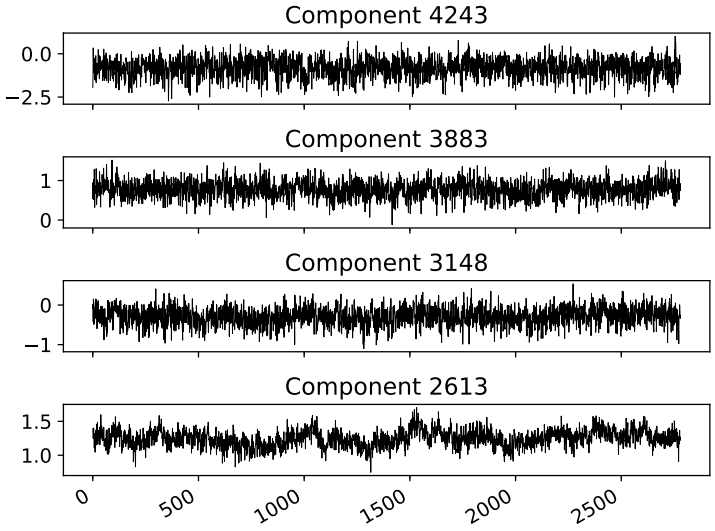
(A) Posterior mean (LA) (B) Posterior standard deviation (LA)

(C) Posterior mean (MCMC) (D) Posterior standard deviation (MCMC)

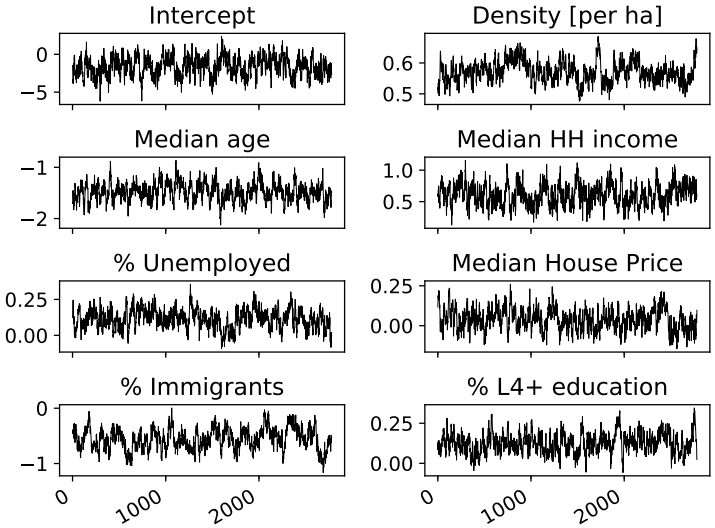
FIGURE D.2: Inference of the posterior of the latent function, $p(\mathbf{f}|\mathbf{y})$,in a non-parametric setting for theft from the person 2016.



(A) Traceplot for the samples from $p(\boldsymbol{\theta}|\mathbf{y})$.



(B) Traceplot for the samples from $p(\mathbf{f}|\mathbf{y})$.



(C) Traceplot for the samples from $p(\boldsymbol{\beta}|\mathbf{y})$.

FIGURE D.3: Traceplots for MCMC inference of a fully-Bayesian semi-parametric model of burglary

Bibliography

Abramowitz, Milton and Irene A. Stegun, eds. (2013). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. eng. 9. Dover print. Dover books on mathematics. OCLC: 935935300. New York, NY: Dover Publ. ISBN: 978-0-486-61272-0.

Adler, Robert J. and Jonathan E. Taylor (2007). *Random fields and geometry*. Springer monographs in mathematics 115. New York: Springer. ISBN: 978-0-387-48112-8 978-0-387-48116-6.

Agnew, Robert (2008). "Strain Theory". In: *Encyclopedia of Social Problems*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. ISBN: 978-1-4129-4165-5 978-1-4129-6393-0. DOI: 10.4135/9781412963930.n550.

Beccalossi, Chiara (2010). "Lombroso, Cesare: The Criminal Man". en. In: *Encyclopedia of Criminological Theory*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc., pp. 561–566. ISBN: 978-1-4129-5918-6 978-1-4129-5919-3. DOI: 10.4135/9781412959193.n155.

Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. Information science and statistics. New York: Springer. ISBN: 978-0-387-31073-2.

Blackman, Shane (2014). "Subculture Theory: An Historical and Contemporary Assessment of the Concept for Understanding Deviance". en. In: *Deviant Behavior* 35.6, pp. 496–512. ISSN: 0163-9625, 1521-0456. DOI: 10.1080/01639625.2013.859049.

Carpenter, Andrew N. (2010). "Beccaria, Cesare: Classical School". en. In: *Encyclopedia of Criminological Theory*. 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. ISBN: 978-1-4129-5918-6 978-1-4129-5919-3. DOI: 10.4135/9781412959193.n19.

Cox, D. R. (1955). "Some Statistical Methods Connected with Series of Events". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 17.2, pp. 129–164.

Deadman, Derek (2003). "Forecasting residential burglary". en. In: *International Journal of Forecasting* 19.4, pp. 567–578. ISSN: 01692070. DOI: 10.1016/S0169-2070(03)00091-8.

Diggle, Peter J. et al. (2013). "Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm". en. In: *Statistical Science* 28.4, pp. 542–563. ISSN: 0883-4237. DOI: 10.1214/13-STS441.

Dong, Kun et al. (2017). "Scalable Log Determinants for Gaussian Process Kernel Learning". In: *Advances in Neural Information Processing Systems*.

Fiedler, Miroslav (1971). "Bounds for the Determinant of the Sum of Hermitian Matrices". In: *Proceedings of the American Mathematical Society* 30.1, p. 27. ISSN: 00029939. DOI: 10.2307/2038212.

Filippone, Maurizio and Mark Girolami (2014). "Pseudo-Marginal Bayesian Inference for Gaussian Processes". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.11, pp. 2214–2226. ISSN: 0162-8828, 2160-9292. DOI: 10.1109/TPAMI.2014.2316530.

Flaxman, Seth et al. (2015). "Fast Kronecker Inference in Gaussian Processes with non-Gaussian Likelihoods". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning*. Vol. 37. ICML'15. Lille, France: JMLR.org, pp. 607–616.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

Gardner, Jacob R et al. (2018). "Product Kernel Interpolation for Scalable Gaussian Processes". In: *AISTATS*.

Gelman, Andrew (2014). *Bayesian data analysis*. Third edition. Chapman & Hall/CRC texts in statistical science. Boca Raton: CRC Press. ISBN: 978-1-4398-4095-5.

Gelman, Andrew and Donald B. Rubin (1992). "Inference from Iterative Simulation Using Multiple Sequences". en. In: *Statistical Science* 7.4, pp. 457–472. ISSN: 0883-4237. DOI: 10.1214/ss/1177011136.

Golub, Gene H. and Charles F. Van Loan (2013). *Matrix computations*. Fourth edition. Johns Hopkins studies in the mathematical sciences. Baltimore: The Johns Hopkins University Press. ISBN: 978-1-4214-0794-4.

Grimmett, Geoffrey and David Stirzaker (2001). *Probability and random processes*. 3rd ed. Oxford ; New York: Oxford University Press. ISBN: 978-0-19-857223-7 978-0-19-857222-0.

Higham, Nicholas J. (2008). *Functions of Matrices: Theory and Computation*. en. Society for Industrial and Applied Mathematics. ISBN: 978-0-89871-646-7 978-0-89871-777-8. DOI: 10.1137/1.9780898717778.

Hirschi, Travis (1974). *Causes of delinquency*. eng. 1. paperback ed., 3. print. OCLC: 174209059. Berkley: Univ. of California Press. ISBN: 978-0-520-01901-0.

Jewkes, Yvonne and Gayle Letherby, eds. (2002). *Criminology: a reader*. en. OCLC: ocm48885121. London ; Thousand Oaks, Calif: SAGE. ISBN: 978-0-7619-4710-3 978-0-7619-4711-0.

Johnson, Shane D. and Lucia Summers (2015). "Testing Ecological Theories of Offender Spatial Decision Making Using a Discrete Choice Model". en. In: *Crime & Delinquency* 61.3, pp. 454–480. ISSN: 0011-1287, 1552-387X. DOI: 10.1177/0011128714540276.

Kurland, Justin, Shane D. Johnson, and Nick Tilley (2014). "Offenses around Stadiums: A Natural Experiment on Crime Attraction and Generation". en. In: *Journal of Research in Crime and Delinquency* 51.1, pp. 5–28. ISSN: 0022-4278, 1552-731X. DOI: 10.1177/0022427812471349.

Marchant, Roman et al. (2018). "Applying machine learning to criminology: semi-parametric spatial-demographic Bayesian regression". en. In: *Security Informatics* 7.1. ISSN: 2190-8532. DOI: 10.1186/s13388-018-0030-x.

Mohler, George (2013). "Modeling and estimation of multi-source clustering in crime and security data". In: *The Annals of Applied Statistics* 7.3, pp. 1525–1539. ISSN: 1932-6157. DOI: 10.1214/13-AOAS647.

Moller, Jesper, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen (1998). "Log Gaussian Cox Processes". en. In: *Scandinavian Journal of Statistics* 25.3, pp. 451–482. ISSN: 0303-6898, 1467-9469. DOI: 10.1111/1467-9469.00115.

Office For National Statistics, National Records Of Scotland, and Northern Ireland Statistics And Research Agency (2016). *2011 Census aggegate data (Data downloaded: 1 June 2016)*.

Osgood, D. Wayne (2000). "Poisson-Based Regression Analysis of Aggregate Crime Rates". In: *Journal of Quantitative Criminology* 16.1, pp. 21–43. ISSN: 07484518, 15737799.

Pratt, Travis, Jacinta Gau, and Travis Franklin (2018). "Key Ideas in Criminology and Criminal Justice". In: Thousand Oaks: SAGE Publications, Inc. DOI: 10.4135/9781483388045.

Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. OCLC: ocm61285753. Cambridge, Mass: MIT Press. ISBN: 978-0-262-18253-9.

Roberts, Gareth O. and Jeffrey S. Rosenthal (2009). "Examples of Adaptive MCMC". en. In: *Journal of Computational and Graphical Statistics* 18.2, pp. 349–367. ISSN: 1061-8600, 1537-2715. DOI: 10.1198/jcgs.2009.06134.

Rocque, Michael, Brandon C. Welsh, and Adrian Raine (2012). "Biosocial criminology and modern crime prevention". en. In: *Journal of Criminal Justice* 40.4, pp. 306–312. ISSN: 00472352. DOI: 10.1016/j.jcrimjus.2012.05.003.

Rogers, Simon and Mark Girolami (2017). *A first course in machine learning*. eng. Second Edition. Chapman & Hall/CRC machine learning & pattern recognition series. OCLC: 967702297. Boca Raton London New York: CRC Press, Taylor & Francis Group, a Chapman & Hall book. ISBN: 978-1-4987-3848-4 978-1-4987-3856-9.

Rue, Håvard et al. (2017). "Bayesian Computing with INLA: A Review". In: *Annual Review of Statistics and its Application* 4.1. ISSN: 2326-8298.

Saatçi, Yunus (2012). "Scalable inference for structured Gaussian process models". PhD Thesis. Citeseer.

Shewchuk, Jonathan Richard (1994). *An introduction to the conjugate gradient method without the agonizing pain*. Carnegie-Mellon University. Department of Computer Science.

Stein, Michael L (1999). *Interpolation of spatial data: some theory for kriging*. English. OCLC: 968504419. Place of publication not identified: Springer. ISBN: 978-1-4612-1494-6.

Taddy, Matthew A. (2010). "Autoregressive Mixture Models for Dynamic Spatial Poisson Processes: Application to Tracking Intensity of Violent Crime". en. In: *Journal of the American Statistical Association* 105.492, pp. 1403–1417. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/jasa.2010.ap09655.

Wilson, Andrew Gordon et al. (2014). "Fast Kernel Learning for Multidimensional Pattern Extrapolation". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Cambridge, MA, USA: MIT Press, pp. 3626–3634.

Young, G. A and Richard L Smith (2005). *Essentials of statistical inference: G.A. Young, R.L. Smith*. English. OCLC: 61410200. Cambridge, UK; New York: Cambridge University Press. ISBN: 978-0-511-12616-1 978-0-511-12402-0 978-0-521-83971-6 978-0-511-75539-2.

Young, Jock (1992). ""Ten Points of Realism" in Rethinking Criminology: The Realist Debate". In: *Rethinking Criminology: The Realist Debate*. Ed. by Jock Young and Roger Matthews. London: Sage. ISBN: 0-8039-8621-1.