On evaluating the Watanabe-Akaike information criteria for Bayesian modelling of point referenced spatial data

University of Southampton

Faculty of Social Sciences

School of Mathematical Sciences

Ho Man Theophilus Chan

ORCID ID 0000-0001-6821-4206

Thesis for the degree of Doctor of Philosophy

July 2024

Abstract

Bayesian modelling of point referenced data estimates the relationships among geospatial variables and enables predictions at unobserved locations. The Watanabe-Akaike information criterion (WAIC) is a model selection criterion for determining the best model configuration from a set of competing candidate models. However, the traditional formulation of the WAIC assumes conditional independence in the outcome variables, an assumption violated by the spatial dependence often exhibited by point referenced spatial data. Consequently, spatial models for point referenced data violate the conditional independence assumption. To address this problem, this thesis introduces the $\mathrm{WAIC}_{\mathrm{NF}}$, a novel approach employing likelihoods for non-factorisable models that consider conditional dependencies for WAIC calculation.

We apply the $\mathrm{WAIC}_{\mathrm{NF}}$ to real-world spatial modelling using the 2018 Nigeria Demographic and Health Survey data, focusing on the coverage of the first-dose of the measles-containing vaccine (MCV1). We construct models with different covariance functions and fit them using the integrated nested Laplace approximation – stochastic partial differential equation (INLA-SPDE) method. We observe notable differences in the $\mathrm{WAIC}_{\mathrm{NF}}$ values, in comparison with the WAIC values computed by default using the R-INLA package. Additionally, we extend our $\mathrm{WAIC}_{\mathrm{NF}}$ application to the MCV1 dataset by fitting spatial models with the nearest-neighbour Gaussian process (NNGP) method in Stan, effectively identifying the spatial model with the optimal covariance function specifications for the MCV1 dataset.

This thesis contributes to existing knowledge on model selection methods for point referenced spatial data with the $\mathrm{WAIC}_{\mathrm{NF}}$. Our findings highlight its effectiveness for model selection, particularly when choosing among spatial models with different covariance functions. Its integration with two Bayesian spatial modelling platforms — INLA-SPDE in R and NNGP in Stan — enhances its utility and provides a robust model selection framework for Bayesian modelling of point referenced spatial data.

Declaration of Authorship

I, Ho Man Theophilus Chan, declare that this thesis titled "On evaluating the Watanabe-Akaike information criteria for Bayesian modelling of point referenced spatial data", and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given, With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed . . . .

Date . . . .

Acknowledgements

I would like to thank my supervisors, Professor Sujit Sahu and Dr Edson Utazi, for their patience, guidance and friendship throughout my PhD journey. To Sujit, for agreeing to take me under his wing, and for providing expert advice, boundless suggestions and invaluable recommendations. To Edson, for generously offering me endless opportunities and for continuously motivating me. Their combined mentorship has been pivotal in shaping both my academic and personal growth.

To my colleagues, for their constant words of encouragement. To my dear friends, for providing me with the much-needed balance to my work. To my parents and my loved one, for their unwavering support, patience and encouragement through every step of my academic journey. Without them, I would not have reached this point in my career. To my brother, Boaz, let us complete our PhDs strong together.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Review of classic model selection methods

Statistical modelling can help statisticians better understand data by estimating relationships between variables of interest, inferring underlying properties of the data and facilitating predictions. The utility of a statistical model lies in its ability to capture the essential features of a dataset. While there is no “perfect” model, when a model fits the data well, it can offer a useful and good representation of the inherent structure and characteristics of the dataset. Therefore, it is crucial to select an appropriate model for the specific dataset under consideration.

For instance, when dealing with a spatial dataset where the outcomes are distributed across a geographic domain, a simple linear regression model may be employed as the first choice. However, a simple linear regression model often prove inadequate in capturing the inherent spatial dependencies within the spatial dataset. Spatial models, designed to account for these spatial characteristics, may be a more suitable alternative. Despite the intuitive appeal of spatial models, it is imperative to substantiate the decision for employing a spatial model over a simple linear regression model. For this, statisticians rely on robust model selection methods. These techniques play an important role in affirming that the chosen model offers a more accurate representation of the data compared to the other candidates. Revisiting the spatial dataset example, model selection methods can help determine whether a spatial model provides a better fit to the spatial dataset when compared to a simple linear regression model.

Consider another scenario where independent variables, referred to as covariates or predictors, are introduced into the model. The selection of covariates necessitates the use of model selection methods to identify the optimal combination. This process involves a delicate balance between the advantages gained from including additional variables and the potential drawbacks associated with their omission. This decision impacts how well the model captures the underlying structure and characteristics of the data.

The scenarios presented above motivate the need for model selection methods. The objective of model selection is to identify a model that is good enough where it can be "useful" (Box, 1976). Conversely, the objective of model selection is to avoid selecting models that are obviously poor.

When comparing two models, one of which is nested within the other, the model with more covariates typically provides a better fit to the data. Although this may intuitively suggest a preference for models with more covariates, such a preference has inherent disadvantages, primarily related to increased model complexity. Models with additional covariates often exhibit greater model complexity, making them more challenging to interpret and compute. Moreover, adhering to the scientific principle of parsimony implies a preference for simpler models unless the expansion of the model and the increase in complexity can be adequately substantiated.

In the non-Bayesian framework, models can be compared using the coefficient of determination, denoted as R^2. It serves as a measure of goodness-of-fit by calculating the ratio between the sum of squares for the regression and the total sum of squares. However, R^2 does not account for model complexity and consistently favours models with more variables, thus contradicting the aforementioned scientific principle of parsimony, and resulting in computational and interpretation challenges.

Model selection methods in the non-Bayesian framework that account for model complexity include the adjusted R^2, Mallows C_p (Mallows, 1964, 2000) and the Akaike information criterion (Akaike, 1973). These model selection methods account for model complexity by incorporating the number of covariates into their calculations. These selection methods strike a balance between the fit of the model to the data and its complexity. Another tool for model selection is the F-test. It is often employed to identify the model that best fits the data. The F-test is commonly used in the context of stepwise, forward and backward model selection procedures for nested models (Kadane and Lazar, 2004).

In the Bayesian framework, the preferred method for model comparison and selection is the Bayes factor (Kass and Raftery, 1995). However, practical implementation of the Bayes factor is challenging in most cases due to the complexities associated with high-dimensional integration problems. Complications in implementing the Bayes factor also arises when improper prior distributions are specified, or when dealing with models that involve large datasets, both of which are commonly encountered in practice (Gelfand and Ghosh, 1998).

Instead, an easier to implement model selection criterion is the deviance information criterion (DIC), often referred to as the "Bayesian version of the Akaike information criterion" due to their similar formula and asymptotic properties. We will further elaborate this in Chapter 2. However, the DIC has faced criticism for not being "fully Bayesian", since its formulation involves a plug-and-use factor with a Bayes estimate (Gelman et al., 2014). A more fully Bayesian model selection criterion is the Watanabe-Akaike information criterion (WAIC). Similar to the Akaike information criterion and Mallows C_p, the DIC and WAIC both not only assess the fit of the model, but also account for the model complexity.

The predictive model choice criterion (PMCC) is another model selection method in the Bayesian context. It is based on the minimum posterior predictive loss approach (Gelfand and Ghosh, 1998). The PMCC appears to be appropriate for many classes of hierarchical models and for correlated data models, such as spatial models, because it depends directly on the posterior predictive distribution rather than the likelihood function. Similar to the DIC and WAIC, the PMCC comprises two components. A component to measure the fit of the model on the data, and a component that accounts for model complexity.

1.2 Motivation

Models for spatial data present a fundamental challenge to the traditional model comparison strategies discussed in the preceding section due to their violation of the independence assumption (Hoeting et al., 2006). In spatial datasets, outcomes often exhibit spatial dependence, and capturing this spatial dependence constitutes a key characteristic of spatial models. As a result, models for spatial data violate the assumption of independence. For example, while the calculation of the Akaike information criterion (AIC) relies on likelihood functions that assume conditional independence among the outcomes of the data, they continue to be a prevalent choice for model selection in practical spatial related applications (Liang, 2012, Vahedi Saheli and Effati, 2021). We will reserve a more detailed exploration of the AIC for Chapter 2, but it is important to acknowledge that failure to account for spatial dependence among spatial data can result in model selection criteria favouring overly complex models. This oversight, in turn, leads to greater uncertainty regarding model parameters, poorer prediction accuracy and misguides inferential conclusions (Duncan and Mengersen, 2020).

In the field of ecological sciences, the spatial leave-one-out cross-validation (SLOO) method is a popular model selection method (Pohjankukka et al., 2017). While primarily employed for variable selection tasks within the non-Bayesian framework, the SLOO method was developed to address the limitation of the conventional leave-one-out cross validation (LOO), that is when dealing with structured data like spatial datasets (Le Rest et al., 2014). The SLOO method involves four key steps: The first step removes one observation from the dataset. Next, remove all observations that are spatially correlated with the removed observation in the first step. Subsequently, fit a model using the remaining data and make a prediction at the location of the observation removed in the first step, based on the estimated model parameters. Finally, calculate a score between the observed and predicted value. Le Rest et al. noted that the SLOO criterion calculation used by Pohjankukka et al. (2017) for selecting data sampling density for new research area relies on the likelihood instead of the classical sum of square of errors. As a result, in the absence of spatial autocorrelation, the SLOO method aligns with the AIC (Le Rest et al., 2014). Although the SLOO method represents a valid approach for model selection for spatial data, we do not further explore it in detail within this thesis due to our primary focus on Bayesian model selection methods. However, we will provide further detail regarding LOO in Chapter 2.

Within the Bayesian framework, model selection methods include the deviance information criterion (DIC) and the Watanabe-Akaike information criterion (WAIC). The computation of these criteria rely on likelihood functions that assume independence among the outcomes of the data; we will further elaborate this in Chapter 2. Specifically, the WAIC requires the dataset to be partitioned into independent parts (Gelman et al., 2014). Again, however, outcomes from spatial data exhibit spatial dependence, which violates the required independence assumption of the likelihood functions. Spatial dependence is a defining characteristic of spatial datasets and is a key characteristic captured in spatial models. This inherent spatial dependence fundamentally contradicts the independence assumption required for DIC and WAIC calculation.

Bayesian model selection methods for spatial models remains relatively understudied within the existing literature, leading to a lack of guidance regarding the appropriate formulations of model selection methods when there is spatial dependence among the outcomes of the data. While popular Bayesian spatial model fitting methods provide access to the WAIC, detail on the explicit calculation and implementation are lacking in available resources, as we will further discuss in Chapter 3. Furthermore, even when WAIC calculations are conducted, their reliance on conditionally independent likelihood functions persist as a significant challenge.

In response to this challenge, this thesis introduces an innovative approach to the WAIC computation. This alternative approach utilises the likelihood functions of non-factorisable models. As we will elaborate in Chapter 3, these likelihood functions serve as a basis for calculating the $\mathrm{WAIC}_{\mathrm{NF}}$, a novel criterion providing a practical solution to Bayesian model selection in the presence of spatial data exhibiting conditional dependencies.

1.3 Thesis structure

This thesis is organised as follows.

In Chapter 2, we will provide the essential background information necessary for a comprehensive understanding of this thesis. We will start by introducing point referenced spatial data and their fundamental properties. This introduction will lay the groundwork for our exploration of variograms and covariance functions. On the covariance functions, we will emphasise the Matérn covariance function and the exponential covariance function, both of which will reappear throughout the subsequent chapters. These functions will assume important roles in Chapter 4, where we will employ them in simulation examples, as well as in Chapters 5 and 6, where we will employ them in real-world spatial modelling scenarios. Chapter 2 will also delve into the construction of models for point referenced spatial data and the methods employed for fitting them within the Bayesian framework. Particularly, we will focus on the integrated nested Laplace approximation (INLA) method and the Stan method, which we will utilise in Chapters 5 and 6 respectively. To conclude Chapter 2, we will introduce various model selection techniques, and focus particular on the Watanabe-Akaike information criteria (WAIC) and the Pareto-smoothed importance sampling leave-one-out cross validation transformed to the deviance scale (PSIS-LOOIC).

In Chapter 3, we will address the issue related to the formulation of the WAIC, which will have been introduced in Chapter 2. We will then introduce the readers to non-factorisable models and their associated likelihood functions, and we will provide the explicit derivations of these non-factorisable model likelihoods. We will conclude this chapter by presenting the algorithm for the computation of the non-factorisable model likelihoods. Furthermore, we will describe how these likelihoods can be implemented to calculate the WAIC and PSIS-LOOIC, which we will call the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, respectively. The WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ will be the primary focus for the remainder of the thesis. In Chapter 4 we will illustrate their practical utility through simulation examples. In Chapters 5 and 6 we will apply them to real-world spatial modelling scenarios to demonstrate their relevance and effectiveness.

Chapter 4 illustrates the practical utility of the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ through simulation examples conducted in the statistical programming language R. The primary objective of the simulation examples is to investigate the difference in performance between our proposed selection criteria and existing selection criteria, namely the WAIC calculated by INLA and the WAIC and PSIS-LOOIC calculated from the log likelihoods extracted from Stan. Specifically, we will investigate two selection tasks: model selection and variable selection. For the model selection task, our objective is to evaluate the ability of these selection criteria to correctly identify the covariance functions, as introduced in Chapter 2, that underlie the generated datasets. For the variable selection task, the objective is to evaluate the ability of these selection criteria to correctly identify the combinations of covariates employed in generating the datasets. These investigations will help us understand both strengths and limitations of the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ for model selection for models of point referenced spatial data.

In Chapter 5, we will apply the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, introduced in Chapter 3, to a real-world spatial modelling scenario. The dataset under consideration is the 2018 Nigeria Demographic Health Survey program dataset that focuses on the coverage of the first-dose measles-containing vaccine (MCV1). Our approach will involve constructing the candidate spatial models for the MCV1 dataset, and incorporating the various covariance functions introduced in Chapter 2. Subsequently, we will employ the INLA method, as detailed in Chapter 2, to fit these models. Finally, we will select the optimal model using the proposed WAIC_{NF} and the WAIC calculated by INLA. This analysis will enhance our understanding of the practical applications and effectiveness of the WAIC_{NF} within the context of real-world spatial modeling.

We will begin Chapter 6 with a discussion on the methodologies proposed in the literature to address the challenges posed by large spatial datasets. The methodologies that we discuss include the stochastic partial differential equation approach, which we implemented in conjunction with the INLA method within Chapter 5. However, the primary focus of this chapter will be on the nearest-neighbour Gaussian process (NNGP) approach. We will extend the application of the proposed WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ to the MCV1 dataset, utilising the NNGP approach implemented in Stan. Chapter 6 will conclude with a summary of the findings.

In Chapter 7, we will conclude this thesis by providing a summary of the results and findings from preceding chapters. We will close this chapter with a brief discussion on the potential extensions of the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ in the spatiotemporal context.

Chapter 2

Background

2.1 Spatial data

Spatial data encompass observations within a defined study domain and consist of geolocation information such as latitude and longitude coordinates from the geographic coordinate system or northing and easting from the Universal Transverse Mercator coordinate system. Spatial data typically fall into one of three distinct categories: areal data, point pattern data, or point referenced data.

2.1.1 Areal data

Suppose the study domain $\mathcal{D}$ is partitioned into a finite number of subregions. Areal data comprise the aggregated outcomes from each these subregion. In literature, areal data are commonly referred to as lattice data (Cressie, 2015, Moraga, 2019, Gómez-Rubio, 2020) and are frequently depicted as checkerboards in trivial examples. In practical applications, however, areal data are often observed as irregular shapes. The partition of $\mathcal{D}$ in practical applications typically corresponds to administrative units such as states, counties or districts. The aggregation of the outcomes can be summary statistics such as the sum, the mean or the median of the observations.

A popular application of areal data analysis is estimating disease risks. In this context, the aggregated outcomes are ratios between the observed number of cases of the disease under investigation and the expected number of cases within each subregion. These ratios are referred to as standard mortality rates (SMRs). While SMRs provide useful exploratory information, they are susceptible to misinterpretation. Notably, subregions with smaller populations may yield extreme SMRs since population size and disease rarity are not accounted for in the calculation of the SMRs (Moraga, 2019). To address this issue, researchers commonly employ the Besag York Mollié (BYM) model (Besag et al., 1991).

The BYM model incorporates random effect components to account for both structured and unstructured randomness present in the data. The spatial random effect captures spatial correlation within the dataset, where outcomes in neighbouring subregions may exhibit stronger associations compared to outcomes in subregions located farther apart. The BYM model also allow the incorporation of additional information that may further explain disease risks. Information borrowed from the random effects and covariates help mitigate the extreme SMRs resulting from small populations or rarer diseases (Gelfand et al., 2010), providing results conducive to more reasoned conclusions.

Practical applications of areal data analyses are found in various research domains including political sciences (y Perdomo, 2004, Harbers, 2017), econometrics (Pineda-Ríos et al., 2019, Laurent and Margaretic, 2021) and environmental sciences (Wang et al., 2018, Lee et al., 2020).

2.1.2 Point pattern data

Point pattern data consist of mapped point locations within a study domain, which collectively represent a spatial pattern. Point pattern data analyses emphasise on understanding the randomness associated with the location of the points (Banerjee et al., 2014). For example, point pattern data analyses involve examining whether these points are randomly dispersed throughout the study domain or if they are located in a non-random pattern, such as clustering patterns.

Point processes is an important concept related to point pattern data analyses. Point processes are stochastic models that describe the locations of events of interest. When considering a given point pattern dataset as realisations from a particular point process, a point process model can be used to identify the distributions of the locations, estimate the intensity of the events and learn more about the correlation between the spatial locations and spatial variables (Moraga, 2019). For example, the homogeneous Poisson process is a point process model where realisations are equally likely to occur at any location within the study domain, independently of the locations of other events. Another example is the Log-Gaussian Cox process (Møller et al., 1998, Diggle et al., 2013), where realisations exhibit varying intensity, and the locations of the events follow a probability distribution.

Practical applications of point pattern data analyses are found in various research domains, such as astronomy (Babu and Feigelson, 1996), forestry (Stoyan and Penttinen, 2000), ecology (Perry et al., 2006, Wiegand and Moloney, 2013) and epidemiology (Gatrell et al., 1996). Other examples of applications of point pattern data analyses as well as an overview of the applications can be found in Møller and Waagepetersen (2003).

2.1.3 Point referenced data

Point referenced data, also referred to as geostatistical data in literature (Moraga, 2019, Gómez-Rubio, 2020), encompass observations recorded within a specified study domain $\mathcal{D}$. In the context of point referenced data, $\mathcal{D}$ represents a single continuous surface that allows outcomes to be observed anywhere within its boundaries. Notably, the locations of point referenced data are fixed. Examples of such fixed locations include weather stations tasked with monitoring air quality indices or hospitals collecting vaccination records.

A main objective in point referenced data analysis is making predictions at locations where observations have not been recorded. Often, this involves constructing a prediction surface over $\mathcal{D}$ by utilising kriging techniques within a non-Bayesian framework or posterior predictive distributions from a Bayesian modelling approach. Bayesian modelling of point referenced data has experienced an increase in popularity, driven by advancements of modern statistical computation tools such as the integrated nested Laplace approximation method (Rue et al., 2009) which has facilitated practical implementation.

Practical applications of point referenced data analyses include modelling healthcare and development indicators (Pezzulo et al., 2023, Utazi et al., 2023) and environmental variables (Sahu et al., 2006, Hammond et al., 2020, Sahu et al., 2020). The primary focus of this thesis will be on point referenced data.

2.2 Definitions and conventions

Following the traditional convention from literature, a location or site in a specified study domain $\mathcal{D}$ is denoted $\mathbf{s}$ and a collection of n sites is written as $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. Capitalised letters denote random variables and lower case letters are reserved for realisations of the corresponding random variable unless specified. For example, if Y is a random variable representing ozone concentration, then y is a realisation of this random variable and is a numerical value with some unit of ozone concentration.

In this thesis, if a random variable ξ follows from a distribution F, we write $\xi \sim F$. For example,

$$\boldsymbol{\omega} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}).$$

In this example, $\boldsymbol{\omega}$ is distributed as an n-dimensional multivariate normal distribution $N_n(\cdot)$ with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$. We use the subscript $\boldsymbol{\omega}$ in the covariance $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ to explicitly denote that the covariance matrix is related to $\boldsymbol{\omega}$. The bold font is reserved for vectors, so explicitly, $\boldsymbol{\omega} = (\omega_1, \omega_2, \ldots, \omega_n)'$ and $\mathbf{0} = (0_1, 0_2, \ldots, 0_n)'$, where $'$ denotes the transpose operator. We will provide formal definitions for the components within this example equation later in Section 2.4.3. For now, we emphasise on the notations that will be used in this thesis.

2.3 Properties of point referenced spatial data

A process is spatial when for $d \geq 2$,

$$\{Y(\mathbf{s}) : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d\},$$

and realisations of this spatial process are point referenced spatial data. The notation $Y(\mathbf{s})$ denotes a random variable Y at site $\mathbf{s}$.

2.3.1 Stationarity

There are three types of stationarity: strong stationarity, weak stationarity and intrinsic stationarity. We begin this discussion by first denoting $\mathbf{h}$ as a distance vector.

Strong stationarity

A spatial process has strong stationarity if the random variables $\big(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\big)'$ have the same distribution as $\big(Y(\mathbf{s}_1 + \mathbf{h}), \ldots, Y(\mathbf{s}_n + \mathbf{h})\big)'$ for any $\mathbf{h}$ (Banerjee et al., 2014). In other words, if shifting the sites by $\mathbf{h}$ does not change the distribution of the random variables, the spatial process is strongly stationary.

Weak stationarity

In practice, strong stationarity does not always hold for spatial processes. Weak stationarity is a more relaxed assumption for spatial processes than strong stationarity. A spatial process is weakly stationary if the distribution of random variables $\big(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\big)'$ and $\big(Y(\mathbf{s}_1 + \mathbf{h}), \ldots, Y(\mathbf{s}_n + \mathbf{h})\big)'$ have a constant mean,

$$E\big[Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\big] = E\big[Y(\mathbf{s}_1 + \mathbf{h}), \ldots, Y(\mathbf{s}_n + \mathbf{h})\big],$$

and the covariance function between $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_i + \mathbf{h})$ is a function dependent only on $\mathbf{h}$, that is $C\big(Y(\mathbf{s}_i),\ Y(\mathbf{s}_i + \mathbf{h})\big) = C(\mathbf{h})$ (Banerjee et al., 2014). It should be noted that a strongly stationary spatial process is also a weakly stationary spatial process. However, a weakly stationary spatial process is not necessarily a strongly stationary spatial process with the exception of the Gaussian process.

Intrinsic stationarity

A spatial process has intrinsic stationarity if the variance of $Y(\mathbf{s}) - Y(\mathbf{s} + \mathbf{h})$ does not depend on the location $\mathbf{s}$ and is only dependent on $\mathbf{h}$. Let $\mathrm{Var}(\cdot)$ denote the variance. Banerjee et al. (2014) defines

$$\mathrm{Var}\big(Y(\mathbf{s} + \mathbf{h}) - Y(\mathbf{s})\big) = 2\gamma(\mathbf{h}),$$

where $2\gamma(\mathbf{h})$ is called the variogram and $\gamma(\mathbf{h})$ is called the semivariogram. Notice how the semivariogram is a function dependent only on $\mathbf{h}$. We will come back to the variogram and semivariogram later in Section 2.3.3.

Nonstationarity

Spatial processes can also be nonstationary. Banerjee et al. (2014) discussed strategies for nonstationary spatial processes. For example, one strategy involves adjusting the mean structure while specifying the covariance function from those proposed by Cressie and Johannesson (2008). Another strategy involves deformation (Sampson and Guttorp, 1992) where the specified study domain is transformed to a new domain such that stationarity holds. For a comprehensive exploration of nonstationarity, we recommend Gneiting (2002), Schmidt and O'Hagan (2003), and Banerjee et al. (2014). We note that the ensuing discussion within this thesis will focus only on stationary spatial processes.

2.3.2 Isotropy

If the semivariogram $\gamma(\mathbf{h})$ depends only on the length of the distance vector $\mathbf{h}$, denoted as $||\mathbf{h}||$, the semivariogram is isotropic. More specifically, we define $||\mathbf{h}||$ as the L2 norm, the Euclidean distance between two locations. When the distance $||\mathbf{h}||$ is short, $Y(\mathbf{s}+\mathbf{h})$ and $Y(\mathbf{h})$ are expected to be very similar. Conversely, as the distance $||\mathbf{h}||$ increases, the similarity between $Y(\mathbf{s}+\mathbf{h})$ and $Y(\mathbf{h})$ is expected to diminish. Isotropy is a useful property as it assumes that the spatial correlation function $\rho(\cdot)$ relies exclusively on the Euclidean distance between sites and avoids the need for additional directional parameters.

If directions of $\mathbf{h}$ are taken into consideration, the semivariogram $\gamma(\mathbf{h})$ is anisotropic. Under anisotropy, $||\mathbf{h}||$ may vary in each direction. Banerjee et al. (2014) discussed two strategies for anisotropic spatial processes. The first strategy involves separating the variogram into directional components, such as latitude and longitude. The second strategy involves the consideration of nested models, where the angles associated with $\mathbf{h}$ are categorised into classes with different variograms specified for each class of angles.

The incorporation of additional components to account for directions of $\mathbf{h}$ provide insights into the directional behaviour of the variogram. However, introducing more parameters into the variogram also increases the complexity of the function. For a complex variogram, it may be difficult to identify and learn about the additional directional parameters unless a lot of locations are available, which may not always be the case. Furthermore, the extent of possible contour shapes induced by anisotropy is limited by the positive definitiveness of the spatial correlation function (Banerjee et al., 2014). We note that the discussion in the remainder of this thesis will focus on isotropic spatial processes.

2.3.3 Semivariograms and covariance functions

In Section 2.3.1, we have directly provided the definition of the variogram. Here, we revisit the definition and provide the explicit derivations. Let $E(\cdot)$ denote the expectation, $\mathrm{Var}(\cdot)$ denote the variance, $C(\cdot)$ denote the covariance function and assume that $E\big(Y(\mathbf{s}+\mathbf{h})-Y(\mathbf{s})\big)=0$. The derivation starts from the variance of the difference between $Y(\mathbf{s}+\mathbf{h})$ and $Y(\mathbf{s})$, that is

$$\begin{aligned}
\mathrm{Var}\big(Y(\mathbf{s}+\mathbf{y})-Y(\mathbf{s})\big) &= E\Big(Y(\mathbf{s}+\mathbf{h})-Y(\mathbf{s})-E\big(Y(\mathbf{s}+\mathbf{h})-Y(\mathbf{s})\big)\Big)^2,\\
&= E\Big(Y(\mathbf{s}+\mathbf{h})-Y(\mathbf{s})-E\big(Y(\mathbf{s}+\mathbf{h})\big)+E\big(Y(\mathbf{s})\big)\Big)^2,\\
&= E\Big(Y(\mathbf{s}+\mathbf{h})-E\big(Y(\mathbf{s}+\mathbf{h})\big)-Y(\mathbf{s})+E\big(Y(\mathbf{s})\big)\Big)^2,\\
&= E\Big(Y(\mathbf{s}+\mathbf{h})-E\big(Y(\mathbf{s}+\mathbf{h})\big)-\big(Y(\mathbf{s})-E\big(Y(\mathbf{s})\big)\big)\Big)^2,\\
&= E\Big(\big(Y(\mathbf{s}+\mathbf{h})-E\big(Y(\mathbf{s}+\mathbf{h})\big)\big)^2+\big(Y(\mathbf{s})-E\big(Y(\mathbf{s})\big)\big)^2\\
&\qquad -2\big(Y(\mathbf{s}+\mathbf{h})-E\big(Y(\mathbf{s}+\mathbf{h})\big)\big)\big(Y(\mathbf{s})-E\big(Y(\mathbf{s})\big)\big)\Big)^2,\\
&= \mathrm{Var}\big(Y(\mathbf{s}+\mathbf{h})\big)+\mathrm{Var}\big(Y(\mathbf{s})\big)-2C\big(Y(\mathbf{s}+\mathbf{h}),Y(\mathbf{s})\big).
\end{aligned}$$

Recall that under weak stationarity, $C\big(Y(\mathbf{s}+\mathbf{h}),Y(\mathbf{s})\big)=C(\mathbf{h})$. An implication of the weak stationarity property is $\mathrm{Var}\big(Y(\mathbf{s}+\mathbf{h})\big)=\mathrm{Var}\big(Y(\mathbf{s})\big)=C(\mathbf{0})$. We continue the derivation above

$$\begin{aligned}
\mathrm{Var}\big(Y(\mathbf{s}+\mathbf{h})-Y(\mathbf{s})\big) &= \mathrm{Var}\big(Y(\mathbf{s}+\mathbf{h})\big)+\mathrm{Var}\big(Y(\mathbf{s})\big)-2C\big(Y(\mathbf{s}+\mathbf{h}),Y(\mathbf{s})\big),\\
&= C(\mathbf{0})+C(\mathbf{0})-2C(\mathbf{h}),\\
&= 2C(\mathbf{0})-2C(\mathbf{h}),\\
&= 2\big(C(\mathbf{0})-C(\mathbf{h})\big),\\
&= 2\gamma(\mathbf{h}).
\end{aligned}$$

Therefore, the definitions of the variogram $2\gamma(\mathbf{h})$ and the semivariogram $\gamma(\mathbf{h})$ in their simplest form are

$$2\gamma(\mathbf{h}) = 2\big(C(\mathbf{0}) - C(\mathbf{h})\big), \tag{2.1}$$

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}). \tag{2.2}$$

To simplify the notation, let us define $d = ||\mathbf{h}||$ for the following discussion. The semivariogram $\gamma(d)$ is useful for visualising spatial variances. The two most commonly used semivariograms are:

Exponential semivariogram

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2(1 - \exp(-\phi d)) & \text{if } d > 0, \\ 0 & \text{if } d = 0. \end{cases} \tag{2.3}$$

Matérn semivariogram

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2\left(1 - \frac{(\sqrt{2\nu}d\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu}d\phi)\right) & \text{if } d > 0, \\ 0 & \text{if } d = 0, \end{cases} \tag{2.4}$$

where $\Gamma(\cdot)$ denotes the mathematical Gamma function, $K_\nu(\cdot)$ denotes the modified Bessel function of the second kind of order ν, ν is the smoothness parameter and ϕ is the spatial decay parameter.

The asymptotic properties of the semivariogram functions reveal several important properties. The asymptotic value of the semivariogram $\lim_{d\to 0}\gamma(d) = \tau^2$ is referred to as the nugget. Notice that in Equations (2.3) and (2.4) there is a discontinuity within the semivariogram functions. As the separation distance approaches zero, the semivariogram converges to τ^2, yet when $d = 0$, the resulting value is zero. Banerjee et al. explain that the semivariogram is set to zero when $d = 0$ to avoid attributing all errors as spatial errors. As the separation distance becomes very small, the errors begin to reflect spatial residuals. However, at $d = 0$, there may be errors unaccounted for by the spatial residuals such as measurement errors, collection errors, replication errors or micro-scale errors (Banerjee et al., 2014).

The asymptotic value of the semivariogram $\lim_{d\to\infty}\gamma(d) = \tau^2 + \sigma^2$ is referred to as the sill. Notice that there are two components involved within the sill: the nugget τ^2 and the σ^2 parameter called the partial sill. The value of d when the semivariogram reaches the sill is called the range. All $\gamma(d)$ values for d less than the range are spatially autocorrelated. Since the still is obtained when $d \to \infty$, the range is infinite and all $\gamma(d)$ values are spatially autocorrelated. For practical interpretation, the effective range serves as a more meaningful metric of d. The effective range indicates the distance where spatial autocorrelation is negligible. In practice, this is when the correlation function $\rho(\cdot)$ reaches a value of 0.05.

The semivariogram also enables the derivation of the covariance function. For modelling purposes, the covariance function needs to be specified for the likelihood function to be written for parameter estimation. The variogram function and the covariance function have the following relationship. Recall, the semivariogram is given as

$$\gamma(d) = C(\mathbf{0}) - C(d).$$

Assuming that the spatial process is ergodic, $C(d) \to 0$ as $d \to \infty$. The ergodic assumption follows American geographer Waldo R. Tobler's first law of geography, which states that objects located closer to each other are more closely related than objects located further apart (Tobler, 1970). Data values close together in the geographical space tend to be more alike than data values that are further apart. Observation on a variable at location i carries some information about what is observed for the same variable in areas that are close to i (Haining and Li, 2020). Using the ergodic assumption in Equation (2.2) gives

$$\begin{aligned}\lim_{d\to\infty}\gamma(d) &= C(\mathbf{0}) - \lim_{d\to\infty} C(d),\\ &= C(\mathbf{0}) - 0,\\ &= C(\mathbf{0}).\end{aligned}$$

The ergodic assumption is useful as it allows the relationship between the semivariogram function and the covariance function to be rewritten in terms of the semivariograms

$$\begin{aligned}\gamma(d) &= C(\mathbf{0}) - C(d),\\ C(d) &= C(\mathbf{0}) - \gamma(d),\\ C(d) &= \lim_{u\to\infty}\gamma(u) - \gamma(d).\end{aligned}$$

In the derivation above, $\gamma(u)$ is also a semivariogram where we define $u = ||\mathbf{h}||$. The notation change is to explicitly show that the limit should only be taken for the first term on the right hand side of the equation. From the exponential semivariogram (2.3) and the Matérn semivariogram (2.4), the exponential covariance function and the Matérn covariance function are given as follows:

Exponential covariance function

$$C(d) = \begin{cases} \sigma^2 \exp(-\phi d) & \text{if } d > 0, \\ \tau^2 + \sigma^2 & \text{if } d = 0. \end{cases} \tag{2.5}$$

Matérn covariance function

$$C(d) = \begin{cases} \sigma^2 \frac{(\sqrt{2\nu} d\phi)^{\nu}}{2^{\nu-1}\Gamma(\nu)} K_{\nu}(\sqrt{2\nu} d\phi) & \text{if } d > 0, \\ \tau^2 + \sigma^2 & \text{if } d = 0. \end{cases} \tag{2.6}$$

The explicit derivations of the covariance functions (2.5) and (2.6) are provided in Appendix A. Covariance functions consist of two components: the variance σ^2 (also referred to as the partial sill) and the correlation function $\rho(\cdot)$. For example, it is clear from (2.5) that the correlation function $\rho(\cdot) = \exp(-\phi d)$ for $d > 0$. Recall that the effective range is the distance at which spatial autocorrelation becomes negligible, often specified as $\rho(\cdot) = 0.05$ in practical applications. From the correlation function of the exponential covariance function (2.5), the effective range is derived by solving for d in $\exp(-\phi d) = 0.05$, which is approximately $3/\phi$.

The exponential covariance function is a popular choice for the covariance function. Its appeal lies in its simplicity, as it only requires the specification of the spatial decay parameter ϕ. Additionally, the exponential covariance function offers desirable attributes such as ease of computation and straightforward interpretation of the effective range.

The exponential covariance function (2.5) is a special case of the Matérn covariance function (2.6) when the smoothness parameter $\nu = 1/2$. Abramowitz and Stegun demonstrated that the Matérn covariance function can be expressed as the product of an exponential component and a polynomial component of order p

$$C_{p+1/2}(d) = \sigma^2 \exp\big(-\sqrt{2p+1}d\phi\big)\frac{p!}{(2p)!}\sum_{i=0}^{p}\frac{(p+i)!}{i!(p-i)!}\big(2\sqrt{2p+1}d\phi\big)^{p-i} \tag{2.7}$$

for $d > 0$ (Abramowitz and Stegun, 1948). In (2.7), the mathematical symbol ! denotes the factorial operator. When $p = 0$, (2.7) is equal to the exponential covariance function (2.5).

Other notable special cases of the Matérn covariance function include $\nu = 3/2$ and the limit $\nu \to \infty$. We can use (2.7) to derive the Matérn covariance function with $\nu = 3/2$ by specifying $p = 1$

$$C_{3/2}(d) = \sigma^2\big(1 + \sqrt{3}d\phi\big)\exp\big(-\sqrt{3}d\phi\big).$$

When $\nu \to \infty$, the Matérn covariance function converges to a Gaussian covariance function (Banerjee et al., 2014) given as

$$C(d) = \sigma^2 \exp(-d^2\phi^2).$$

While the exponential and Matérn covariance functions are prominent choices, various alternative covariance functions are available. Banerjee et al. (2014) provide a comprehensive table of these available covariance functions.

2.4 Modelling point referenced spatial data

Estimating unknown parameters, such as the regression coefficient $\boldsymbol{\beta}$, is a common task in statistical modelling. In the non-Bayesian framework, $\boldsymbol{\beta}$ can be derived using ordinary least squares, a method that minimises the sum of square of the differences between the observations and the output of the model. Within the Bayesian framework, $\boldsymbol{\beta}$ is inferred through Bayes' theorem, which incorporates prior beliefs about $\boldsymbol{\beta}$, and yields the posterior distribution of $\boldsymbol{\beta}$. This posterior distribution encapsulates the uncertainty associated with $\boldsymbol{\beta}$. Explicit modelling within the Bayesian framework allows us to make inference on the parameters of interest and evaluate the uncertainty associated with any individual inferential statement (Sahu, 2022).

2.4.1 Kriging

A primary objective of point referenced data analyses is making predictions at an unobserved or new location, denoted as $\mathbf{s}_0$. In the non-Bayesian framework, spatial prediction using the minimum mean-squared error approach is called "kriging", named after South African statistician and mining engineer Danie G. Krige for his contribution to the development of empirical methods for geostatistical data (Matheron, 1963).

Let us denote the random variable at $\mathbf{s}_0$ as $Y(\mathbf{s}_0)$. Kriging involves minimising $E\big(Y(\mathbf{s}_0) - \big(\sum_{i=1}^{n} l_i Y(\mathbf{s}_i) + \delta_0\big)\big)^2$, where δ_0 denotes a minimal error and l_i denotes a weight parameter. Notably, the term $\sum_{i=1}^{n} l_i Y(\mathbf{s}_i)$ incorporates distinct weights for each observation, with observations in closer proximity to $\mathbf{s}_0$ assigned higher weights compared to those located farther away. Banerjee et al. (2014) demonstrated that by leveraging the intrinsic stationarity property described in Section 2.3.1 and applying substitutions, the function to be minimised can be reformulated as $-\sum_{i=0}^{n}\sum_{j=0}^{n} a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j)$, where $a_0 = 1$, $a_i = -l_i$ and $\gamma(\cdot)$ is the semivariogram function (2.2) introduced in Section 2.3.3. As a result, kriging requires prior knowledge of the parameters associated with the semivariogram function, such as the spatial decay parameter ϕ in (2.3) or the smoothness parameter ν in (2.4). The optimal values for l_i can be obtained through the solution of a constrained optimisation problem using Lagrange multipliers. Further derivations by Banerjee et al. revealed that $\sum_{i=1}^{n} l_i Y(\mathbf{s}_i)$ is the best linear unbiased predictor, and the uncertainty in the prediction is the predictive mean squared error calculated from $E\big(Y(\mathbf{s}_0) - \sum_{i=1}^{n} Y(\mathbf{s}_i)\big)^2$ (Banerjee et al., 2014).

Another strategy to determine $Y(\mathbf{s}_0)$ is by evaluating the conditional mean of $Y(\mathbf{s}_0)$. We begin by defining the following. Suppose there are n sites, we define $\mathbf{Y} = \big(Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n)\big)'$ as the vector of random variables at the sites and define $\mathbf{y} = \big(y(\mathbf{s}_1), y(\mathbf{s}_2), \ldots, y(\mathbf{s}_n)\big)'$ as the vector of observations at the sites. From multivariate normal theory, we have

$$\begin{pmatrix} Y(\mathbf{s}_0) \\ \mathbf{Y} \end{pmatrix} \sim N_{n+1}\left(\begin{pmatrix} \mu_0 \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix} \right),$$

where $N_{n+1}(\cdot)$ denotes an $(n+1)$-dimensional multivariate normal distribution, $\boldsymbol{\mu}$ is a vector of means of $\mathbf{Y}$ and μ_0 is the mean of $Y(\mathbf{s}_0)$. From the distribution above, we have

$$Y(\mathbf{s}_0)|\mathbf{Y} \sim N(\bar{\mu}, \bar{\Sigma}),$$

where

$$\bar{\mu} = E\big(Y(\mathbf{s}_0)|\mathbf{Y} = \mathbf{y}\big) = \mu_0 + \Omega_{12}\Omega_{22}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \tag{2.8}$$

$$\bar{\Sigma} = \mathrm{Var}\big(Y(\mathbf{s}_0)|\mathbf{Y} = \mathbf{y}\big) = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21} \tag{2.9}$$

(Eaton, 1983). Equation (2.8) is the conditional mean of $Y(\mathbf{s}_0)$ given $\mathbf{Y} = \mathbf{y}$, and Equation (2.9) is the conditional variance. The elements within the covariance matrix are given as follows. Ω_{11} denotes the variance of $Y(\mathbf{s}_0)$ and Ω_{22} denotes the $n \times n$ covariance matrix of $\mathbf{Y}$ with elements calculated from (2.5), (2.6) or any other covariance function. Ω_{12} denotes an n-dimensional row vector with elements given by the covariance between $Y(\mathbf{s}_0)$ and $Y(\mathbf{s}_i)$ for $i = 1, \ldots, n$, and Ω_{21} denotes an n-dimensional column vector and is the transpose of Ω_{12}.

Kriging is a non-Bayesian approach for spatial prediction. Given that the focus of this thesis is on Bayesian approaches, we will not further discuss kriging.

2.4.2 Bayes' theorem and Bayesian inference

Suppose θ is an unknown parameter that we want to estimate, and we have observations $\mathbf{y} = (y_1, \ldots, y_n)'$ for random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. We can utilise the Bayes' theorem

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}, \tag{2.10}$$

where $p(\cdot)$ denotes the probability density function (PDF) for continuous random variables or the probability mass function (PMF) for discrete random variables. In (2.10), the component $p(\mathbf{y}|\theta)$ is known as the likelihood function, the component $p(\theta)$ is known as the prior distribution and the component $p(\theta|\mathbf{y})$ is known as the posterior distribution. The prior distribution represents the prior belief on the unknown parameter θ. The component $p(\mathbf{y})$ is called the marginal likelihood function and can be expanded as

$$p(\mathbf{y}) = \int_{-\infty}^{\infty} p(\mathbf{y}|\theta)p(\theta)d\theta.$$

Bayes' theorem (2.10) is often shown in the simpler form

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta).$$

This simplification is possible because $p(\mathbf{y})$ is free of the unknown parameter θ. As a result, we know that the posterior distribution is proportional to the likelihood function and the prior distribution up to some normalising-constant.

Using the posterior distribution $p(\theta|\mathbf{y})$, we can identify useful measures of centrality including the posterior mean, the posterior median and the posterior mode. They are given as

$$\begin{aligned}
&\hat{\theta} = E(\theta|\mathbf{y}),\\
&\hat{\theta} : \int_{-\infty}^{\hat{\theta}} p(\theta|\mathbf{y})d\theta = 0.5,\\
&\hat{\theta} : p(\hat{\theta}|\mathbf{y}) = \text{arg max}_{\theta}\ p(\theta|\mathbf{y}),
\end{aligned}$$

respectively, where the arg $\max_\theta$ operator is used to find the θ that maximises the posterior distribution $p(\theta|\mathbf{y})$. The posterior distribution $p(\theta|\mathbf{y})$ also allows us to make probability statements about θ. For example, we can find the $\alpha/2$ and $(1-\alpha/2)$ quantiles of $p(\theta|\mathbf{y})$. Denoting these quantiles as q_L and q_U respectively, we have

$$\begin{aligned}
&\int_{-\infty}^{q_L} p(\theta|\mathbf{y})d\theta = \alpha/2,\\
&\int_{q_U}^{\infty} p(\theta|\mathbf{y})d\theta = (1-\alpha/2),
\end{aligned}$$

which implies that $P(q_L < \theta < q_U|\mathbf{y}) = 1-\alpha$. This interval is a $100\times(1-\alpha)\%$ credible set. It should be noted that this interval is symmetric about the mode for symmetric unimodal posterior distributions. For posterior distributions that are not symmetric or unimodal, it is better to take values of θ that have high posterior density greater than some cutoff for the interval. The resulting interval is called the highest posterior density (HPD) interval or the highest density interval (Lambert, 2018). Formally, a $100 \times (1-\alpha)\%$ HPD region for θ is a subset $C \in \Theta$ defined by

$$C = \{\theta : p(\theta|\mathbf{y}) \geq k\},$$

where k is the largest number possible while satisfying

$$\int_{\theta : p(\theta|\mathbf{y}) \geq k} p(\theta|\mathbf{y}) d\theta = 1 - \alpha.$$

For example, if the posterior distribution is multimodal, the HPD interval may be a discontinuous set.

Analyses within the Bayesian framework often require solving integration problems that are typically intractable in closed-form, even when the likelihood function and the prior distribution have closed-form expressions. Although analytical solutions are available, they are often limited to trivial cases, such as when the prior distribution is conjugate with the likelihood function. Markov chain Monte Carlo (MCMC) methods have traditionally been the preferred approach to address this challenge. Instead of attempting to solve integration problems analytically, MCMC methods rely on simulation-based approaches. MCMC methods involve generating samples of θ from a convergent Markov chain, where the stationary distribution corresponds to the posterior distribution $p(\theta|\mathbf{y})$. The generated samples of θ can provide useful information about the unknown parameter of interest. For example, the mean of these samples provides an estimate of the posterior mean, while the variance provides an estimate of the variance of the posterior distribution.

In practical applications, we often want to estimate more than a single unknown parameter. We can represent the unknown parameters as a vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ with up to p unknown parameters. Examples of $\boldsymbol{\theta}$ include the regression coefficients $\boldsymbol{\beta}$ and other latent variables within hierarchical models. Adapting $\boldsymbol{\theta}$ to (2.10), we have

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) d\theta_1 \ldots d\theta_p}.$$

This expression can be extended to incorporate hyperparameters and their respective prior distributions, referred to as hyperprior distributions. A hierarchical structure emerges in Bayesian models when we include hyperparameters and hyperprior distributions (Green et al., 2020).

2.4.3 Hierarchical Bayesian models

Hierarchical Bayesian models provide a coherent framework to model stochastic processes. They allow the incorporation of prior knowledge and properly account for uncertainties at different levels of the model. Hierarchical models can account for within-group similarity and allow for difference between groups for such datasets. Hierarchical models are models with multiple layers with different specifications to account for different sources of variations. When the parameter has many components, it may be useful to specify their joint prior distribution using a common hyper parameter. Hierarchical models are useful as they are flexible and representative of practical complex problems. Berliner (1996) suggested that a hierarchical model can be thought of in three stages:

$$\begin{aligned}&\text{Stage 1.}\ \ [\text{data} \mid \text{process, parameters}]\\ &\text{Stage 2.}\ \ [\text{process} \mid \text{parameters}]\\ &\text{Stage 3.}\ \ [\text{parameters}].\end{aligned}$$

Although the original context of this specification was for Bayesian models of time-series data, it was later adapted for the context of Bayesian models of point referenced spatial data (Gelfand, 2012, Blangiardo and Cameletti, 2015, Sahu, 2022). In the first stage, we have the distribution of the observations. The first stage describes the structure of the conditional distribution of the data, given the underlying process and any parameters of the model. The second stage describes the structure of the underlying process. The spatial dependence structure of interest is primarily modelled in the second stage. Gaussian processes are commonly used in this stage to model the latent spatial structure in the observable data. The third stage sets the priors distributions for the parameters from the first and second stage. The priors distributions specified in a hierarchical model can be vague or even improper since the data are often sufficiently informative. The posterior distribution of the parameter from the first and second stage becomes narrow despite the vagueness of the prior distributions. The posterior distribution would often change very little even if a more specific prior distribution was defined for the hyperparameters (Neal, 1996).

To illustrate the hierarchical structure of Bayesian models of point referenced data, consider the following

$$Y(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \omega(\mathbf{s}_i) + \epsilon(\mathbf{s}_i),$$

where $Y(\mathbf{s}_i)$ is the random variable Y at site $\mathbf{s}_i$ for $i = 1, \ldots, n$, $\mathbf{x}(\mathbf{s}_i)$ are the covariates at site $\mathbf{s}_i$ and $\boldsymbol{\beta}$ is a vector of regression coefficient. The component $\omega(\mathbf{s}_i)$ is the spatial random effect that captures spatial dependence among the data and the component $\epsilon(\mathbf{s}_i)$ is the unstructured random effect. Let us also assume

$$\begin{aligned} \boldsymbol{\omega} &\sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}), \\ \boldsymbol{\epsilon} &\sim N_n(\mathbf{0}, \tau^2 I_n) \end{aligned}$$

for $\boldsymbol{\omega} = \big(\omega(\mathbf{s}_1), \ldots, \omega(\mathbf{s}_n)\big)'$ and $\boldsymbol{\epsilon} = \big(\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_n)\big)'$, where I_n denotes an $n \times n$ identity matrix and $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ denotes an $n \times n$ covariance matrix with elements calculated from some covariance function, such as (2.5) or (2.6). This model is equivalently expressed as

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\omega}, \boldsymbol{\theta} &\sim N_n(X\boldsymbol{\beta} + \boldsymbol{\omega}, \tau^2 I_n), \\ \boldsymbol{\omega}|\boldsymbol{\psi} &\sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}), \end{aligned}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2, \boldsymbol{\psi})'$ and $\boldsymbol{\psi}$ is a vector of parameters required by the covariance function. For example, if the spatial random effect follows a zero-mean n-dimensional multivariate normal distribution with elements of its covariance matrix calculated from the exponential covariance function (2.5), then $\boldsymbol{\psi}$ will include the spatial variance $\sigma^2_{\boldsymbol{\omega}}$ and the spatial decay parameter ϕ, i.e., $\boldsymbol{\psi} = (\sigma^2_{\boldsymbol{\omega}}, \phi)'$. If the elements of the covariance matrix is calculated from a Matérn covariance function (2.6), then $\boldsymbol{\psi}$ will include the the spatial variance $\sigma^2_{\boldsymbol{\omega}}$, the spatial decay parameter ϕ, and the smoothness parameter ν, i.e., $\boldsymbol{\psi} = (\sigma^2_{\boldsymbol{\omega}}, \phi, \nu)'$. When we express our example model this way, the three-staged specification becomes clear

$$\begin{aligned} &\text{Stage 1. } \mathbf{Y}|\boldsymbol{\omega}, \boldsymbol{\theta} \sim N_n(X\boldsymbol{\beta} + \boldsymbol{\omega}, \tau^2 I_n), \\ &\text{Stage 2. } \boldsymbol{\omega}|\boldsymbol{\psi} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}), \\ &\text{Stage 3. Priors distributions on } \boldsymbol{\theta}. \end{aligned}$$

Bayesian hierarchical modelling provides a natural framework to properly assess the uncertainty in the parameter estimates and spatial predictions. Bayesian inference allows us to make probability statements about parameters of interest, including the uncertainty arising from the model (Gelfand, 2012, Bass, 2015). The hierarchical estimates also display a behaviour called “shrinkage towards the mean” where estimates with the most extreme values are shifted the most since using hierarchical models takes the probability mass away from the outlier estimates (Lambert, 2018). Although hierarchical models offer a high degree of flexibility and are well-suited for capturing the intricacies of real-world problems, they present challenges for parameter estimation. Hierarchical models are often too complex for exact inference. Consequently, the most effective approach for parameter estimation for Bayesian hierarchical models is utilising Markov chain Monte Carlo (MCMC) techniques (Robert and Casella, 2004, Lambert, 2018).

As a remark, we may flatten our example hierarchical model above through suitable marginalisation and integration such that

$$\mathbf{Y}|\boldsymbol{\theta} \sim N_n(X\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}} + \tau^2 I_n).$$

Although fitting marginal models using MCMC methods is usually computationally better behaved (Gelfand, 2012), the models specified in the hierarchical structure have the advantage of interpretability.

2.5 Markov chain Monte Carlo methods

2.5.1 Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm (Metropolis et al., 1953, Hastings, 1970) is a Markov chain Monte Carlo (MCMC) method. It is used to generate a sequence of samples that approximate the posterior distribution of the unknown parameters of interest. These samples enable inferences about the parameters, which is a primary goal of Bayesian modelling. To facilitate the description of the MH algorithm, let θ denote an unknown parameter of interest, and let $\mathbf{y} = (y_1, \dots, y_n)'$ represent the observations for the random variables $\mathbf{Y} = (Y_1, \dots, Y_n)'$. For sampling iterations $s = 1, \dots, S$, where S denotes the total number of sampling iterations, the MH algorithm follows these steps:

Step 0. Initialise

Select an initial sample point $\theta^{(0)}$ either randomly or based on prior information. Prior information can include estimates from previous studies or a prior distribution.

Another approach is to use simple models, such as a linear regression model, to provide initial estimates for parameters in a more complex Bayesian model. Choosing $\theta^{(0)}$ wisely is crucial for the efficiency of the algorithm, as a poorly chosen initial point may lead to slower convergence. For instance, if $\theta^{(0)}$ is far from the high probability regions of the posterior distribution, the algorithm will require more iterations to converge to the target distribution.

Step 1. Propose

Generate a candidate sample point ϕ from a proposal distribution $g(\phi|\theta^{(s-1)})$. The proposal distribution should be easy to sample from, and ideally, one that closely resembles the posterior distribution to improve convergence towards the target distribution. An example is a normal distribution centered at $\theta^{(s-1)}$, that is $\phi|\theta^{(s-1)} \sim N(\theta^{(s-1)}, \sigma^2)$, where σ^2 is a predefined variance. It is important to tune σ^2 appropriately. If σ^2 is too large, the acceptance rate will be too low, leading to slow convergence. If it is too small, the chain will make small moves, also resulting in slow convergence.

Step 2. Evaluate

Compute the acceptance probability, defined as

$$\alpha = \min\left(1, \frac{p(\phi|\mathbf{y})}{p(\theta^{(s-1)}|\mathbf{y})}\frac{g(\theta^{(s-1)}|\phi)}{g(\phi|\theta^{(s-1)})}\right).$$

Formally, the acceptance probability is denoted as $A\big(\phi, \theta^{(s-1)}\big)$ to explicitly show that it depends on both ϕ and $\theta^{(s-1)}$. For compact notation, we use α to represent the acceptance probability. The ratio within the acceptance probability is generally called the Metropolis ratio, though it is sometimes referred to as the Hastings ratio (Dunn and Shultis, 2022). In the Metropolis ratio, $p(\phi|\mathbf{y})$ denotes the posterior distribution evaluated at ϕ, and $p(\theta^{(s-1)}|\mathbf{y})$ denotes the posterior distribution evaluated at $\theta^{(s-1)}$. Importantly, this ratio does not depend on the normalising constant, and avoids the need for integration as in Bayes' theorem. The terms $g(\theta^{(s-1)}|\phi)$ and $g(\phi|\theta^{(s-1)})$ represent the proposal distributions. $g(\theta^{(s-1)}|\phi)$ is the probability of choosing $\theta^{(s-1)}$ given the current state of the Markov chain is ϕ, and $g(\phi|\theta^{(s-1)})$ is the probability of generating ϕ given $\theta^{(s-1)}$, as described in Step 1.

Step 3. Accept or reject
Independently sample u from a uniform distribution, $u \sim U(0, 1)$. If $\alpha \geq u$, accept the candidate sample point and set $\theta^{(s)} = \phi$. Conversely, if $\alpha < u$, reject the candidate sample point and set $\theta^{(s)} = \theta^{(s-1)}$.

The MH algorithm repeats Steps 1-3 for a prespecified number of iterations S. Although there are no definitive guidelines for determining the appropriate value of S, it is essential to ensure that the Markov chain has converged and exhibits good mixing. Good mixing implies that the Markov chain explores the target distribution comprehensively and samples the entire posterior distribution effectively. This means the chain does not remain in local modes for extended periods and can transition between different regions of the state space efficiently (Robert et al., 2010, Lambert, 2018). When convergence is achieved, the Markov chain reaches its stationary distribution, which corresponds to the posterior distribution $p(\theta|\mathbf{y})$.

It is also important to monitor the acceptance rate of the algorithm. An optimal acceptance rate facilitates efficient sampling and faster convergence. If the acceptance rate is excessively low or high, the parameters of the proposal distribution should be adjusted accordingly. Asymptotically, the optimal acceptance rate is 0.234 (Gelman et al., 1996), although in practice, the ideal acceptance rate typically ranges between 20% and 40% (Lambert, 2018).

As mentioned in Step 0 of the algorithm, $\theta^{(0)}$ may influence the early iterations of the Markov chain, causing them to be unrepresentative of the target posterior distribution. Burn-in refers to the initial phase of the Markov chain during which these samples are discarded. This process helps reduce bias and ensures that the remaining samples are drawn from the stationary distribution, providing a more accurate representation of the target distribution. Determining the appropriate length of the burn-in period often requires visual inspection of the trace plots to assess when the chain has stabilised.

Successive samples drawn from the MH algorithm may be autocorrelated, leading to greater uncertainties in the parameters of interest. The Effective Sample Size (ESS) is a metric used to determine the number of independent samples obtained from the chain. A higher ESS indicates better mixing and more independent information about the target distribution, while a lower ESS suggests high autocorrelation, fewer independent samples, and potentially poorer mixing. ESS also helps assess whether the number of iterations S is sufficient. If the ESS is low relative to S, it indicates that more iterations are needed to obtain a sufficient number of independent samples. Another way to reduce autocorrelation between successive samples is through thinning. Thinning involves selecting every k-th sample from the Markov chain and discarding the rest, with the value of k chosen based on the autocorrelation of the chain.

2.5.2 Gibbs sampler

The Gibbs sampler (Casella and George, 1992) is an MCMC method that, in its most basic form, is a special case of the MH algorithm where all proposals are accepted. The Gibbs sampler is particularly useful when sampling directly from the joint distribution is challenging, but sampling from the conditional distributions is more feasible. To describe the Gibbs sampling algorithm, let $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)'$ denote a vector of unknown parameters of interest, and let $\mathbf{y} = (y_1, \ldots, y_n)'$ represent the observed data for the random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. For sampling iterations $s = 1, \ldots, S$, the Gibbs sampler follows these steps:

Step 0. Initialise
Select initial sample points $\boldsymbol{\theta}^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \theta_4^{(0)}\right)'$ either randomly or based on prior information. Prior information can include estimates from previous studies or a prior distribution. The choice of $\boldsymbol{\theta}^{(0)}$ may influence the efficiency of the algorithm. For example, if $\boldsymbol{\theta}^{(0)}$ is far from the high probability regions of the posterior distribution, the algorithm will require more iterations to converge to the target distribution.

Step 1. Sample
Update each parameter sequentially by sampling from its conditional distribution given the current values of the other parameters and the observed data $\mathbf{y}$

$$\begin{aligned}
\theta_1^{(s)} &\sim p(\theta_1 | \theta_2^{(s-1)}, \theta_3^{(s-1)}, \theta_4^{(s-1)}, \mathbf{y}), \\
\theta_2^{(s)} &\sim p(\theta_2 | \theta_1^{(s)}, \theta_3^{(s-1)}, \theta_4^{(s-1)}, \mathbf{y}), \\
\theta_3^{(s)} &\sim p(\theta_3 | \theta_1^{(s)}, \theta_2^{(s)}, \theta_4^{(s-1)}, \mathbf{y}), \\
\theta_4^{(s)} &\sim p(\theta_4 | \theta_1^{(s)}, \theta_2^{(s)}, \theta_3^{(s)}, \mathbf{y}).
\end{aligned}$$

The Gibbs sampler repeats step 1 for a prespecified number of iterations S or until the Markov chain has reached convergence. When convergence is achieved, the Markov chain reaches its stationary distribution, which corresponds to the joint posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. In practical applications, the procedure outlined above can be generalised to any number of unknown parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$.

Notice that unlike the MH algorithm discussed in Section 2.5.1, where an acceptance probability is calculated to determine whether a proposed candidate sample point is accepted or rejected, the Gibbs Sampler "accepts" all proposals. In the Gibbs Sampler, each proposed candidate is immediately used to sample other parameters. For instance, in Step 1, after generating $\theta_1^{(s)}$, this value is immediately used to generate $\theta_2^{(s)}$. The proposal distribution in Gibbs Sampling is implicitly defined by the conditional posterior distribution of each parameter given the current values of the other parameters. When a parameter is updated, a new value is drawn directly from its conditional posterior distribution, which effectively serves as the proposal distribution for that parameter during that step. This ensures that each step in the Gibbs Sampler always produces valid samples from the target distribution without the need for an acceptance-rejection mechanism (Lambert, 2018).

The Gibbs sampler in practical applications comes with several challenges. One of the primary difficulties is deriving the conditional distributions. The Gibbs sampler relies on the fact that these conditional distributions can be sampled more easily than the joint distribution. However, in many cases, closed-form expressions for these conditional distributions are not available, making them mathematically complex or computationally expensive to determine. Additionally, high-dimensional parameter spaces require a large number of sampling iterations for the algorithm to effectively explore the parameter space, leading to greater computational demand and slower convergence. Highly correlated unknown parameters can also result in slow mixing in the Gibbs sampling algorithm, as discussed by Roberts and Sahu (1997). Despite these challenges, the Gibbs sampler remains a valuable tool for Bayesian modeling and inference. It can be implemented in statistical software such as JAGS (Just Another Gibbs Sampler) (Plummer et al., 2003) and WinBUGS (Lunn et al., 2000), which are based on the BUGS (Bayesian inference Using Gibbs Sampling) project (Gilks et al., 1994, Spiegelhalter et al., 1996).

2.5.3 Hamiltonian Monte Carlo and Stan

The Hamiltonian Monte Carlo (HMC) algorithm is an MCMC method that efficiently explores parameter spaces by exploiting the local geometric information of the posterior distribution, allowing for unconstrained movements. Originally developed by Duane et al. (1987) under the name “Hybrid Monte Carlo” for applications in quantum chromodynamics within theoretical physics, the algorithm found its first statistical application in neural network models through the work of Neal (1996). It was later renamed “Hamiltonian Monte Carlo”, as documented in a review by Neal et al. (2011). The HMC algorithm effectively explores the parameter space by leveraging the local features of the posterior distribution, such as its “peaks” and “troughs”, to gain a deeper understanding of the posterior distribution. The HMC algorithm can be implemented in the probabilistic programming language Stan (Stan Development Team, 2023).

The operation of the HMC algorithm is analogous to an object sliding on a frictionless surface with varying elevations (Neal et al., 2011, Lambert, 2018). On this surface, there are peaks and troughs. When the object is pushed off a peak, it descends into a trough. However, because the object has momentum, it has enough energy to ascend to another peak. Without momentum, the object would only descend. Momentum allows the object to explore both peaks and troughs efficiently. Additionally, the effect of gravity causes the object to spend less time exploring the peaks and more time exploring the troughs.

In this analogy, the frictionless surface corresponds to the negative logarithm of the posterior distribution (NLP), which is the inverse of the posterior distribution. The peaks on the surface represent areas of low probability in the posterior distribution, while the troughs represent areas of high probability. The concept of the object spending less time exploring peaks and more time in troughs is advantageous for the algorithm, as it implies that more time is allocated to regions of high posterior density and less to regions of low posterior density. For example, a multimodal posterior distribution presents a challenge due to its multiple regions of high posterior density separated by low-probability areas. This complexity makes efficient exploration difficult. By incorporating momentum, the HMC algorithm can efficiently escape from one high-density region to another, effectively exploring complex posterior distributions.

In the HMC algorithm, “Hamiltonian” refers to the concept of total energy in a system from physics. It is denoted as $H(\theta, m)$, a function of the parameters θ and momentum m. Specifically, the Hamiltonian comprises two components: potential energy, which depends on the location of θ within the parameter space, and kinetic energy, which depends on the momentum m. In the context of HMC algorithm, the potential energy, denoted as $U(\theta)$, corresponds to the NLP. The kinetic energy, denoted as $KE(m)$, is proportional to the square of the momentum.

$$\begin{aligned} H(\theta, m) &= U(\theta) + KE(m), \\ &= -\log\big(p(\theta|\mathbf{y})\big) + \frac{m^2}{2}. \end{aligned} \tag{2.11}$$

As the posterior density $p(\theta|\mathbf{y})$ increases, the potential energy $U(\theta)$ decreases (becomes more negative) because there is less gravitational potential energy available to be converted to kinetic energy in areas of high posterior density (i.e., the troughs). Conversely, in areas of low posterior density (i.e., the peaks), a gentle push will initiate a rapid descent. The HMC algorithm samples from the joint distribution of θ and m, which has the PDF

$$p(\theta, m|\mathbf{y}) \propto \exp\big(-H(\theta, m)\big).$$

The objective of the HMC algorithm is to fully explore the NLP landscape while avoid getting stuck in areas of high posterior density, which can lead to bias results in the posterior distribution. For sampling iterations $s = 1, \ldots, S$, the HMC algorithm follow the steps:

Step 0. Initialise

Select an initial sample point $\theta^{(0)}$ either randomly or based on prior information. Prior information can include estimates from previous studies or a prior distribution.

Step 1. Sample momentum

Sample the momentum m from a multivariate normal distribution with mean zero and identity covariance matrix, $m \sim \mathcal{N}(0, I)$.

Step 2. Propose

Propose a candidate sample point $\tilde{\theta}$ and a new momentum m^* by simulating the

Hamiltonian dynamics over a period T using the leapfrog integrator. In this step, let $q^{(0)} = \theta^{(s-1)}$ and $m^{(0)} = m$. For leapfrog steps $l = 0, \ldots, L-1$, the leapfrog integrator follows the steps:

Step 2.1. Half step update of the momentum

Using the current momentum $m^{(l)}$, step size ϵ and gradient of the potential energy with respect to the current sample point $\nabla U(q^{(l)})$, update the momentum by half a step

$$m^{(l+\epsilon/2)} = m^{(l)} - \frac{\epsilon}{2} \nabla U(q^{(l)}).$$

Step 2.2. Full step update of the location

Using the current sample point $q^{(l)}$, half step updated momentum $m^{(l+\epsilon/2)}$ from Step 2.1, and step size ϵ, update the sample point by one step

$$q^{(l+\epsilon)} = q^{(l)} + \epsilon m^{(l+\epsilon/2)}.$$

Step 2.3. Another half step update of the momentum

Using the half step updated momentum $m^{(l+\epsilon/2)}$ from Step 2.1, step size ϵ and the gradient of the potential energy with respect to the full step updated sample point $q^{(l+\epsilon)}$ from Step 2.2, update the momentum by another half a step

$$m^{(l+\epsilon)} = m^{(l+\epsilon/2)} - \frac{\epsilon}{2} \nabla U\big(q^{(l+\epsilon)}\big).$$

Repeat Steps 2.1-2.3 for $L-1$ leapfrog steps, where L is the number of steps determined by the total duration divided by the step size $L = T/\epsilon$, then obtain the candidate sample point $\tilde{\theta} = q^{(L)}$ and the new momentum $m^* = m^{(L)}$.

Step 3. Evaluate

Compute the acceptance probability

$$\alpha = \min \left(1,\ \frac{\exp\big(-H(\tilde{\theta}, m^*)\big)}{\exp\big(-H(\theta^{(s-1)}, m)\big)}\right),$$

where $\theta^{(s-1)}$ denotes the current sample point, $\tilde{\theta}$ denotes the candidate sample point calculated from the previous step, m denotes the current momentum and m^* denotes the new momentum calculated from the previous step. $H(\cdot)$ is the Hamiltonian given in Equation (2.11).

Step 4. Accept or reject
Independently sample u from a uniform distribution, $u \sim U(0,1)$. If $\alpha \geq u$, accept the new location and set $\theta^{(s)} = \tilde{\theta}$. Otherwise, return to the initial location and set $\theta^{(s)} = \theta^{(s-1)}$.

The HMC algorithm repeats Steps 1-4 for a prespecified S number of sampling iterations or when the Markov chain has reached convergence. When convergence has been achieved, the algorithm provides samples that approximate the joint posterior distribution.

The choice of the total duration T and step size ϵ in the leapfrog integrator is crucial for the convergence of the HMC algorithm. A T that is too short leads to slower exploration, while a T that is too long can cause the algorithm to retrace its path, hindering exploration. The No U-Turn Sampler (NUTS) developed by Hoffman et al. (2014) addresses this by stopping when a U-turn is detected. Shorter T values are recommended for highly curved posterior distributions, while longer T values are suitable for flatter ones (Lambert, 2018). Similarly, an ϵ that is too large results in a low acceptance rate, while an ϵ that is too small wastes computation time and leads to slow exploration. For practical implementations, Neal et al. (2011) suggested randomly choosing ϵ from a distribution, using adaptive tuning methods to achieve an optimal acceptance rate of approximately 0.65, and monitoring diagnostics like trace plots.

2.6 The INLA method

The integrated nested Laplace approximation (INLA) method (Rue et al., 2009) offers a powerful alternative to traditional MCMC methods for Bayesian inference. Unlike MCMC, which relies on iterative sampling, the INLA method employs a combination of analytical approximations and numerical integration schemes (Martino and Riebler, 2019). This approach reduces computational burdens, particularly when dealing with large datasets, and helps avoid potential convergence issues often encountered with MCMC methods (Blangiardo and Cameletti, 2015, Moraga, 2019). By leveraging these techniques, INLA provides an efficient and scalable solution for approximating posterior distributions in Bayesian hierarchical models, making it especially useful in applications such as spatial and spatio-temporal modelling.

Recall from Section 2.4.3 that one of the primary goals of a Bayesian hierarchical model for point referenced spatial data is to estimate the spatial random effects $\boldsymbol{\omega} = \big(\omega(\mathbf{s}_1), \ldots, \omega(\mathbf{s}_n)\big)'$, and the unknown parameters $\boldsymbol{\theta}$. For simplicity, let ω_i denote the spatial random effects, such that $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)'$ in this section. In this context, the objective is to estimate

$$p(\omega_i|\mathbf{y}) = \int_{-\infty}^{\infty} p(\omega_i|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \tag{2.12}$$

$$p(\theta_k|\mathbf{y}) = \int_{-\infty}^{\infty} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-k}, \tag{2.13}$$

for $i = 1, \ldots, n$ and $k = 1, \ldots, p$, where p is the number of unknown parameters. The parameter $\boldsymbol{\theta}_{-k}$ denotes the vector of unknown parameters without θ_k.

To achieve the objective, we may consider evaluating the approximations of $p(\omega_i|\mathbf{y})$ and $p(\theta_k|\mathbf{y})$, denoted as $\tilde{p}(\omega_i|\mathbf{y})$ and $\tilde{p}(\theta_k|\mathbf{y})$ respectively. With the approximations, we can re-express (2.12) and (2.13) as

$$\tilde{p}(\omega_i|\mathbf{y}) = \int_{-\infty}^{\infty} \tilde{p}(\omega_i|\boldsymbol{\theta}, \mathbf{y}) \tilde{p}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta},$$

$$\tilde{p}(\theta_k|\mathbf{y}) = \int_{-\infty}^{\infty} \tilde{p}(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}_{-k},$$

where the integrations with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_{-k}$ are performed using numerical integration methods. To obtain $\tilde{p}(\boldsymbol{\theta}|\mathbf{y})$,

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto \left. \frac{p(\mathbf{y}|\boldsymbol{\omega}, \boldsymbol{\theta}) p(\boldsymbol{\omega}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{\tilde{p}_G(\boldsymbol{\omega}|\boldsymbol{\theta}, \mathbf{y})} \right|_{\boldsymbol{\omega} = \boldsymbol{\omega}^*(\boldsymbol{\theta})} = \tilde{p}(\boldsymbol{\theta}|\mathbf{y}), \tag{2.14}$$

where $\tilde{p}_G(\boldsymbol{\omega}|\boldsymbol{\theta}, \mathbf{y})$ denotes the Gaussian approximation of $p(\boldsymbol{\omega}|\boldsymbol{\theta}, \mathbf{y})$, and $\boldsymbol{\omega}^*(\boldsymbol{\theta})$ is the mode of the full conditional distribution for a given $\boldsymbol{\theta}$. To obtain $\tilde{p}(\omega_i|\boldsymbol{\theta}, \mathbf{y})$,

$$p(\omega_i|\boldsymbol{\theta}, \mathbf{y}) \propto \left. \frac{p(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})}{\tilde{p}_G(\boldsymbol{\omega}|\boldsymbol{\theta}, \mathbf{y})} \right|_{\boldsymbol{\omega} = \boldsymbol{\omega}^*(\boldsymbol{\theta})} = \tilde{p}(\omega_i|\boldsymbol{\theta}, \mathbf{y}). \tag{2.15}$$

However, the Gaussian approximation $\tilde{p}(\boldsymbol{\omega}|\boldsymbol{\theta}, \mathbf{y})$ is often too strong and tends to give poor results since the conditional distributions are typically skewed or heavy tailed (Blangiardo and Cameletti, 2015).

Another approach to achieve the objective is to partition $\boldsymbol{\omega}$, such that $\boldsymbol{\omega} = (\omega_i, \boldsymbol{\omega}_{-i})'$, where $\boldsymbol{\omega}_{-i}$ denotes the vector $\boldsymbol{\omega}$ without ω_i, and then apply the approximation

$$p(\omega_i|\boldsymbol{\theta}, \mathbf{y}) \propto \left. \frac{p(\boldsymbol{\omega}, \boldsymbol{\theta}|\mathbf{y})}{\tilde{p}_G(\boldsymbol{\omega}_{-i}|\omega_i, \boldsymbol{\theta}, \mathbf{y})} \right|_{\boldsymbol{\omega}_{-i}=\boldsymbol{\omega}^*_{-i}(\omega_i,\boldsymbol{\theta})} = \tilde{p}(\omega_i|\boldsymbol{\theta}, \mathbf{y}), \tag{2.16}$$

where $\tilde{p}_G(\boldsymbol{\omega}_{-i}|\omega_i, \boldsymbol{\theta}, \mathbf{y})$ is the Gaussian approximation of $p(\boldsymbol{\omega}_{-i}|\omega_i, \boldsymbol{\theta}, \mathbf{y})$ and $\boldsymbol{\omega}^*_{-i}(\omega_i, \boldsymbol{\theta})$ is its mode. While this approach performs better compared to the previously described method, because $\tilde{p}_G(\boldsymbol{\omega}_{-i}|\omega_i, \boldsymbol{\theta}, \mathbf{y})$ is often close to normal, it is computationally expensive (Blangiardo and Cameletti, 2015).

The third approach to achieve the objective involves utilising the simplified Laplace approximation approach (Blangiardo and Cameletti, 2015). This method employs the Taylor series expansion up to the third order to approximate $p(\omega_i|\boldsymbol{\theta}, \mathbf{y})$. The Laplace approximation enables the estimation of the posterior distribution using normal distributions. In essence, the concept of the Laplace approximation is to approximate a well-behaved unimodal function with a Gaussian density. Let us demonstrate the Laplace approximation with the following examples.

Example 1. The Laplace approximation of a χ^2 *distribution*
The PDF of a χ^2 distribution with k degrees of freedom is

$$p(x; k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} \exp(-x/2),$$

where $\Gamma(\cdot)$ is the mathematical Gamma function. The Laplace approximation requires the log PDF, the first derivative of the log PDF and the second derivative of the log PDF

$$\begin{aligned} \log p(x; k) &= \left(\frac{k}{2} - 1\right) \log x - \frac{x}{2} - \log\left(2^{k/2}\Gamma(k/2)\right), \\ \frac{\partial}{\partial x} \log p(x; k) &= \frac{\frac{k}{2} - 1}{x} - \frac{1}{2}, \\ \frac{\partial^2}{\partial x^2} \log p(x; k) &= \frac{1 - \frac{k}{2}}{x^2}. \end{aligned}$$

The Laplace approximation of any distribution is a normal distribution with mean $\hat{x}$ and variance $\hat{\sigma}^2 = -1/\frac{\partial^2}{\partial x^2} \log p(\hat{x})$. The mean of the Laplace approximate normal distribution $\hat{x}$ is obtained by evaluating the first derivative at zero

$$\begin{aligned} \frac{\frac{k}{2}-1}{\hat{x}} - \frac{1}{2} &= 0, \\ \hat{x} &= k-2. \end{aligned}$$

The variance of the Laplace approximate normal distribution $\hat{\sigma}^2$ is calculated as

$$\begin{aligned} \hat{\sigma}^2 &= -1/\frac{\partial^2}{\partial x^2} \log p(\hat{x}), \\ &= \frac{\hat{x}^2}{\frac{k}{2}-1}, \\ &= \frac{\left(2(\frac{k}{2}-1)\right)^2}{\frac{k}{2}-1}, \\ &= \frac{4\left(\frac{k}{2}-1\right)^2}{\frac{k}{2}-1}, \\ &= 2(k-2). \end{aligned}$$

Hence, the simplified Laplace approximation of a χ^2 distribution with k degrees of freedom is

$$\chi_k^2 \overset{LA}{\sim} N(\hat{x} = k-2, \hat{\sigma}^2 = 2(k-2)),$$

where $\overset{LA}{\sim}$ denotes the simplified Laplace approximation.

Example 2. The Laplace approximation of a Gamma distribution
The log PDF of a Gamma distribution is given by

$$\begin{aligned} \log p(x; \alpha, \beta) &= \log\left(\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)\right), \\ &= (\alpha-1)\log x - \beta x + \alpha \log \beta - \log \Gamma(\alpha). \end{aligned}$$

The first and second derivative of the log PDF are

$$\begin{aligned}\frac{\partial}{\partial x}\log p(x;\alpha,\beta) &= \frac{(\alpha-1)}{x}-\beta,\\ \frac{\partial^2}{\partial x^2}\log p(x;\alpha,\beta) &= \frac{-(\alpha-1)}{x^2}.\end{aligned}$$

The mean of the Laplace approximate normal distribution $\hat{x}$ is obtained as follows

$$\begin{aligned}\frac{(\alpha-1)}{\hat{x}}-\beta = 0,&\\ \hat{x} &= \frac{\alpha-1}{\beta}.\end{aligned}$$

The variance of the Laplace approximate normal distribution $\hat{\sigma}^2$ is obtained as follows

$$\begin{aligned}\hat{\sigma}^2 &= -1/\frac{\partial^2}{\partial x^2}\log p(\hat{x}),\\ &= \frac{\hat{x}^2}{(\alpha-1)},\\ &= \frac{(\alpha-1)^2}{(\alpha-1)\boldsymbol{\beta}^2},\\ &= \frac{(\alpha-1)}{\boldsymbol{\beta}^2}.\end{aligned}$$

Hence, the Laplace approximation of the Gamma distribution is

$$\Gamma(\alpha,\beta) \overset{LA}{\sim} N\left(\hat{x} = \frac{\alpha-1}{\beta}, \hat{\sigma}^2 = \frac{\alpha-1}{\beta^2}\right).$$

In the examples above, the Laplace approximation works reasonably well for large k and α. The INLA approach can be implemented in R (R Core Team, 2021) through the `INLA` package (Rue et al., 2009, Martins et al., 2013). By default, the simplified Laplace approximation is used when implementing INLA in R. However, the other two approximation strategies are also available.

In summary, the INLA method first explores the marginal posterior distribution of the hyperparameters to locate the mode. It then generates a set of relevant points along with corresponding weights to approximate the distribution of the parameters of interest. Each marginal posterior can be obtained through interpolations based on these computed values.

2.6.1 The stochastic partial differential equation method

One of the major challenges in employing MCMC methods for Bayesian models of point referenced spatial data is handling large covariance matrices. The computational complexity of algebraic operations involving $n \times n$ dense covariance matrices is $O(n^3)$, making these methods impractical for large datasets. The stochastic partial differential equation (SPDE) method addresses this problem by providing a more computationally efficient way to represent Gaussian fields (GFs), particularly those with Matérn covariance structures. By transforming the problem into solving a differential equation, the SPDE method reduces computational complexity, providing a feasible approach for handling large spatial datasets.

A Gaussian Field (GF) can be thought of as a continuous surface where spatial data can occur at any point, making it challenging and computationally expensive to account for every single point. The SPDE method addresses this by describing the spatial process in terms of its local properties, which can then be used to efficiently reconstruct the global structure. This is achieved by partitioning the continuous surface into triangles, creating a Gaussian Markov Random Field (GMRF), also referred to as a mesh. This discretisation simplifies computation by focusing only on the vertices of the triangles, reducing the computational cost.

To facilitate the description of the SPDE method, consider the following model

$$\begin{aligned} Y(\mathbf{s}_i) &= \eta(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \\ \eta(\mathbf{s}_i) &= \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \omega(\mathbf{s}_i), \end{aligned}$$

where $Y(\mathbf{s}_i)$ represents the random variable at location $\mathbf{s}_i$ for $i = 1, \ldots, n$ and $\epsilon(\mathbf{s}_i)$ is the unstructured error term, assumed to be independently and identically normally distributed with mean zero and variance σ_ϵ^2. The linear predictor $\eta(\mathbf{s}_i)$ comprises the covariates $\mathbf{x}(\mathbf{s}_i)$, the regression coefficients $\boldsymbol{\beta}$ and the spatial random effect $\omega(\mathbf{s}_i)$, which is assumed to follow a zero-mean Gaussian process with a Matérn covariance. The SPDE method follows the steps:

Step 1. Define the SPDE

Formulate the SPDE to which the GF is a solution. Lindgren et al. demonstrated that a GF $\omega(\mathbf{s})$ with Matérn covariance is a solution to the linear fractional SPDE

$$(\kappa^2 - \Delta)^{\alpha/2}\omega(\mathbf{s}) = \mathcal{W}(\mathbf{s}),$$

where $(\kappa^2 - \Delta)^{\alpha/2}$ is a pseudo-differential operator with parameters κ and α. Here, κ controls the spatial range, α relates to the smoothness, Δ is the Laplacian operator, and $\mathcal{W}(\mathbf{s})$ is a spatial Gaussian white noise process (Lindgren et al., 2011).

Step 2. Discretise the spatial domain

To solve the SPDE numerically, the spatial domain is discretised using a mesh. This involves a process known as "triangularisation", where the continuous GF is partitioned into triangles, creating a discretised GMRF, also referred to as a mesh. The level of detail in the mesh directly impacts both the computational cost and the accuracy of the representation of the original GF. A finer mesh, with more triangles, offers a closer approximation to the original GF but increases computational demands. Conversely, a coarser mesh reduces computational cost but may result in a less accurate representation. Although there are guidelines for constructing the mesh (Krainski et al., 2018, Righetto et al., 2020), this process remains subjective. Generally, the mesh should have an outer boundary with larger triangles in the outer region and smaller triangles within the inner region. The precise influence of the mesh on model predictions is still not fully understood and requires further investigation.

Step 3. Link to the Gaussian field

The spatial random effect using the discretised GMRF is given by

$$\omega(\mathbf{s}) = \sum_{g=1}^{G} \varphi_g(\mathbf{s})\tilde{\omega}_g,$$

where G is the total number of vertices on the mesh, $\{\varphi_g\}$ is the set of basis function and $\{\tilde{\omega}_g\}$ are zero-mean Gaussian distributed weights. The basis functions are chosen to be piecewise linear on each triangle, such that φ_g is 1 at the vertex and 0 elsewhere (Miller et al., 2020). The GMRF representation of the spatial random effect can be incorporated into the linear predictor as

$$\eta(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \sum_{g=1}^{G} \tilde{\mathbf{A}}_{ig}\tilde{\boldsymbol{\omega}},$$

where $\tilde{\mathbf{A}}$ is an $n \times G$ sparse matrix and $\tilde{\boldsymbol{\omega}} = (\tilde{\omega}_1, \ldots, \tilde{\omega}_G)'$. The sparse matrix $\tilde{\mathbf{A}}$ maps the GMRF weights $\tilde{\boldsymbol{\omega}}$ from the n observations to the G triangulation nodes on the mesh. The model can then be expressed compactly as

$$\begin{aligned} \mathbf{Y} &\sim N_n(\boldsymbol{\eta}, \sigma_{\boldsymbol{\epsilon}}^2 I_n), \\ \boldsymbol{\eta} &= X\boldsymbol{\beta} + \tilde{\mathbf{A}}\tilde{\boldsymbol{\omega}}, \end{aligned}$$

where $\mathbf{Y} = \big(Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n)\big)'$, $\boldsymbol{\eta} = \big(\eta(\mathbf{s}_1), \ldots, \eta(\mathbf{s}_n)\big)'$, I_n denotes the $n \times n$ identity matrix and X denotes the $n \times p$ design matrix.

Step 4. Implement with INLA
Use the `INLA` package (Rue et al., 2009, Lindgren et al., 2011, Lindgren and Rue, 2015) in R to implement the SPDE method and fit the model. The mesh is created using the `inla.mesh.2d()` function, and the SPDE is defined using the `inla.spde2.pcmatern()` function. The `inla()` function is then used to fit the model to the data.

A limitation of the SPDE method is the restriction of the smoothness parameter α to $0 < \alpha \leq 2$, as addressed by Lindgren et al. (2011). This constraint ensures desirable mathematical properties and computational efficiency. For $\alpha \geq 2$, the null space of the pseudo-differential operator becomes non-trivial, posing challenges in obtaining results for alternative α values and implying an implicit assumption of appropriate boundary conditions for the SPDE. Maintaining $0 < \alpha \leq 2$ preserves the positive definiteness of the resulting covariance matrix, which is crucial for computational performance. As α increases, the covariance function becomes smoother, making the process more differentiable. Allowing $\alpha > 2$ could lead to overly smoothed covariance functions that might not adequately capture spatial variations, and it would also increase the computational burden. In practical terms, the Matérn covariance function with $0 < \alpha \leq 2$ performs well for a wide variety of spatial datasets (Lindgren et al., 2011).

2.7 Bayesian model selection

Modelling is a fundamental data analytic task and is a way to approximate reality (Shibata, 1989, Gelfand and Dey, 1994, Marin and Robert, 2014). The responsibilities of a statistician, when working with models, include assessing their utility, comparing their predictive performance and exploring directions for improvements. Even when all models under consideration have mismatches with the data, the simplest model can provide information towards the next step in model building (Gelman et al., 2014, Haining and Li, 2020).

2.7.1 Bayes factor

Bayesian model selection begins with the formal Bayes approach, which, in the context of two models, leads to the calculation of the Bayes factor (Gelfand and Dey, 1994). The Bayes factor is the standard solution for Bayesian model selection (Lewis and Raftery, 1997) and should be the only consideration within the orthodox Bayesian perspective (Sahu, 2022). Suppose there is a choice between two hypotheses, denoted as H_1 and H_2, which correspond to the underlying assumptions of models M_1 and M_2 respectively, for observations $\mathbf{y} = (y_1, \ldots, y_n)'$. The prior predictive distribution, denoted as $p(\mathbf{y}|M_i)$ for $i = 1, 2$, can be expressed as

$$p(\mathbf{y}|M_i) = \int_{-\infty}^{\infty} p_i(\mathbf{y}|\theta_i) p_i(\theta_i) d\theta_i, \tag{2.17}$$

where $p_i(\mathbf{y}|\theta_i)$ denotes the likelihood function, and $p_i(\theta_i)$ denotes the prior distribution. More explicitly,

$$p(\mathbf{y}|M_i) = \int_{-\infty}^{\infty} p_i(\mathbf{y}|\theta_i, M_i) p_i(\theta_i|M_i) d\theta_i.$$

The Bayes factor is defined as

$$\text{BF} = \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)}, \tag{2.18}$$

which is a ratio of the prior predictive distributions. The Bayes factor can take on any positive value and has many interpretations, as documented by Jeffreys (1961), Raftery (1996), Lee and Wagenmakers (2014). In general, a BF $>$ 1 favours the model in the numerator, which is M_1 in this example. Conversely, a BF < 1 provides evidence in support of the model in the denominator, which is M_2 in this example.

Suppose the models M_1 and M_2 have prior probabilities $P(M_1)$ and $P(M_2)$ respectively, with $P(M_1)+P(M_2) = 1$. Following Bayes theorem, the posterior probabilities $P(M_1|\mathbf{y})$ and $P(M_2|\mathbf{y})$, for M_1 and M_2, are calculated as

$$P(M_1|\mathbf{y}) = \frac{p(\mathbf{y}|M_1)P(M_1)}{p(\mathbf{y}|M_1)P(M_1) + p(\mathbf{y}|M_2)P(M_2)},$$

$$P(M_2|\mathbf{y}) = \frac{p(\mathbf{y}|M_2)P(M_2)}{p(\mathbf{y}|M_1)P(M_1) + p(\mathbf{y}|M_2)P(M_2)},$$

with $P(M_1|\mathbf{y}) + P(M_2|\mathbf{y}) = 1$. From these results, the posterior odds ratio is given as

$$\frac{P(M_1|\mathbf{y})}{P(M_2|\mathbf{y})} = \frac{p(\mathbf{y}|M_1)P(M_1)}{p(\mathbf{y}|M_2)P(M_2)},$$

$$= \frac{p(\mathbf{y}|M_1)}{p(\mathbf{y}|M_2)} \times \frac{P(M_1)}{P(M_2)}.$$

The expression above shows that the Bayes factor is the multiplicative factor used to covert the prior odds ratio to the posterior odds ratio. Rearranging the terms in the expression can also show that the Bayes factor is the ratio of posterior odds ratio and prior odds ratio. The Bayes factor provides a measure of whether the data has improved the odds for M_1 relative to M_2.

Bayes factors can be used for Bayesian hypothesis testing by quantifying the relative support for different hypotheses based on observed data and drawing conclusions about the posterior probabilities (O'Hagan, 2006). This enables informative and accurate interpretations of the evidence favouring or against different hypotheses, in contrast to the common misunderstandings surrounding p-values in frequentist null hypothesis significance tests. In such tests, p-values are often misconstrued as the probability that a null hypothesis H_0 is true or false, which is an inaccurate interpretation based on the underlying calculations.

However, Bayes factor are often not implemented in practical applications. As discussed by Zhu and Carlin, Bayes factor is difficult to compute and interpret for high-dimensional hierarchical models and for models with improper prior distributions (Zhu and Carlin, 2000). Bayes factor is also sensitive to the choice of the prior distribution for the parameter, denoted as $p_i(\theta_i)$ in (2.17). To address these challenges, alternative approaches to the Bayes factor have been proposed; see O'Hagan (1995) and Berger and Pericchi (1996) for example. We conclude our discussion on Bayes factor and will shift our focus to criterion-based model selection methods and cross validation methods for Bayesian model selection.

2.7.2 Information criteria

The concept of information criteria involves information theory and statistical analysis. Specifically, the information theory part is based on the Kullback-Leibler (K-L) divergence, which measures the difference between two probability distributions (Kullback and Leibler, 1951). Since model selection revolves around the task of approximation (Akaike, 1973, Buraham and Anderson, 1998), guided by the principle that "all models are wrong, but some are useful" (Box, 1976), the objective is to quantify information loss. The K-L information can be used to measure the amount of information lost when approximating the true underlying probability distribution with a model (Portet, 2020). The statistical analysis part comes from Akaike (1973) discovering an asymptotically unbiased estimator of the expected relative K-L divergence, which he termed "an information criterion". This was later named the "Akaike information criterion" (AIC) for his contribution to its development (Buraham and Anderson, 1998).

The AIC is a popular model choice criterion as it produces a metric for easy model comparisons. It comprises a goodness-of-fit component and a penalty component. The goodness-of-fit component is represented by the log likelihood function, also known as the log predictive density (Gelman et al., 2014). The log likelihood function is a general summary of the predictive fit and can assess the model fit since prediction accuracy serves as a proxy for model evaluation (Gneiting, 2011). The penalty component is the number of parameters estimated within the model, denoted as p, and act as a corrective measure against overfitting. The penalty component is consistent with the principle of parsimony, emphasising the importance of keeping a model as simple as possible (Johnson and Omland, 2004), as the fit of any model can be improved by increasing the number of parameters (Buraham and Anderson, 1998).

This concept is commonly known as the bias-variance tradeoff. Too few parameters can lead to high bias in the parameter estimates. This generally leads to an underfitted model that may fail to identify all important factors, because it is too simplistic to represent and capture all the nuances and details of the data accurately. Conversely, too many parameters can lead to high variance in the parameter estimates, resulting in an overfitted model. While an overfitted model may fit the observed data well, it often struggles to generalise for unobserved data.

The AIC is defined as

$$\mathrm{AIC} = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}) + 2p, \tag{2.19}$$

where $\log p(\mathbf{y}|\hat{\boldsymbol{\theta}})$ denotes the log likelihood function and $\hat{\boldsymbol{\theta}}$ denotes the maximum likelihood estimate. Portet suggests that the AIC is suitable when the number of observations is relatively large, typically when $(n/40) > p$. In cases where this condition is not met, the corrected AIC (AICc), introduced by Sugiura (1978), is the preferred choice (Portet, 2020). It is noteworthy that the penalty component of the AIC represents a specific case of a more general result derived by Takeuchi (1976), which also gave rise to the Takeuchi information criterion (TIC). Other adaptations of the AIC for other specific cases are discussed by Buraham and Anderson (1998), however, they are rarely used in practical applications due to computational difficulties and issues with stability (Vehtari and Ojanen, 2012).

The Bayesian information criterion (BIC), developed by Schwarz (1978), is another information criterion. The "Bayesian" aspect of the BIC arises from the Bayesian viewpoint of equal priors on the candidate models and vague priors on the parameters given the model. Models selected using BIC often assume the purpose of prediction rather than scientific understanding of the process or the system under study (Buraham and Anderson, 1998). The BIC is calculated as

$$\mathrm{BIC} = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}) + p \log(n).$$

Gelman et al. believe that the BIC is misleading as there is nothing "Bayesian" about its formula (Gelman et al., 2014). While the formula of the BIC may resemble (2.19), it is not an estimator related to the K-L divergence, from which the AIC originated. Instead, the BIC is based on a criterion called minimum description length, which is from coding theory, a branch of information theory (Rissanen, 1989, Yu, 1996, Buraham and Anderson, 1998). Furthermore, the BIC was developed with the assumption that a "true" model exists, and the motivation is to select that "true" model. Compared to the AIC, the BIC gives a larger penalty per parameter, and favours simpler models (Gelman et al., 2014). In practical applications, models with lower AIC or BIC values are preferred.

Beyond linear models or models involving flat prior distributions, the number of estimated parameters cannot be used as a penalty component. Informative prior distributions and models with hierarchical structure tend to reduce overfitting compared to simple least squares (Gelman et al., 2014). In situations where it is difficult to identify p, such as hierarchical models, the AIC and the BIC cannot be used for model comparisons (Spiegelhalter et al., 1998). We conclude our discussion on the AIC and BIC, and will now focus on the Bayesian variants of information criteria.

Deviance information criterion

The deviance information criterion (DIC), developed by Spiegelhalter et al. (1998), is often referred to as a generalisation of the AIC for hierarchical models, or a Bayesian version of the AIC (Spiegelhalter et al., 1998, Zhu and Carlin, 2000, Wheeler et al., 2010, Gelman et al., 2014, Sahu, 2022). To see why, observe the definition of the DIC

$$\text{DIC} = -2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\text{Bayes}}) + 2p_{\text{DIC}}. \tag{2.20}$$

The DIC is a Bayesian version of the AIC. There is resemblance between (2.20) and (2.19) in that they both comprise a goodness-of-fit component and a penalty component. The first notable difference between the AIC and the DIC is their representation of goodness-of-fit. In the AIC, this is the log likelihood function given the maximum likelihood estimate, denoted as $\log p(\mathbf{y}|\hat{\boldsymbol{\theta}})$. In the DIC, the goodness-of-fit component is the log likelihood function given some Bayes estimate, denoted as $\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\text{Bayes}})$. The Bayes estimate is typically the posterior mean, $\hat{\boldsymbol{\theta}}_{\text{Bayes}} = E(\boldsymbol{\theta}|\mathbf{y})$, as the posterior mean maximises the log predictive density if it is the same as the posterior mode.

The DIC is a generalisation of the AIC for hierarchical models. The development of the DIC stems from the need for Bayesian metrics that can assess model complexity and goodness-of-fit, particularly, in the context of comparing models with arbitrary structures, such as models with hierarchical structure. The DIC is a useful tool for models where the number of parameters is either not clearly defined or unknown (Spiegelhalter et al., 2002, Wheeler et al., 2010).

The reduction in uncertainty, due to estimation, is given as

$$d_\theta\{\mathbf{y},\boldsymbol{\theta},\tilde{\boldsymbol{\theta}}(\mathbf{y})\} = -2\log p(\mathbf{y}|\boldsymbol{\theta}) + 2\log p(\mathbf{y}|\tilde{\boldsymbol{\theta}}(\mathbf{y})),$$

where $\tilde{\theta}(\mathbf{y})$ is an estimator of the parameter θ. $d_{\boldsymbol{\theta}}$ can be thought of as the reduction in degree of "overfitting" due to the estimator $\tilde{\boldsymbol{\theta}}(\mathbf{y})$ adapting to the data $\mathbf{y}$. The unknown parameters in $d_{\boldsymbol{\theta}}$ can be estimated by its posterior expectation with respect to $p(\boldsymbol{\theta}|\mathbf{y})$,

$$\begin{aligned} p_D\{\mathbf{y},\boldsymbol{\Theta},\tilde{\boldsymbol{\theta}}(\mathbf{y})\} &= E_{\boldsymbol{\theta}|\mathbf{y}}\big(d_\theta\{\mathbf{y},\boldsymbol{\Theta},\tilde{\boldsymbol{\theta}}(\mathbf{y})\}\big), \\ &= E_{\boldsymbol{\theta}|\mathbf{y}}\big(-2\log p(\mathbf{y}|\boldsymbol{\theta}) + 2\log p(\mathbf{y}|\tilde{\boldsymbol{\theta}}(\mathbf{y}))\big), \\ &= 2\log p(\mathbf{y}|\tilde{\boldsymbol{\theta}}(\mathbf{y})) + E_{\boldsymbol{\theta}|\mathbf{y}}\big(-2\log p(\mathbf{y}|\boldsymbol{\theta})\big), \end{aligned}$$

where p_D is a measure of complexity for the effective number of parameters (Spiegelhalter et al., 1998, 2002). p_D is also referred to as the "effective number of parameters", as originally termed by Moody (1991).

Returning to (2.20), the component p_{DIC} is defined as

$$p_{\text{DIC}} = p_D\{\mathbf{y},\boldsymbol{\Theta},\tilde{\boldsymbol{\theta}}(\mathbf{y})\} = 2\big(\log p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\text{Bayes}}) - E_{\boldsymbol{\theta}|\mathbf{y}}(\log p(\mathbf{y}|\boldsymbol{\theta}))\big). \tag{2.21}$$

The penalty component of the DIC can be thought of as the difference between the posterior mean of the deviance and the deviance at the posterior means of the parameters of interest (Spiegelhalter et al., 2002). The p_{DIC} is often less than the total number of model parameters, due to the borrowing of strength across individual level parameters in hierarchical models (Zhu and Carlin, 2000). If the posterior mean is far away from the posterior mode, the p_{DIC} will be negative. A penalty component that produces a negative value is counterproductive, since it is placed to correct for overfitting. To address this, Gelman et al. proposed an alternative penalty component, denoted as $p_{\text{DIC alt}}$. It is defined as

$$p_{\text{DIC alt}} = 2\text{Var}_{\boldsymbol{\theta}|\mathbf{y}}\big(\log p(\mathbf{y}|\boldsymbol{\theta})\big), \tag{2.22}$$

where $\text{Var}_{\boldsymbol{\theta}|\mathbf{y}}(\cdot)$ denotes the posterior variance. Using (2.22) ensures that the penalty component will be positive and also avoids the use of a "plug-in" estimate, which is typically invariant to reparameterisation. However, (2.22) can be unstable (Vehtari and Ojanen, 2012). In general, the p_{DIC} and $p_{\text{DIC alt}}$ are both accurate if the limit of the fixed model and large n can be derived from the asymptotic chi-squared distribution of the log predictive density (Vehtari and Ojanen, 2012, Gelman et al., 2014). Between the two versions of the penalty component, the p_{DIC} is more numerically stable, but $p_{\text{DIC alt}}$ has the advantage of always being positive. For linear models with uniform prior distributions, both p_{DIC} and $p_{\text{DIC alt}}$ reduce to p, which is the same as the AIC (Gelman et al., 2014). Sahu demonstrates this through a simple example in Chapter 4 of his book (Sahu, 2022).

Spiegelhalter et al. suggest that the DIC should be used as a method to screen for alternative formulations of the model to produce a list of candidate models for further consideration. They also expect the DIC to be strongly related to the cross-validatory assessment (Spiegelhalter et al., 1998), a topic that will be further explored in Section 2.7.3.

Watanabe-Akaike information criterion

A statistical model is "regular" if its parameters are mapped one-to-one to probability distributions and if its Fisher information matrix is positive definite (Watanabe and Opper, 2010). If a statistical model is not regular, it is "singular". A statistical model with hierarchical structure and latent variables is typically singular (Watanabe and Opper, 2010, Watanabe, 2010). In singular statistical models, the maximum likelihood estimator does not satisfy asymptotic normality. Hence, the maximum likelihood method is not appropriate. The number of parameters in models with hierarchical and mixture structures tend to increase with sample size (Gelman et al., 2014). The Watanabe-Akaike information criterion, also referred to as the "widely applied information criterion", was developed by Watanabe and Opper (2010) for such models.

The WAIC is considered to be an improvement over the DIC (Vehtari et al., 2017b). The WAIC has a more desirable property of averaging over the posterior distribution, rather than conditioning on a point estimate $\hat{\boldsymbol{\theta}}_{\text{Bayes}}$. This renders the WAIC a fully Bayesian metric as it makes use of the complete posterior distribution. Furthermore, the WAIC incorporates the posterior predictive density into its goodness-of-fit component. It is noteworthy that the WAIC is asymptotically equal to

Bayesian cross-validation. In a prediction context, the WAIC assesses the predictions used for new data (Gelman et al., 2014).

The WAIC is defined as follows

$$\text{WAIC} = -2\sum_{i=1}^{n} \log\left(\int_{-\infty}^{\infty} p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}\right) + 2p_{\text{WAIC}},$$

where the first term is really the posterior predictive distribution, denoted as $p(y_i|\mathbf{y})$. The definition can be re-expressed as

$$\text{WAIC} = -2\sum_{i=1}^{n} \log p(y_i|\mathbf{y}) + 2p_{\text{WAIC}}. \tag{2.23}$$

The penalty component of the WAIC comes in two versions: one obtained by using the posterior mean and the other based on the posterior variance

$$p_{\text{WAIC1}} = 2\sum_{i=1}^{n}\Bigg(\log E_{\boldsymbol{\theta}|\mathbf{y}}(p(y_i|\boldsymbol{\theta})) - E_{\boldsymbol{\theta}|\mathbf{y}}(\log p(y_i|\boldsymbol{\theta}))\Bigg), \tag{2.24}$$

$$p_{\text{WAIC2}} = \sum_{i=1}^{n} \text{Var}_{\boldsymbol{\theta}|\mathbf{y}}\big(\log p(y_i|\boldsymbol{\theta})\big). \tag{2.25}$$

Gelman et al. (2014) recommend using p_{WAIC2} over p_{WAIC1}, because the series expansion of p_{WAIC2} bears a closer resemblance to the series expansion of leave-one-out cross-validation. Moreover, in practical applications, p_{WAIC2} gives results close to those obtained from leave-one-out cross-validation.

For practical purposes, the WAIC is calculated with S posterior draws $\boldsymbol{\theta}^{(s)}$, such that

$$\text{WAIC} = -2\sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}^{(s)})\right) + 2p_{\text{WAIC}}, \tag{2.26}$$

where p_{WAIC} is either

$$p_{\text{WAIC1}} = 2\sum_{i=1}^{n}\Bigg(\log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}^{(s)})\right) - \frac{1}{S}\sum_{s=1}^{S}\log p(y_i|\boldsymbol{\theta}^{(s)})\Bigg), \tag{2.27}$$

$$p_{\text{WAIC2}} = \sum_{i=1}^{n} V_{s=1}^{S}\big(\log p(y_i|\boldsymbol{\theta}^{(s)})\big), \tag{2.28}$$

where $V_{s=1}^{S}(\cdot)$ represents the sample variance, i.e.,

$$V_{s=1}^{S}(a_s) = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2.$$

In practice, the candidate model with the lowest WAIC value is preferred.

The WAIC relies on partitioning the data into n parts, which is often not a simple task for structured data, such as time-series data and spatial data (Gelman et al., 2014). Additionally, the WAIC assumes that the partitioned data are both disjointed and, ideally, conditionally independent. This assumption poses a limitation when dealing with models for point referenced spatial data. To address this limitation, we will introduce the WAIC_{NF} in Chapter 3.

2.7.3 Leave-one-out cross validation

The prediction accuracy of a model can be used to measure the performance of the model and compare against models. Measures of predictive accuracy for probabilistic predictions are called scoring rules, and the logarithmic score is the most commonly used scoring rule in model selection (Bernardo, 1979, Gneiting and Raftery, 2007). The ideal measure of a model's fit is its out-of-sample predictive performance for new data produced from the true data-generating process. Gelman et al. (2014) define the expected log predictive density for a new data point (elpd) as

$$\text{elpd} = \int_{-\infty}^{\infty} \log p_{\text{post}}(\tilde{y}_i) f(\tilde{y}_i) d\tilde{y}_i,$$

where $p_{\text{post}}(\cdot)$ denotes a probability that averages over the posterior distribution of the unknown parameters $\boldsymbol{\theta}$.

Directly computing the elpd using the expression above can be difficult, since often time the true distribution f is unknown. Instead, various estimation methods are available to approximate the elpd. One approach involves estimating the elpd for new data using the log predictive density with existing data. This is referred to as the within-sample predictive accuracy. While this approach provides a summary that is quick and easy to understand, it generally leads to overfitting, which can be corrected by subtracting the effective number of parameters. In other words, the goal is to estimate the expected out-of-sample prediction error using bias-corrected adjustment of within-sample error. Examples of this adjusted within-sample predictive accuracy approach include the AIC, DIC and WAIC, as detailed in Section 2.7.2.

Another approach to assess out-of-sample predictive accuracy involves partitioning the dataset into training and testing subsets. The model is then fitted to the training data, and its predictive accuracy is evaluated using the testing data. This technique is commonly referred to as cross-validation. While cross-validation avoids overfitting, it can become computationally expensive, as it typically requires many data partitions and model fitting to obtain a stable estimate.

Leave-one-out cross validation (LOO) is a special case of cross-validation. Let $\mathbf{y} = (y_1, \ldots, y_n)'$ denote the observations, where y_i denotes the ith observation and $\mathbf{y}_{-i}$ denotes all observations except for y_i. LOO is when the model is fitted to $\mathbf{y}_{-i}$ and evaluated with y_i. The Bayesian LOO is given as

$$p(y_i|\mathbf{y}_{-i}) = \int_{-\infty}^{\infty} p(y_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{-i})d\boldsymbol{\theta}. \tag{2.29}$$

Assuming that the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}_{-i})$ is summarised by S simulation draws of $\boldsymbol{\theta}^{(s)}$, the log predictive density can be calculated as $\log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}^{(s)})\right)$. Gelman et al. re-express (2.29) as

$$\text{LOO} = \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}^{(is)})\right), \tag{2.30}$$

where $\boldsymbol{\theta}^{(is)}$ denotes the $s = 1, \ldots, S$ posterior simulations of $\boldsymbol{\theta}$ for data point $i = 1, \ldots, n$. It is important to note that both cross-validation and LOO assume that data are partitioned into disjointed and, ideally, conditionally independent parts (Gelman et al., 2014). Similar to the WAIC, this assumption presents a limitation when applied to models for point referenced spatial data.

PSIS-LOO

The Pareto smoothed importance sampling leave-one-out cross validation (PSIS-LOO) was developed by Vehtari et al. (2015) for a better approximation of the Bayesian LOO (2.29). As mentioned before, calculating the exact LOO requires fitting the model n times, which can be computationally expensive. Gelfand et al. (1992) discovered that if the test data, also referred to as hold out points, are conditionally independent, the Bayesian LOO can be evaluated using importance ratios to get the importance sampling leave-one-out cross-validation (IS-LOO). The IS-LOO predictive density is given as

$$p(y_i|\mathbf{y}_{-i}) \approx \frac{\sum_{s=1}^{S} r_i^{(s)} p(y_i|\boldsymbol{\theta}^{(s)})}{\sum_{s=1}^{S} r_i^{(s)}},$$

where $r_i^{(s)}$ denotes the importance ratio and is given as

$$r_i^{(s)} = \frac{1}{p(y_i|\boldsymbol{\theta}^{(s)})} \propto \frac{p(\boldsymbol{\theta}^{(s)}|\mathbf{y}_{-i})}{p(\boldsymbol{\theta}^{(s)}|\mathbf{y})}.$$

The IS-LOO predictive density can be further expressed as

$$p(y_i|\mathbf{y}_{-i}) \approx \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|\boldsymbol{\theta}^{(s)})}}.$$

However, directly applying this induces instability, because $r_i^{(s)}$ can have high or infinite variance (Vehtari et al., 2017b). To resolve this problem, Ionides proposed a modification of the importance ratios by truncated weighting. More explicitly, instead of using $r_i^{(s)}$, use

$$w_i^{(s)} = \min(r_i^{(s)}, \sqrt{S}\bar{r}_i),$$

where $\bar{r}_i = \frac{1}{S}\sum_{s=1}^{S} r_i^{(s)}$ (Ionides, 2008). Using these truncated weights gives a mean square error close to an estimate with a case specific optimal truncation level. However, the downside in using this approach is that it introduces a bias that can potentially be large.

Vehtari et al. proposed the PSIS-LOO, an approach that is more accurate and reliable, and one that is more robust in the finite case with weak priors or influential observations (Vehtari et al., 2015). The procedure of the PSIS-LOO algorithm is outlined as follows.

Suppose there are $s = 1, \ldots, S$ posterior draws and $S = 100$. The inputs of the algorithm are the raw importance ratios $r_s = (r_1, \ldots, r_{100})'$ ordered from the lowest to highest value, and the outputs are the Pareto smoothed importance weights $w_s = (w_1, \ldots, w_{100})'$. First, determine the number of largest r_s to extract. This number is denoted as M, and the general guideline is $M = \min(0.2S, 3\sqrt{S})$. Since

$S = 100$, then $M = 20$. Next, set $\tilde{w}_s = r_s$, for $s = 1, \ldots, S - M$. Since $S = 100$ and $M = 20$, $(\tilde{w}_1, \ldots, \tilde{w}_{80})' = (r_1, \ldots, r_{80})'$. Estimate the parameters $\hat{k}$ and $\hat{\sigma}$ in the generalised Pareto distribution with $(r_{81}, \ldots, r_{100})'$ using the algorithm by Zhang and Stephens (2009) with a weakly informative Gaussian prior distribution. Then set

$$\tilde{w}_{S-M+z} = F^{-1}\left(\frac{z - 1/2}{M}\right)$$

for $z = 1, \ldots, M$. The notation $F^{-1}(\cdot)$ denotes the inverse cumulative distribution function (CDF) of the generalised Pareto distribution, and is given as

$$F^{-1}\left(\frac{z - 1/2}{M}\right) = u + \frac{\hat{\sigma}}{\hat{k}}\left(\left(1 - \frac{z - 1/2}{M}\right)^{\hat{k}} - 1\right)$$

where u are the M extracted importance ratios, $u = (r_{81}, \ldots, r_{100})$. To guarantee finite variance for the estimate, Vehtari et al. (2017b) included an additional step

$$w_s = \min(\tilde{w}_s, S^{3/4}\bar{w}),$$

where $\bar{w} = \frac{1}{S}\sum_{s=1}^{S} \tilde{w}_s$. Note that while not explicitly denoted, this algorithm is performed for each i hold out data point.

Vehtari et al. define the PSIS-LOO estimate of the LOO expected pointwise predictive density as

$$\widehat{\text{elpd}}_{\text{PSIS-LOO}} = \sum_{i=1}^{n} \log\left(\frac{\sum_{s=1}^{S} w_i^{(s)} p(y_i|\boldsymbol{\theta}^{(s)})}{\sum_{s=1}^{S} w_i^{(s)}}\right). \tag{2.31}$$

Additionally, (2.31) be transformed to the deviance scale by

$$\text{PSIS-LOOIC} = -2\widehat{\text{elpd}}_{\text{PSIS-LOO}}, \tag{2.32}$$

where PSIS-LOOIC stands for PSIS-LOO information criterion. PSIS-LOOIC is useful for comparing against other information criteria, including the DIC and WAIC.

An important consideration is that when the estimated $\hat{k}$ from the generalised Pareto distribution exceeds 0.7, the resulting importance sampling estimates are likely to exhibit instability. In such instances, Vehtari et al. suggest several potential solutions. These include drawing samples directly from $p(\boldsymbol{\theta}^{(s)}|\mathbf{y}_{-i})$ for the problematic ith observation, employing a K-fold cross-validation approach, or utilising a more robust model (Vehtari et al., 2017b). Additionally, this algorithm relies on the assumption that the n hold-out data points are conditionally independent in the data model (Gelman et al., 2014). This is the same problem described for the WAIC in Section 2.7.2, and it poses a limitation when applied to applied to models of point referenced spatial data. To address this limitation, we will introduce the PSIS-LOOIC$_{\mathrm{NF}}$ in Chapter 3.

Chapter 3

Calculating the WAIC for non-factorisable models

The two primary objectives of this chapter are the following. The first objective is to highlight the challenge that arises when applying the computation of the Watanabe-Akaike information criterion (WAIC) and the Pareto smoothed importance sampling leave-one-out cross-validation information criterion (PSIS-LOOIC) to "non-factorisable" models. The second objective is to introduce the novel approach, centered on the non-factorisable model likelihood, for WAIC and PSIS-LOOIC computation, specifically in the context of Bayesian models of point referenced spatial data.

3.1 The challenge posed by non-factorisable models in WAIC computation

First, recall the important detail about the WAIC and PSIS-LOOIC from the discussion in Sections 2.7.2 and 2.7.3. The WAIC (2.23) is given by

$$\text{WAIC} = -2\sum_{i=1}^{n} \log p(y_i|\mathbf{y}) + 2p_{\text{WAIC}},$$

where $\log p(y_i|\mathbf{y})$ denotes the posterior predictive distribution, and p_{WAIC} is the penalty component. The penalty component is defined in two ways, (2.24) and (2.25), given by

$$p_{\mathrm{WAIC1}} = 2\sum_{i=1}^{n}\Big(\log E_{\boldsymbol{\theta}|\mathbf{y}}(p(y_i|\boldsymbol{\theta})) - E_{\boldsymbol{\theta}|\mathbf{y}}(\log p(y_i|\boldsymbol{\theta}))\Big),$$
$$p_{\mathrm{WAIC2}} = \sum_{i=1}^{n}\mathrm{Var}_{\boldsymbol{\theta}|\mathbf{y}}\big(\log p(y_i|\boldsymbol{\theta})\big),$$

where $E_{\boldsymbol{\theta}|\mathbf{y}}(\cdot)$ denotes the posterior average and $\mathrm{Var}_{\boldsymbol{\theta}|\mathbf{y}}(\cdot)$ denotes the posterior variance. In practical applications, the WAIC and its penalty components are calculated using S posterior draws of the unknown parameters $\boldsymbol{\theta}$, (2.26), (2.27) and (2.28). Explicitly, they are

$$\mathrm{WAIC} = -2\sum_{i=1}^{n}\log\left(\frac{1}{S}\sum_{s=1}^{S}p(y_i|\boldsymbol{\theta}^{(s)})\right) + 2p_{\mathrm{WAIC}},$$
$$p_{\mathrm{WAIC1}} = 2\sum_{i=1}^{n}\left(\log\left(\frac{1}{S}\sum_{s=1}^{S}p(y_i|\boldsymbol{\theta}^{(s)})\right) - \frac{1}{S}\sum_{s=1}^{S}\log p(y_i|\boldsymbol{\theta}^{(s)})\right),$$
$$p_{\mathrm{WAIC2}} = \sum_{i=1}^{n}V_{s=1}^{S}\big(\log p(y_i|\boldsymbol{\theta}^{(s)})\big),$$

for $s = 1, \dots, S$, and where $V_{s=1}^{S}(\cdot)$ denotes the sample variance.

The PSIS-LOO estimate of the LOO expected pointwise predictive density (2.31) is given by

$$\widehat{\mathrm{elpd}}_{\mathrm{PSIS-LOO}} = \sum_{i=1}^{n}\log\left(\frac{\sum_{s=1}^{S}w_i^{(s)}p(y_i|\boldsymbol{\theta}^{(s)})}{\sum_{s=1}^{S}w_i^{(s)}}\right),$$

where $w_i^{(s)}$ are the Pareto smoothed importance weights calculated following the outline described in Section 2.7.3. The PSIS-LOOIC (2.32) is given by

$$\text{PSIS-LOOIC} = -2\widehat{\mathrm{elpd}}_{\mathrm{PSIS-LOO}} = -2\sum_{i=1}^{n}\log\left(\frac{\sum_{s=1}^{S}w_i^{(s)}p(y_i|\theta^{(s)})}{\sum_{s=1}^{S}w_i^{(s)}}\right).$$

An important aspect about the formula of both the WAIC and PSIS-LOOIC is that they assume the data to be partitioned into disjointed and, ideally, conditionally independent parts (Gelman et al., 2014). A model that satisfies this assumption is called a "factorised model". More specifically, if the observation model is formulated directly as the product of the pointwise observations, it is called a "factorised model"

(Vehtari et al., 2018, Bürkner et al., 2021). A factorised model is a model that can have a full (log) likelihood function expressed as

$$\begin{aligned} p(\mathbf{y}|\psi) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{\psi}), \\ \log p(\mathbf{y}|\psi) &= \sum_{i=1}^{n} \log p(y_i|\boldsymbol{\psi}), \end{aligned}$$

where $\boldsymbol{\psi}$ denotes some model parameters and $\mathbf{y} = (y_1, \ldots, y_n)'$. This implies that the observations y_i are conditionally independent of one another. Conversely, a "non-factorisable" model is when the observations are not conditionally independent, and the full likelihood function cannot be written in the form shown above.

Response values from models of structured data, such as models of point referenced spatial data, often exhibit conditional dependence. We would expect spatial dependence among the sites in a point referenced spatial dataset, as nearby points tend to influence each other. Models of structured data are often characterised by multivariate normal distributions with some structured covariance matrix that does not factorise. Within the context of point referenced spatial data, this covariance matrix has dimensions corresponding to the number of sites, and its elements are calculated from covariance functions such as (2.5) and (2.6), as discussed in Section 2.3.3. The dependence of observations on other observations from a different spatial unit is one of the key features that spatial models aim to capture (Hooten and Hobbs, 2015). This key characteristic is a reason that a model may be considered as a non-factorisable model.

Bürkner et al. noted that, conceptually, neither a factorised model nor the conditional independence assumption are necessary prerequisites for LOO (Bürkner et al., 2021). This conceptually extends to the calculation of the WAIC and PSIS-LOOIC. However, imposing models of structured data to follow factorised model strategies for LOO potentially introduces computational inefficiency and numerical instability (Bürkner et al., 2021). In cases where non-factorisable model strategies are at our disposal, their adoption is warranted, particularly when handling models inherently characterised as non-factorisable. Moreover, utilising non-factorisable model strategies for LOO under the appropriate context can provide computational efficiency and numerical stability.

To summarise, the challenge posed by non-factorisable models in the calculation of the WAIC and PSIS-LOOIC is that it does not satisfy the conditional independence assumption. We will provide an approach to address this challenge in Section 3.5.

3.2 Multivariate calculation of the WAIC

In this section, we make our first attempt to calculate the WAIC for models of point referenced spatial data. We employ the strategy of directly applying multivariate normal distributions to the calculation of the WAIC. Consider the following,

$$\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 H), \tag{3.1}$$

where random variables $Y = (Y_1, \ldots, Y_n)'$ follows an n-dimensional multivariate normal distribution with mean structure $X\boldsymbol{\beta}$ and covariance matrix $\sigma^2 H$. The mean structure comprise an $n \times p$ design matrix, denoted as X, and a $p \times 1$ column vector of regression coefficients, denoted as $\boldsymbol{\beta}$. Suppose σ^2 is a known value and H is an $n \times n$ identity matrix, denoted as I_n. Let us assume the prior distribution of $\boldsymbol{\beta}$ to be

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, \sigma^2 M^{-1}), \tag{3.2}$$

where $N_p(\cdot)$ denotes a p-dimensional multivariate normal distribution with mean structure $\boldsymbol{\beta}_0$ and covariance matrix $\sigma^2 M^{-1}$. The mean structure $\boldsymbol{\beta}_0$ is a $p \times 1$ column vector of known values. Suppose σ^2 in the covariance matrix is the same known value, and M is a $p \times p$ identity matrix, denoted as I_p. This implies that $M = M^{-1} = I_p$. To assist the following derivations, define $\lambda^2 = 1/\sigma^2$.

The PDFs of (3.1) and (3.2) are

$$\begin{aligned} p(\mathbf{y}|\boldsymbol{\beta}, \lambda^2) &= \left(\frac{\lambda^2}{2\pi}\right)^{\frac{n}{2}} \det(H)^{-\frac{1}{2}} \exp\left(-\frac{\lambda^2}{2}(\mathbf{y} - X\boldsymbol{\beta})' H^{-1} (\mathbf{y} - X\boldsymbol{\beta})\right), \\ p(\boldsymbol{\beta}|\lambda^2) &= \left(\frac{\lambda^2}{2\pi}\right)^{\frac{p}{2}} \det(M)^{\frac{1}{2}} \exp\left(-\frac{\lambda^2}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' M (\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right), \end{aligned}$$

respectively, where $\det(\cdot)$ denotes the determinant. Furthermore, the posterior mean and posterior variance of $\boldsymbol{\beta}$ are

$$E(\boldsymbol{\beta}|\mathbf{y}, \lambda^2) = \boldsymbol{\beta}^*, \tag{3.3}$$

$$\text{Var}(\boldsymbol{\beta}, |\mathbf{y}, \lambda^2) = \frac{1}{\lambda^2}(M^*)^{-1}, \tag{3.4}$$

where

$$\boldsymbol{\beta}^* = (M^*)^{-1}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0),$$
$$M^* = X'H^{-1}X + M.$$

The posterior predictive distribution for a new observation, denoted as $\tilde{Y}_0$, is given as

$$\tilde{Y}_0|\mathbf{y}, \lambda^2 \sim N\Bigg(\mathbf{g}'\boldsymbol{\beta}^*, \frac{1}{\lambda^2}\big(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g}\big)\Bigg),$$

with the posterior mean and posterior variance of $\tilde{Y}_0$ given as

$$E(\tilde{Y}_0|\mathbf{y}, \lambda^2) = \mathbf{g}'\boldsymbol{\beta}^*, \tag{3.5}$$

$$\text{Var}(\tilde{Y}_0|\mathbf{y}, \lambda^2) = \frac{1}{\lambda^2}\big(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g}\big), \tag{3.6}$$

where

$$\mathbf{g}' = (\mathbf{x}_0' - \boldsymbol{\Sigma}_{12}H^{-1}X),$$
$$\delta^2 = (1 - \boldsymbol{\Sigma}_{12}H^{-1}\boldsymbol{\Sigma}_{21}).$$

and $\mathbf{x}_0$ denotes the corresponding elements of the regression variables for $\tilde{Y}_0$. In the expressions above, $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}'$ and $\boldsymbol{\Sigma}_{12}$ is an n-dimensional row vector with elements $\text{Cor}(\tilde{Y}_0, Y_i)$ for $i = 1, \ldots, n$, where $\text{Cor}(\cdot)$ denotes the correlation function. See Appendix B for the derivation of the posterior mean (3.3) and variance (3.4), and Appendix C for the derivation of the posterior mean (3.5) and posterior variance (3.6) of the posterior predictive distribution.

The WAIC (2.23) is given by

$$\text{WAIC} = -2\sum_{i=1}^{n}\log p(y_i|\mathbf{y}) + 2p_{\text{WAIC}},$$

where the penalty component p_{WAIC} is defined in two ways, (2.24) and (2.25), given by

$$p_{\text{WAIC1}} = 2\sum_{i=1}^{n}\left(\log E_{\boldsymbol{\theta}|\mathbf{y}}(p(y_i|\boldsymbol{\theta})) - E_{\boldsymbol{\theta}|\mathbf{y}}(\log p(y_i|\boldsymbol{\theta}))\right),$$

$$p_{\text{WAIC2}} = \sum_{i=1}^{n}\text{Var}_{\boldsymbol{\theta}|\mathbf{y}}\big(\log p(y_i|\boldsymbol{\theta})\big),$$

Derivation of p_{WAIC1}. Employing our multivariate calculation strategy, we write the p_{WAIC1} as

$$\begin{aligned} p_{\text{WAIC1}} &= 2\left(\log E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(p(\mathbf{y}|\boldsymbol{\beta},\lambda^2)\big) - E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(\log p(\mathbf{y}|\boldsymbol{\beta},\lambda^2)\big)\right), \\ &= 2\left(\log p(\tilde{y}_0|\mathbf{y},\lambda^2) - E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(\log p(\mathbf{y}|\boldsymbol{\beta},\lambda^2)\big)\right), \end{aligned}$$

where the $\boldsymbol{\beta}$ is set in place of $\boldsymbol{\theta}$ as it is the unknown parameter of interest. To make the derivation clearer, the two components within the outmost brackets are calculated separately before combining back together. The first component requires the log PDF of the posterior predictive distribution, which is given as

$$\log p(\tilde{y}_0|\mathbf{y},\lambda^2) = \frac{1}{2}\log\left(\frac{\lambda^2}{2\pi}\right) - \frac{1}{2}\log(\delta^2+\mathbf{g}'(M^*)^{-1}\mathbf{g}) - \frac{\lambda^2}{2}(\delta^2+\mathbf{g}'(M^*)^{-1}\mathbf{g})(\tilde{y}_0-\mathbf{g}'\boldsymbol{\beta}^*)^2.$$

The second component requires the help of the following. Let $Z = \mathbf{y} - X\boldsymbol{\beta}$ and $A = H^{-1}$. By definition,

$$E[Z'AZ] = \text{tr}(A\Sigma) + \mu' A\mu, \tag{3.7}$$

where $\text{tr}(\cdot)$ denotes the trace; see Mathai and Provost (1992) for more information regarding quadratic forms of random variables. In (3.7), μ denotes the posterior mean and Σ denotes the posterior variance of Z. They are calculated as follows

$$\begin{aligned}
\mu &= E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}(Z),\\
&= E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}(\mathbf{y} - X\boldsymbol{\beta}),\\
&= \mathbf{y} - XE_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}(\boldsymbol{\beta}),\\
&= \mathbf{y} - X\boldsymbol{\beta}^*. \qquad (3.8)
\end{aligned}$$

$$\begin{aligned}
\Sigma &= \mathrm{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}(Z),\\
&= \mathrm{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}(\mathbf{y} - X\boldsymbol{\beta}),\\
&= \mathrm{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}(-X\boldsymbol{\beta}),\\
&= (-X)\mathrm{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}(\boldsymbol{\beta})(-X)',\\
&= X\Sigma_\beta X', \qquad (3.9)
\end{aligned}$$

where Σ_β denotes the variance of the posterior distribution (3.4). With this information, the second component is derived as

$$\begin{aligned}
&E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(\log p(\mathbf{y}|\boldsymbol{\beta},\lambda^2)\big)\\
&= E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\left(\frac{n}{2}\log\left(\frac{\lambda^2}{2\pi}\right) - \frac{1}{2}\log\big(\det(H)\big) - \frac{\lambda^2}{2}(\mathbf{y} - X\boldsymbol{\beta})'H^{-1}(\mathbf{y} - X\boldsymbol{\beta})\right),\\
&= \frac{n}{2}\log\left(\frac{\lambda^2}{2\pi}\right) - \frac{1}{2}\log\big(\det(H)\big) - \frac{\lambda^2}{2}E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\Big((\mathbf{y} - X\boldsymbol{\beta})'H^{-1}(\mathbf{y} - X\boldsymbol{\beta})\Big),\\
&= \frac{n}{2}\log\left(\frac{\lambda^2}{2\pi}\right) - \frac{1}{2}\log\big(\det(H)\big) - \frac{\lambda^2}{2}\Big(\mathrm{tr}\big(H^{-1}X\Sigma_\beta X\big) + (\mathbf{y} - X\boldsymbol{\beta}^*)'H^{-1}(\mathbf{y} - X\boldsymbol{\beta}^*)\Big).
\end{aligned}$$

Combining the two parts gives us the derivation of p_{WAIC1} using the multivariate calculation strategy

$$\begin{aligned}
p_{\mathrm{WAIC1}} = {} & \log\left(\frac{\lambda^2}{2\pi}\right) - \log(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g}) - \lambda^2(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g})(\tilde{y}_0 - \mathbf{g}'\boldsymbol{\beta}^*)^2\\
& - n\log\left(\frac{\lambda^2}{2\pi}\right) + \log\big(\det(H)\big) + \lambda^2\Big(\mathrm{tr}\big(H^{-1}X\Sigma_\beta X\big) + (\mathbf{y} - X\boldsymbol{\beta}^*)'H^{-1}(\mathbf{y} - X\boldsymbol{\beta}^*)\Big).
\end{aligned} \tag{3.10}$$

Derivation of p_{WAIC2}. Employing our multivariate calculation strategy, we write the

p_{WAIC2} as

$$p_{\text{WAIC2}} = \text{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(\log p(\mathbf{y}|\boldsymbol{\beta},\lambda^2)\big),$$

where the $\boldsymbol{\beta}$ is set in place of $\boldsymbol{\theta}$ as it is the unknown parameter of interest. The derivation requires the definition of the variance of random variables in the quadratic form. Again, let $Z = \mathbf{y} - X\boldsymbol{\beta}$ and $A = H^{-1}$. By definition,

$$\text{Var}[Z'AZ] = 2\text{tr}(A\Sigma A\Sigma) + 4\mu' A\Sigma A\mu, \tag{3.11}$$

where μ and Σ are the posterior mean and posterior variance of Z as derived in (3.8) and (3.9) respectively. With this information, the derivation is given as follows

$$\begin{aligned}
p_{\text{WAIC2}} &= \text{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(\log p(\mathbf{y}|\boldsymbol{\beta},\lambda^2)\big),\\
&= \text{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\bigg(\frac{n}{2}\log\left(\frac{\lambda^2}{2\pi}\right) - \frac{1}{2}\log\big(\det(H)\big) - \frac{\lambda^2}{2}(\mathbf{y} - X\boldsymbol{\beta})'H^{-1}(\mathbf{y} - X\boldsymbol{\beta})\bigg),\\
&= \text{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\bigg(- \frac{\lambda^2}{2}(\mathbf{y} - X\boldsymbol{\beta})'H^{-1}(\mathbf{y} - X\boldsymbol{\beta})\bigg),\\
&= \frac{\lambda^4}{4}\text{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\bigg((\mathbf{y} - X\boldsymbol{\beta})'H^{-1}(\mathbf{y} - X\boldsymbol{\beta})\bigg),\\
&= \frac{\lambda^4}{4}\bigg(2\text{tr}\big(H^{-1}(X\Sigma_\beta X')H^{-1}(X\Sigma_\beta X')\big) + 4(\mathbf{y} - X\boldsymbol{\beta}^*)'H^{-1}(X\Sigma_\beta X')H^{-1}(\mathbf{y} - X\boldsymbol{\beta}^*)\bigg).
\end{aligned} \tag{3.12}$$

Returning to the WAIC (2.23), we rewrites the formula as follows

$$\begin{aligned}
\text{WAIC} &= 2\log p(\tilde{y}_0|\mathbf{y},\lambda^2) + 2p_{\text{WAIC}},\\
&= 2\bigg(\frac{1}{2}\log\left(\frac{\lambda^2}{2\pi}\right) - \frac{1}{2}\log(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g}) - \frac{\lambda^2}{2}(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g})(\tilde{y}_0 - \mathbf{g}'\boldsymbol{\beta}^*)^2\bigg) + 2p_{\text{WAIC}},\\
&= \log\left(\frac{\lambda^2}{2\pi}\right) - \log(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g}) - \lambda^2(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g})(\tilde{y}_0 - \mathbf{g}'\boldsymbol{\beta}^*)^2 + 2p_{\text{WAIC}},
\end{aligned}$$

where p_{WAIC} can be either (3.10) or (3.12).

To verify (3.10) and (3.12), let $p = 1$ and $X = \mathbf{1}_{n\times 1}$, which denotes an n-dimensional column vector of ones. Since H is the identity matrix, Σ_{12} is a null vector, $\delta^2 = 1$ and $\mathbf{g}' = \mathbf{x}_0'$. The initial posterior predictive distribution accounts for one new observation. Since the p_{WAIC1} calculation requires a summation of n parts, the verification here follows by summing the first part by n parts. Let $\mathbf{x}_0' = X_0 = X = c$ and

$\tilde{y}_0 = \mathbf{y}$.

$$
\begin{aligned}
p_{\text{WAIC1}} &= n\log\left(\frac{\lambda^2}{2\pi}\right) - n\log\left(\frac{\lambda^2}{2\pi}\right) \\
&- \log\big(\det(H + X_0(M^*)^{-1}X_0')\big) + \log\big(\det(H)\big) \\
&- \lambda^2(\mathbf{y} - X_0\beta^*)'(H + X_0(M^*)^{-1}X_0')^{-1}(\mathbf{y} - X_0\beta^*) \\
&+ \lambda^2\big(\text{tr}(H^{-1}X\Sigma_\beta X')\big) + \lambda^2\big((\mathbf{y} - X\beta^*)'H^{-1}(\mathbf{y} - X\beta^*)\big), \\
\\
&= -\log\big(\det(H + X_0(M^*)^{-1}X_0')\big) \\
&- \lambda^2(\mathbf{y} - X_0\beta^*)'(H + X_0(M^*)^{-1}X_0')^{-1}(\mathbf{y} - X_0\beta^*) \\
&+ \lambda^2\big(\text{tr}(H^{-1}X\Sigma_\beta X')\big) + \lambda^2\big((\mathbf{y} - X\beta^*)'H^{-1}(\mathbf{y} - X\beta^*)\big), \\
\\
&= -\log\big(\det(I_{n\times n} + (M^*)^{-1}J_{n\times n})\big) \\
&- \lambda^2(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*)'(I_{n\times n} + (M^*)^{-1}J_{n\times n})^{-1}(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*) \\
&+ \lambda^2\Sigma_\beta\big(\text{tr}(J_{n\times n})\big) + \lambda^2\sum_{i=1}^{n}(y_i - \beta^*)^2.
\end{aligned}
$$

The notation $J_{n\times n}$ is used to denote an $n \times n$ matrix of ones. Notice that in the verification above, β^* is used instead of $\boldsymbol{\beta}^*$. Since $p = 1$, β^* is scalar and not a vector. This also implies that M^*, a $p \times p$ matrix, is also scalar. To continue with the verification, define the following

$$
\begin{aligned}
\det(aI + bJ) &= (a + nb)a^{n-1}, \\
(aI + bJ)^{-1} &= \frac{1}{a}I - \frac{b}{a(a + nb)}J,
\end{aligned}
$$

where I is the identity matrix and J is a matrix of ones.

$$\begin{aligned}
p_{\text{WAIC1}} &= -\log\big(\det(I_{n\times n} + (M^*)^{-1}J_{n\times n})\big) \\
&\quad - \lambda^2(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*)'(I_{n\times n} + (M^*)^{-1}J_{n\times n})^{-1}(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*) \\
&\quad + \lambda^2\Sigma_\beta\big(\text{tr}(J_{n\times n})\big) + \lambda^2\sum_{i=1}^{n}(y_i - \beta^*)^2, \\
&= -\log\big(1 + n(M^*)^{-1}\big) + \lambda^2\Sigma_\beta n + \lambda^2\sum_{i=1}^{n}(y_i - \beta^*)^2 \\
&\quad - \lambda^2(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*)'\bigg(I_{n\times n} - \frac{(M^*)^{-1}}{1 + n(M^*)^{-1}}J_{n\times n}\bigg)(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*), \\
&= -\log\big(1 + n(M^*)^{-1}\big) + \lambda^2\Sigma_\beta n + \lambda^2\sum_{i=1}^{n}(y_i - \beta^*)^2 - \lambda^2\sum_{i=1}^{n}(y_i - \beta^*)^2 \\
&\quad + \lambda^2\bigg(\frac{(M^*)^{-1}}{1 + n(M^*)^{-1}}\bigg)(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*)'(J_{n\times n})(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*), \\
&= -\log\big(1 + n(M^*)^{-1}\big) + \lambda^2\Sigma_\beta n + \frac{\lambda^2(M^*)^{-1}}{1 + n(M^*)^{-1}}\bigg(\sum_{i=1}^{n}(y_i - \beta^*)\bigg)^2.
\end{aligned}$$

Note that in the verification above, when using the quadratic form definition $Z'AZ = \sum_i\sum_j z_i z_j a_{ij}$, if $a_{ij} = (J_{n\times n})_{ij}$, $(J_{n\times n})_{ij} = 1$ for all i and $j = 1, \ldots, n$. As a result, the double summation can be condensed to the square of one summation. However, $a_{ij} = (I_{n\times n})_{ij}$, $(I_{n\times n})_{ij} = 1$ only on the diagonal elements, i.e., when $j = i$ for $i = 1, \ldots, n$. As a result, the double summation can be condensed to one summation, but as the summation of the squared terms. In other words,

$$\begin{aligned}
(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*)'(J_{n\times n})(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*) &= \bigg(\sum_{i=1}^{n}(y_i - \beta^*)\bigg)^2, \\
(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*)'(I_{n\times n})(\mathbf{y} - \mathbf{1}_{n\times 1}\beta^*) &= \sum_{i=1}^{n}(y_i - \beta^*)^2.
\end{aligned}$$

Recall that $M^* = X'H^{-1}X + M$. This implies that $M^* = 2$ in context of the verification, which means

$$p_{\text{WAIC1}} = -\log\big(1+2n\big) + \lambda^2\Sigma_\beta n + \frac{2\lambda^2}{1+2n}\left(\sum_{i=1}^{n}(y_i-\beta^*)\right)^2. \tag{3.13}$$

This verification result should match the calculation of p_{WAIC1} in the univariate case, as shown by Gelman et al. (2014) and Sahu (2022), which is

$$p_{\text{WAIC1}} = n\log\left(\frac{\sigma^2}{\sigma^2+\sigma_p^2}\right) + n\frac{\sigma_p^2}{\sigma^2} + \frac{\sigma_p^2}{\sigma^2(\sigma^2+\sigma_p^2)}\sum_{i=1}^{n}(y_i-\mu_p)^2. \tag{3.14}$$

In the case of (3.14), the posterior mean and posterior variance are denoted as μ_p and σ_p^2 respectively, whereas the notation β^* and Σ_β are used in (3.13).

The verification of (3.12) is the following

$$\begin{aligned}
p_{\text{WAIC2}} &= -\frac{\lambda^4}{4}\left(2\text{tr}\big(H^{-1}(X\Sigma_\beta X')H^{-1}(X\Sigma_\beta X')\big) + 4(\mathbf{y}-X\boldsymbol{\beta}^*)'H^{-1}(X\Sigma_\beta X')H^{-1}(\mathbf{y}-X\boldsymbol{\beta}^*)\right)\\
&= \frac{\lambda^4}{2}\left((\Sigma_\beta)^2\text{tr}\big((J_{n\times n})(J_{n\times n})\big) + 2\Sigma_\beta(\mathbf{y}-\mathbf{1}_{n\times 1}\boldsymbol{\beta}^*)'(J_{n\times n})(\mathbf{y}-\mathbf{1}_{n\times 1}\boldsymbol{\beta}^*)\right),\\
&= \frac{\lambda^4}{2}\left((\Sigma_\beta)^2n^2 + 2\Sigma_\beta\left(\sum_{i=1}^{n}(y_i-\beta^*)\right)^2\right),\\
&= \frac{\lambda^4(\Sigma_\beta)^2}{2}n^2 + \lambda^4\Sigma_\beta\left(\sum_{i=1}^{n}(y_i-\beta^*)\right)^2.
\end{aligned} \tag{3.15}$$

Likewise, the verification result should match that of the univariate derivation by Sahu (2022) which is

$$p_{\text{WAIC2}} = n\frac{\sigma_p^4}{2\sigma^4} - \frac{\sigma_p^2}{\sigma^4}\sum_{i=1}^{n}(y_i-\mu_p)^2. \tag{3.16}$$

The verification results indicate a discrepancy between the univariate and multivariate outcomes. Specifically, (3.13) does not align with (3.14), and (3.15) does not align with (3.16). This suggests that, although our multivariate calculation strategy for the WAIC represents a novel approach, it may not be the appropriate method. The mismatches arise from the inherent differences in the partitioning and summation processes in the original formulation of the WAIC. In particular, the order of operations involving the posterior variance and summation in the context of p_{WAIC2}

differs between the multivariate and univariate cases. These differences are due to the distinct ways these models handle dependencies and correlations within the data. Consequently, the results are not expected to align, as they essentially measure different aspects of the data and model fit.

3.3 The calculation of the WAIC in INLA

From the technical aspect, the WAIC and its penalty component, which is referred to as the effective number of parameters, can be calculated in INLA by invoking the `control.compute = list(waic = T)` option within the `inla()` function within R. However, attempting to understand how the `inla()` function calculates the WAIC proves to be challenging. In our first attempt, we turn to the available INLA documentation (Martino and Rue, 2009) and online resources (Rue et al., 2013). Within the online resources, the developers refer to the the WAIC calculations described by Gelman et al. (2014). Gelman et al. provides their definition of the WAIC as

$$\text{WAIC} = -2(\widehat{\text{elppd}}_{\text{WAIC}}), \tag{3.17}$$

$$\widehat{\text{elppd}}_{\text{WAIC}} = \text{lppd} - p_{\text{WAIC}}, \tag{3.18}$$

where $\widehat{\text{elppd}}_{\text{WAIC}}$ denotes the estimate of the expected log pointwise predictive density for a new dataset (Gelman et al., 2014). The lppd denotes the log pointwise predictive density and is given as

$$\text{lppd} = \sum_{i=1}^{n} \log \int p(y_i|\boldsymbol{\theta}) p_{\text{post}}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{3.19}$$

where $p_{\text{post}}(\cdot)$ denotes the probability that averages over the posterior distribution. Additionally, the developers of INLA explicitly mentioned that the effective number of parameters are calculated following Equation (11) in Gelman et al. (2014), which is

$$p_{\text{WAIC2}} = \sum_{i=1}^{n} \text{Var}_{\text{post}}\big(\log p(y_i|\boldsymbol{\theta})\big),$$

where $\text{Var}_{\text{post}}(\cdot)$ denotes the posterior variance. However, beyond these information, there are not much other useful information that can aid our understanding on the explicit calculation of the WAIC within the `inla()` function. Particularly, we are interested to understand how INLA obtain the $\log p(y_i|\boldsymbol{\theta})$ in the p_{WAIC2} equation mentioned above.

In our second attempt, we delve into the R code and try to understand how the WAIC is explicitly calculated internally within the `inla()` function. We found the following,

$$\begin{aligned}\texttt{waic} &= -2(\texttt{sum}(\log(\texttt{po.res} - \texttt{sum}(\texttt{po2.res})))),\\ \texttt{p.eff} &= \texttt{sum}(\texttt{po2.res}).\end{aligned}$$

Although the expressions above are code representation of the formula, they provide enough information to infer that

$$\begin{aligned}\texttt{po.res} &= \sum_{i=1}^{n} \log \int p(y_i|\boldsymbol{\theta}) p_{\text{post}}(\boldsymbol{\theta}) d\boldsymbol{\theta},\\ \texttt{po2.res} &= \sum_{i=1}^{n} \text{Var}_{\text{post}}\big(\log p(y_i|\boldsymbol{\theta})\big).\end{aligned}$$

The calculations above are performed in the background of the `inla()` function whenever the user requires the WAIC to be calculated. Ideally, the derivations of `po.res` and `po2.res` could be further investigated. However, they require components that are read in from binary files, and the explicit calculations are tucked under layers of source code.

In our investigation on the code of the WAIC within the `inla()` function, we found an interesting comment under the calculation of the WAIC that reads, “yes, here we use the po results” (Rue et al., 2014). In the context of cross-validation methods, we hypothesise that “po” denotes predictive ordinates.

Although there are no information specifically on “predictive ordinates”, there are plenty of literature on the conditional predictive ordinate (CPO). However, there is an option to calculate the CPO in the `inla()` function, and we found that the calculation of the CPO from the code is not entirely the same as `po.res` and `po2.res`. Furthermore, calculating the CPO within the `inla()` function also requires components read in from binary files, which presents a difficulty when attempting to understand if CPO is related to “po”. Further investigation into the `po.res` and `po2.res` is required to fully understand how the WAIC is explicitly calculated within the `inla()` function.

We previously mentioned that the `inla()` function calculates the conditional predictive ordinate (CPO). The CPO is a cross-validation method that estimates the leave-one-out predictive distribution and measures the predictive ability of the fitted model (Pettit, 1990, Gelfand et al., 1992, Geisser, 1993, Gelfand and Dey, 1994).

For $i = 1, \ldots, n$, the CPO for observation y_i is given as

$$CPO_i = p(y_i|\mathbf{y}_{-i}) = \int_{-\infty}^{\infty} p(y_i|\mathbf{y}_{-i}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}_{-i}) d\boldsymbol{\theta},$$

where $\mathbf{y}_{-i}$ denotes the all observations without y_i, and

$$CPO = \prod_{i=1}^{n} CPO_i.$$

A larger CPO_i value indicates better adjustment, meaning that y_i is very likely to be under the current model. Conversely, a smaller CPO_i value indicates that y_i is likely an outlier, or a high leverage observation (Pettit, 1990). A model with a larger CPO value suggests better predictive performance. Hence, the candidate model with the largest CPO value is the preferred model. However, it is noted that CPO values are often close to zero, therefore an alternative criteria is often employed (Draper and Krnjajic, 2006, Cai et al., 2013). The $NLLK_{\text{CV}}$ represents the negative cross-validatory log likelihood function, and is given as

$$NLLK_{CV} = -\sum_{i=1}^{n} \log CPO_i.$$

Since a large CPO value indicate agreement between the observation and the model, a model with a smaller $NLLK_{CV}$ value implies a better fit. The candidate model with the smallest $NLLK_{CV}$ value has the best predictive performance (Cai et al., 2013, Ayalew et al., 2021).

Calculating CPO_i requires refitting the model each time up to n times. Furthermore, the closed form of CPO_i is usually unavailable (Cai et al., 2013). Instead, we can obtain an approximation from the Monte Carlo estimates of CPO_i through the MCMC samples from the posterior distribution,

$$\widehat{CPO}_i = \left(\frac{1}{S} \sum_{s=1}^{S} \frac{1}{p(y_i|\boldsymbol{\theta}^{(s)})} \right)^{-1},$$

where $\boldsymbol{\theta}^{(s)}$ is a sample from the posterior distribution $p(\boldsymbol{\theta}|y_i)$, and S indicates the total number of posterior samples (Gelfand et al., 1992, Cai et al., 2013).

While the CPO involves a harmonic mean which yields a numerically unstable estimator in practical applications (Hooten and Hobbs, 2015), the `inla()` function is able to flag problematic cases; see Held et al. (2010). However, numerical problems may still occur when the CPO are computed by the INLA method. Specifically, some of the CPO values may not be reliable due to numerical problems when evaluating $p(y_i|\mathbf{y}_{-i})$. INLA evaluates the CPO with a numerical integration step which involves the full conditional component (Held et al., 2010). However, the INLA method approximates this full conditional component with either the Gaussian approximation, Laplace approximation, or the simplified Laplace approximation based on the skew-normal distribution (Azzalini and Capitanio, 1999, Martino and Rue, 2009, Rue et al., 2009), as detailed in Section 2.6. Essentially, the accuracy of the numerical integration is dependent on the accuracy of the approximation.

3.4 The calculation of the WAIC and PSIS-LOOIC in Stan

The WAIC and PSIS-LOOIC can be calculated with the `loo` package (Vehtari et al., 2017a) after fitting models with Stan, using the `sampling()` function in the `rstan` package (Stan Development Team, 2020). More specifically, a model is first written in the Stan language, then compiled in R. Afterwards the `sampling()` function draws MCMC samples from the model. In the initial code of the model, which is written in Stan language, a generated quantities block have to be included in order to calculate the WAIC. Within this generated quantities block, the log likelihood is calculated. The `extract_log_lik()` function within the `loo` package can be used to extract the calculated log likelihood. Finally, using this extracted log likelihood, the WAIC and PSIS-LOOIC can be calculated with the `waic()` and `loo()` functions in the `loo` package. The `waic()` and `loo()` functions require an $S \times n$ matrix with elements $\log p(y_i|\theta^{(s)})$, for $i = 1, \ldots, n$ and $s = 1, \ldots, S$, where S denotes the total number of MCMC samples, and n denotes the number of sites.

The WAIC (3.17) is calculated from the expected log pointwise predictive density (3.18), which requires a component represented by the the log pointwise predictive density (3.19) and a penalty component (2.25) (Gelman et al., 2014, Vehtari et al., 2017b). In practical applications, the two components can be calculated from the MCMC samples of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$, given as

$$\begin{aligned} \text{lppd} &= \sum_{i=1}^{n} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(y_i|\boldsymbol{\theta}^{(s)})\right), \\ p_{\text{WAIC2}} &= \sum_{i=1}^{n} V_{s=1}^{S}\big(\log p(y_i|\boldsymbol{\theta}^{(s)})\big), \end{aligned}$$

where $V_{s=1}^{S}(\cdot)$ represents the sample variance.

The PSIS-LOOIC is calculated using Equation (2.31). It requires the Pareto smoothed truncated weights, which is calculated by following the procedure described in Section 2.7.3. Additionally, the penalty component of the PSIS-LOOIC can be derived as

$$\begin{aligned} p_{\text{PSIS}-\text{LOOIC}} &= \widehat{\text{elppd}}_{\text{PSIS}-\text{LOO}} - \text{lppd}, \\ \widehat{\text{elppd}}_{\text{PSIS}-\text{LOO}} &= -\text{PSIS-LOOIC}/2, \end{aligned}$$

where lppd denotes log pointwise predictive density (3.19).

The calculation for the WAIC and PSIS-LOO are straightforward and easily implemented with the `loo` package in R. However, the real challenge, as prompted in the beginning of this chapter, have not been addressed. In the following section, we will introduce our approach to calculate the WAIC and PSIS-LOOIC.

3.5 Calculating the WAIC and PSIS-LOOIC with non-factorisable models likelihoods

Recall from Section 3.1 that a non-factorisable model is one where the the full likelihood cannot be expressed as

$$p(\mathbf{y}|\psi) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\psi}),$$

where $\boldsymbol{\psi}$ denotes some model parameters and $\mathbf{y} = (y_1, \ldots, y_n)'$. A non-factorisable model also implies that the observations y_i are conditionally independent of one another. Models for point referenced spatial data are non-factorisable models, since we expect point referenced spatial data to exhibit conditional independence, where nearby points tend to influence each other. As mentioned in Section 3.1, models of structured data, such as point referenced spatial data, are often characterised by multivariate normal distributions with some structured covariance matrix $\boldsymbol{\Sigma}$ that does not factorise.

According to multivariate normal theory, the conditional distribution for the ith observation, denoted as $p(y_i|\mathbf{y}_{-i}, \boldsymbol{\theta})$, is univariate normal with mean $\tilde{\mu}_i$ and variance $\tilde{\sigma}_i$,

$$\begin{aligned}
\tilde{\mu}_i &= \mu_i + \sigma_{i,-i}\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{y}_{-i} - \boldsymbol{\mu}_{-i}), \\
\tilde{\sigma}_i &= \sigma_{ii} + \sigma_{i,-i}\Sigma_{-i}^{-1}\sigma_{-i,i},
\end{aligned}$$

where $\mathbf{y}_{-i}$ denotes all observations without the ith observation. Additionally, $\sigma_{i,-i}$ and $\sigma_{-i,i}$ denote the ith row and column vectors of $\boldsymbol{\Sigma}$ without the ith element, σ_{ii} denotes the ith diagonal element of $\boldsymbol{\Sigma}$, Σ_{-i}^{-1} denotes the inverse of the covariance matrix without the ith row and column, and $\boldsymbol{\mu}_{-i}$ denotes the mean vector without the ith element. The pointwise log likelihood is then given as

$$\log p(y_i|\mathbf{y}_{-i}, \boldsymbol{\theta}) = -\frac{1}{2}\log(2\pi\tilde{\sigma}_i) - \frac{1}{2}\frac{(y_i - \tilde{\mu}_i)^2}{\tilde{\sigma}_i}. \tag{3.20}$$

However, calculating $\tilde{\mu}_i$ and $\tilde{\sigma}_i$ this way is computationally expensive and inefficient, since $\boldsymbol{\Sigma}$ must be computed for each i, and may further increase computation cost depending on its structure. Instead, Bürkner et al. suggest the following,

$$\tilde{\mu}_i = y_i - \frac{g_i}{\bar{\sigma}_{ii}}, \tag{3.21}$$

$$\tilde{\sigma}_i = \frac{1}{\bar{\sigma}_{ii}}, \tag{3.22}$$

where

$$\begin{aligned}
g_i &= [\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu})]_i, \\
\bar{\sigma}_{ii} &= [\Sigma^{-1}]_{ii}
\end{aligned}$$

(Bürkner et al., 2021), and likewise, the pointwise conditional log likelihood is given by Equation (3.20). This way of calculating $\tilde{\mu}_i$ and $\tilde{\sigma}_i$ is more computationally efficient since $\mathbf{\Sigma}$ is calculated and inverted once only and can be reused for each i.

The derivations of (3.21) and (3.22) are based on Lemma 1 of the work by Sundarajan and Keerthi (2001), and documented in their supplementary materials. Sundarajan and Keerthi demonstrated that, for any finite subset z of a zero-mean Gaussian process with covariance matrix $\mathbf{\Sigma}$, the LOO predictive mean and standard deviation can be computed as follows,

$$\begin{aligned} \tilde{\mu}_i &= z_i - \frac{g_i}{\bar{\sigma}_{ii}}, \\ \tilde{\sigma}_i &= \frac{1}{\bar{\sigma}_{ii}}, \end{aligned}$$

where

$$\begin{aligned} g_i &= [\Sigma^{-1} z]_i, \\ \bar{\sigma}_{ii} &= [\Sigma^{-1}]_{ii}. \end{aligned}$$

It is noteworthy that this proof does not rely on any specific form of the covariance matrix $\mathbf{\Sigma}$, implying its applicability to all zero-mean multivariate normal distributions (Sundarajan and Keerthi, 2001). Following this, if $\mathbf{y}$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Sigma}$, then $(\mathbf{y} - \boldsymbol{\mu})$ also follows a multivariate normal distribution with covariance matrix $\mathbf{\Sigma}$, but with zero-mean. Hence, $(\mathbf{y} - \boldsymbol{\mu})$ can be used in place of z in the proof by Sundarajan and Keerthi described above. In cases where $(y_i - \mu_i)$ has a LOO mean $(y_i - \mu_i) - \frac{g_i}{\bar{\sigma}_{ii}}$, it follows that the LOO mean of y_i is $y_i - \frac{g_i}{\bar{\sigma}_{ii}}$ (Bürkner et al., 2021).

To facilitate the implementation of (3.20), (3.21), and (3.22), we present the following algorithm. The algorithm's syntax closely resembles that of the R programming language. It is essential to note that, before applying this algorithm, one should already possess the MCMC samples of the unknown parameters of interest. These samples can be obtained, for instance, by fitting the model in Stan, as detailed in Section 2.5.3. The inputs of this algorithm comprise the MCMC samples of the unknown parameters of interest, a vector of the observations denoted as $\mathbf{y}$, and the total number of observations denoted as n. The output of this algorithm is a $S \times n$ matrix of non-factorisable model log likelihood values, where S denotes the total number of

MCMC samples and n denotes the total number of observations.

Algorithm 1 Non-factorisable model log likelihood

```
 1: for it in 1 to imax do
 2:
 3:     sigmasq ← i_sigmasq[it]
 4:     tausq ← i_tausq[it]
 5:     phi ← i_phi[it]
 6:
 7:     MC1 ← 2^(1 − nu)/gamma(nu)
 8:     MC2 ← (sqrt(2 ∗ nu) ∗ phi ∗ D)^nu
 9:     MC3 ← besselK(sqrt(2 ∗ nu) ∗ phi ∗ D, nu = nu)
10:     Sigma ← sigmasq ∗ MC1 ∗ MC2 ∗ MC3 + diag(tausq, nrow = n, ncol = n)
11:
12:     Qmat ← solve(Sigma)
13:     meanvec ← as.numeric(ps_xbetas[it, ])
14:     meanmult ← diag(1/diag(Qmat), nrow = n, ncol = n) %*% Qmat
15:     condmean ← y − meanmult %*% (y − meanvec)
16:     condvar ← 1/diag(Qmat)
17:
18:     loglik[it, ] ← dnorm(y, mean = condmean, sd = sqrt(condvar), log = T)
19:
20: end for
```

Algorithm 1 represents a comprehensive set of instructions to be executed iteratively until reaching the total number of MCMC samples, denoted as `imax` or equivalently as S. In the first part of Algorithm 1 (lines 3-5), the algorithm retrieves the current MCMC samples from the unknown parameters of interest. Specifically, these parameters include the spatial variance denoted as σ_{ω}^{2}, the independent and identically distributed variance represented by τ^2, and the spatial decay parameter ϕ. Their respective vectors of MCMC samples are stored as `i_sigmasq`, `i_tausq` and `i_phi`. The values corresponding to the `it` iteration are then extracted and stored as `sigmasq`, `tausq` and `phi`, respectively.

Using `sigmasq`, `tausq` and `phi`, we compute the covariance function, denoted as `Sigma` within the algorithm. In lines 7-10 of Algorithm 1, we demonstrate the computation of the Matérn covariance function (2.6), which was previously introduced in Section 2.3.3. Following this, we calculate the inverse of the resulting covariance matrix, denoted as `Qmat`.

In line 13 of Algorithm 1, we extract the `it` row of `ps_xbetas`. Here, `ps_xbetas` represents the transposed product of the design matrix X of dimensions $n \times p$ and the vector of regression coefficients $\boldsymbol{\beta}$. The regression coefficients $\boldsymbol{\beta}$ are also unknown parameters of interest. After obtaining the MCMC samples, $\boldsymbol{\beta}$ extends to a $p \times S$ matrix, where p denotes the number of variables included in the model and S denotes the total number of MCMC samples. Combining both components and transposing the resulting matrix gives a $S \times n$ matrix, denoted as `ps_xbetas`.

In lines 14-15 of Algorithm 1, we implement the calculation of (3.21), denoted as `condmean` within the algorithm. Subsequently, in line 16, we implement the calculation of (3.22), denoted as `condvar` within the algorithm. Finally, in line 18 of Algorithm 1, we implement (3.20). This completes the `it` iteration out of `imax`.

The resulting `loglik` object obtained from Algorithm 1 is a $S \times n$ matrix. This `loglik` matrix assumes an important role in the computation of what we have termed the WAIC_{NF} and the $\text{PSIS-LOOIC}_{\text{NF}}$. More explicitly, this `loglik` matrix finds application in (2.26) and (2.31). In Chapter 4, we further investigate the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ calculated with the `loglik` matrix from Algorithm 1. Specifically, we compare the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ against the WAIC computed by INLA (Rue et al., 2009, Martins et al., 2013), and the WAIC and PSIS-LOOIC computed from the log likelihoods extracted from Stan (Stan Development Team, 2020).

Chapter 4

Simulation examples

In Chapter 3, we introduced two novel model selection criteria for Bayesian models for point referenced spatial data: the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$. These criteria are computed using the non-factorisable model log likelihoods, as detailed in Algorithm 1 within Section 3.5. In this chapter, our primary objective is to conduct a comparative analysis between the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ and established alternatives. These alternatives include the WAIC computed by INLA (Rue et al., 2009, Martins et al., 2013), as well as the WAIC and PSIS-LOOIC computed from the log likelihoods extracted from Stan (Stan Development Team, 2020). Our investigation will encompass both model selection tasks and variable selection tasks.

4.1 Data simulation

$$\begin{aligned} Y(\mathbf{s}_i) &= \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \omega(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \\ \boldsymbol{\omega} &\sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}), \\ \epsilon(\mathbf{s}_i) &\sim N(0, \tau^2). \end{aligned} \tag{4.1}$$

The point referenced spatial data are generated following the model above. The locations of the point referenced spatial data, denoted as $\mathbf{s}_i,\ i = 1, \ldots, n$, are generated from a unit square. On the right-hand side of (4.1), we have a linear combination that incorporates the covariates, their associated regression coefficients, the spatial random effect and the independent and identically distributed (iid) random effect. The spatial random effects, denoted as $\boldsymbol{\omega} = \big(\omega(\mathbf{s}_1), \ldots, \omega(\mathbf{s}_n)\big)'$, is generated using a zero-mean Gaussian Process (GP). To elaborate, this GP assumes the form of an n-dimensional multivariate normal distribution with mean-zero and an $n \times n$ covariance matrix denoted as $\mathbf{\Sigma}_{\boldsymbol{\omega}}$. The iid random effect $\epsilon(\mathbf{s}_i)$ is generated from a normal distribution with mean-zero and a specified variance parameter denoted as τ^2.

The covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\omega}}$ within the spatial random effect comprise a spatial variance parameter $\sigma^2_{\boldsymbol{\omega}}$ and a correlation matrix component, as detailed in Section 2.3.3. The elements within the correlation matrix are computed through a function that incorporates additional parameters. Specifically, when employing a Matérn function, the additional parameters include the smoothness parameter denoted as ν and the spatial decay parameter denoted as ϕ. More explicitly, the Matérn covariance function (2.6) is used to calculate the elements of $\mathbf{\Sigma}_{\boldsymbol{\omega}}$, and is given as

$$C(d) = \sigma^2 \frac{(\sqrt{2\nu} d\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu} d\phi)$$

for $d > 0$, where $\Gamma(\cdot)$ represents the standard mathematical Gamma function, $K_\nu(\cdot)$ denotes the modified Bessel function of the second kind (Abramowitz and Stegun, 1948) of order ν, and $\sigma^2_{\boldsymbol{\omega}}$ denotes the spatial variance parameter. The parameter ν is referred to as the smoothness parameter as it determines the smoothness of the covariance function, and ϕ is referred to as the spatial decay as it dictates the rate of decay as d increases. Additionally, d denotes the Euclidean between the locations of the point referenced spatial data.

4.1.1 Model selection task: covariance function selection

In this chapter, we conduct model selection tasks with the aim of comparing our proposed model selection criteria, WAIC$_{\mathrm{NF}}$ and PSIS-LOOIC$_{\mathrm{NF}}$ as introduced in Section 3.5, against the WAIC computed by INLA, as well as the WAIC and PSIS-LOOIC computed from the log likelihoods extracted from Stan. The primary focus of our model selection task centres around covariance function selection.

The outline for the design of the covariance function selection is as follows: We generate point referenced spatial data, following the procedure outlined in Section 4.1. The elements within the covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\omega}}$ are calculated based on the specified covariance functions listed below. Subsequently, Bayesian models are constructed for the generated data, utilising the candidate covariance functions from the same list.

These models are fitted using both the INLA method and within the Stan framework. Following this, we compute the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, as well as the WAIC by INLA, and the WAIC and PSIS-LOOIC from the log likelihoods extracted from Stan for all the candidate models. We determine the best candidate model based on the magnitude of the selection criteria. Finally, we ascertain whether the selected candidate model is constructed using the covariance function that aligns with the covariance function employed in generating the data, thereby shedding light on the efficacy of the selection criteria within the context of covariance function selection.

The list of candidate covariance functions is the following:

Exponential covariance function

The Matérn covariance function given above can be simplified if we set the smoothing parameter $\nu = 1/2$. The simplified covariance function is given as

$$\mathbf{\Sigma}_{\boldsymbol{\omega}} = \begin{cases} \sigma^2_{\boldsymbol{\omega}} \exp(-\phi d) & \text{if } d > 0 \\ \tau^2 + \sigma^2_{\boldsymbol{\omega}} & \text{if } d = 0, \end{cases}$$

where $\sigma^2_{\boldsymbol{\omega}}$ is the spatial variance parameter, τ^2 is the nugget, ϕ is the spatial decay parameter and d denotes the Euclidean distance between the sites $\mathbf{s}_i$, for $i = 1, \ldots, n$. The exponential covariance function is widely used because of its simplicity and easy interpretation with the effective range $r \approx 3/\phi$ as described in Section 2.3.3.

Matérn ($\nu = 3/2$) covariance function

The Matérn covariance function also simplifies to a nice form when we set the smoothing parameter $\nu = 3/2$. The covariance function is given as

$$\mathbf{\Sigma}_{\boldsymbol{\omega}} = \begin{cases} \sigma^2_{\boldsymbol{\omega}} (1 + \sqrt{3}\phi d) \exp(-\sqrt{3}\phi d) & \text{if } d > 0 \\ \tau^2 + \sigma^2_{\boldsymbol{\omega}} & \text{if } d = 0, \end{cases}$$

where $\sigma^2_{\boldsymbol{\omega}}$ is the spatial variance parameter, τ^2 is the nugget, ϕ is the spatial decay parameter, and d denotes the Euclidean distance between sites $\mathbf{s}_i$, for $i = 1, \ldots, n$. Compared to the exponential covariance function, the Matérn ($\nu = 3/2$) covariance function decreases at a slower rate as the distance d increases. As a result, it is useful for capturing more moderate spatial correlations over intermediate distances.

Matérn ($\nu = 5/2$) covariance function

The Matérn covariance function also simplifies to a nice form when we set the smoothing parameter $\nu = 5/2$. The covariance function is given as

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \begin{cases} \sigma^2_{\boldsymbol{\omega}}(1 + \sqrt{5}\phi d + \frac{5}{3}\phi^2 d^2)\exp(-\sqrt{5}\phi d) & \text{if } d > 0 \\ \tau^2 + \sigma^2_{\boldsymbol{\omega}} & \text{if } d = 0, \end{cases}$$

where $\sigma^2_{\boldsymbol{\omega}}$ is the spatial variance parameter, τ^2 is the nugget, ϕ is the spatial decay parameter and d denotes the Euclidean distance matrix. The Matérn ($\nu = 5/2$) covariance function decreases at an even slower rate compared to the exponential covariance function and the Matérn ($\nu = 3/2$) covariance function as d increases. The Matérn ($\nu = 5/2$) covariance function captures more long-range spatial dependence.

Spherical covariance function

The spherical covariance function we use as a candidate covariance function in this chapter follows the definition by Banerjee et al. (2014), and is given

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \begin{cases} 0 & \text{if } d \geq 1/\phi \\ \sigma^2_{\boldsymbol{\omega}}\left(1 - \frac{3}{2}\phi d + \frac{1}{2}(\phi d)^3\right) & \text{if } 0 < d < 1/\phi \\ \tau^2 + \sigma^2_{\boldsymbol{\omega}} & \text{if } d = 0, \end{cases}$$

where $\sigma^2_{\boldsymbol{\omega}}$ is the spatial variance parameter, τ^2 is the nugget, ϕ is the spatial decay parameter and d is the Euclidean distance between sites $\mathbf{s}_i$, for $i = 1, \ldots, n$.

Gaussian covariance function

The Gaussian covariance function we use as a candidate covariance function in this simulation study follow the definition by Banerjee et al. (2014), and is given as

$$\boldsymbol{\Sigma}_{\boldsymbol{\omega}} = \begin{cases} \sigma^2_{\boldsymbol{\omega}}\exp\big(-(\phi d)^2\big) & \text{if } d > 0 \\ \tau^2 + \sigma^2_{\boldsymbol{\omega}} & \text{if } d = 0, \end{cases}$$

where $\sigma^2_{\boldsymbol{\omega}}$ is the spatial variance parameter, τ^2 is the nugget, ϕ is the spatial decay parameter and d denotes the Euclidean distance between sites $\mathbf{s}_i$, for $i = 1, \ldots, n$.

4.1.2 Variable selection task

We employ another approach to compare our proposed model selection criteria, the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, against the WAIC computed by INLA, as well as the WAIC and PSIS-LOOIC computed from the log likelihoods extracted from Stan. This approach involves variable selection tasks.

The design of the variable selection task is as follows: We generate point referenced data, following the procedure outlined in Section 4.1. Within the covariates on the right-hand side of (4.1), we specify a combination from the list of covariates combination provided below. Following this, Bayesian models are constructed using the list of combinations of covariates. These models are then fitted using both the INLA method and within the Stan framework. Subsequently, we compute the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, as well as the WAIC by INLA, and the WAIC and PSIS-LOOIC from the log likelihoods extracted from Stan. We select the best candidate model based on the magnitude of the selection criteria. Finally, we ascertain whether the selected candidate model is constructed using the combination of covariates that aligns with the combination of covariates employed in generating the data. This investigation provides insights into the effectiveness of the selection criteria within the context of variable selection.

The covariates are generated as follows,

$$\begin{aligned}
x_1(\mathbf{s}_i) &= 1, \\
x_2(\mathbf{s}_i) &\sim N(0,1), \\
x_3(\mathbf{s}_i) &\sim N(0,1), \\
x_4(\mathbf{s}_i) &\sim N(0,2),
\end{aligned}$$

for all i, where $N(\cdot)$ denotes the normal distribution. Note that the covariates could have been generated as other types, such as binary, gamma-distributed, or heavy-tailed covariates. These alternative distributions are important for applications where they more accurately represent the underlying data. However, for the purposes of this thesis, we focused on Gaussian covariates to maintain consistency with common geospatial modelling practices. We further apply the Gram-Schmidt orthogonalisation process (Cheney and Kincaid, 2009) to the generated covariates to ensure that they are independent; see Appendix D for the explicit calculations of this process.

The point referenced data are generated using one of the following combination of covariates, and the candidate models are subsequently constructed and fitted using these combinations of covariates,

$$\begin{aligned}
\text{f1}: \quad & \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} = x_1(\mathbf{s}_i)\beta_1 + x_2(\mathbf{s}_i)\beta_2, \\
\text{f2}: \quad & \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} = x_1(\mathbf{s}_i)\beta_1 + x_2(\mathbf{s}_i)\beta_2 + x_3(\mathbf{s}_i)\beta_3, \\
\text{f3}: \quad & \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} = x_1(\mathbf{s}_i)\beta_1 + x_2(\mathbf{s}_i)\beta_2 + x_3(\mathbf{s}_i)\beta_3 + x_4(\mathbf{s}_i)\beta_4,
\end{aligned}$$

where we set $\beta_1 = 1,\ \beta_2 = 2,\ \beta_3 = 2,\ \beta_4 = 2$.

4.2 Simulation design

In each dataset we generate, following the procedure outlined in Section 4.1, we specify the following: the sample size n of the simulated dataset, the smoothness parameter ν, the spatial variance parameter $\sigma^2_{\boldsymbol{\omega}}$, the spatial decay parameter ϕ, the variance parameter τ^2, the covariance function of $\Sigma_{\boldsymbol{\omega}}$, as detailed in Section 4.1.1, and the covariates, as detailed in Section 4.1.2. To illustrate, Figure 4.1 provides an example of a simulated dataset with sample size of $n = 30$. This dataset is constructed using covariates from the f1 configuration, and using parameters $\sigma^2_{\boldsymbol{\omega}} = 3$, $\tau^2 = 3$ and $\nu = 1/2$, implying the utilisation of an exponential covariance function.

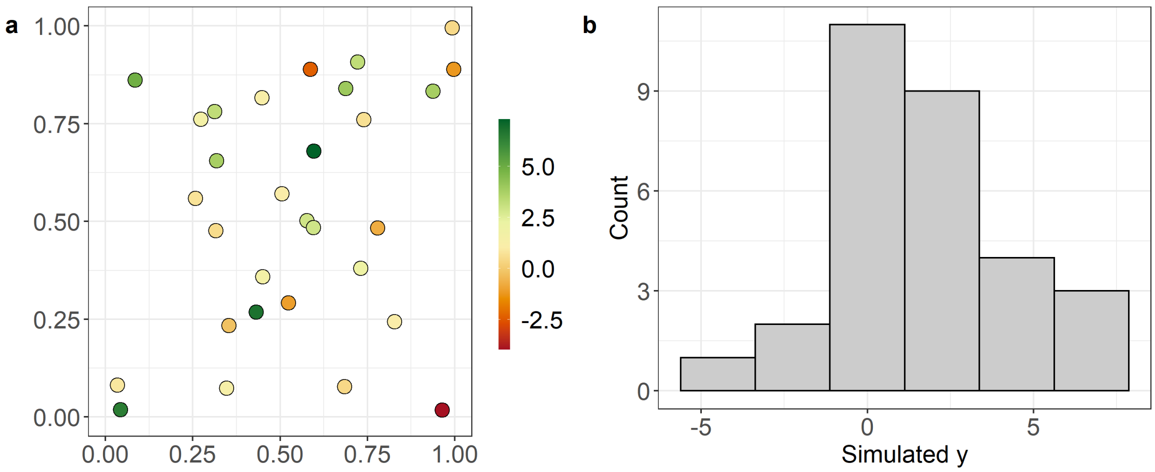

Figure 4.1. Example of point referenced data generated from a (a) unit square, and (b) the histogram of the generated responses.

To facilitate comparison, we generate a total of 100 datasets, using identical configurations. We proceed to construct and fit Bayesian models for these generated datasets. Subsequently, we calculate the non-factorisable model (NF) log likelihoods, following Algorithm 1 as detailed in Section 3.5. We then utilise these log likelihoods to compute both the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$.

When utilising Stan, our procedure involves computing the log likelihoods within the generated quantities code block within Stan. These log likelihoods are then extracted using the `extract_log_lik()` function within the `loo` package. Finally we calculate the WAIC and PSIS-LOO with the `waic()` and `loo()` functions, respectively, also found within the `loo` package. In the case of INLA, it inherently calculates the WAIC, allowing us to extract the WAIC value for comparison against our WAIC_{NF}.

When we compare the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ against the WAIC and PSIS-LOOIC from Stan and INLA, the best candidate model is determined based on the smallest value of the selection criteria. Out of the 100 generated datasets, we ascertain whether the selection criteria correctly identify the candidate model with the generating configurations.

When using the `inla.spde2.pcmatern()` function to use the SPDE approach in INLA, the specification of the smoothness parameter ν is restricted to a limited range. The `alpha` argument within the `inla.spde2.pcmatern()` function corresponds to $\nu = \alpha - d/2$. For point referenced spatial datasets, where $d = 2$, this relationship simplifies to $\nu = \alpha - 1$. For instance, if we intend to fit the model using an exponential covariance function (recalling that a Matérn covariance function with smoothness parameter $\nu = 1/2$ is equivalent to the exponential covariance function), we will need to set `alpha = 3/2` within the `inla.spde2.pcmatern()` function. However, INLA currently restricts the acceptable range for `alpha` to $0 <$ `alpha` < 2, which limits the range of ν values that can be specified.

All simulations were conducted in Stan (Stan Development Team, 2020) using two chains, each with a total of 2000 iterations, including 1000 burn-in iterations. These simulation experiments were executed on a Windows machine equipped with a 4-core Intel processor and 8GB random-access memory (RAM). The software environment employed for the simulations was R version 4.0.4 (R Core Team, 2021).

4.3 Simulation results

Figure 4.2 presents the results of the model selection tasks, specifically pertaining to covariance function selection, as outlined in Section 4.1.1. In Figure 4.2a, we generated point referenced spatial datasets using $n = 10$, spatial variance $\sigma_{\boldsymbol{\omega}}^2 = 3$, iid variance $\tau^2 = 3$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 1/2$ for a Matérn covariance function, which corresponds to the utilisation of an exponential covariance function (2.5). In this experiment, we considered candidate models employing the Gaussian, spherical and exponential covariance functions, with the latter being the one used to generate the datasets. These candidate models were fitted using the Stan framework.

Moving to Figure 4.2b, point referenced spatial datasets were generated using an increased sample size of $n = 15$, spatial variance $\sigma_{\boldsymbol{\omega}}^2 = 2$, iid variance $\tau^2 = 2$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 3/2$ for a Matérn covariance function. In this experiment, we considered candidate models using Matérn covariance functions and smoothness parameters $\nu = 1/2$, $\nu = 5/2$ and $\nu = 3/2$, with the latter being consistent with the smoothness parameter used in generating the datasets. The candidate models were also fitted within the Stan framework.

Similarly, in Figure 4.2c, point referenced spatial datasets were generated using $n = 15$, spatial variance $\sigma_{\boldsymbol{\omega}}^2 = 3$, iid variance $\tau^2 = 3$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 5/2$ for a Matérn covariance function. As in the previous experiment (Figure 4.2b), the candidate models utilised Matérn covariance functions and smoothness parameters $\nu = 1/2$, $\nu = 3/2$ and $\nu = 5/2$, with the latter aligning with the smoothness parameter used to generate the datasets. Once again, the candidate models were fitted using the Stan framework.

Figure 4.3 presents the results from the variable selection tasks detailed in Section 4.1.2. In Figure 4.3a, point referenced spatial datasets were generated using $n = 10$, spatial variance $\sigma_{\boldsymbol{\omega}}^2 = 3$, iid variance $\tau^2 = 3$ and smoothness parameter $\nu = 1/2$ for a Matérn covariance function. The selected covariate combination for this experiment was covariate combination f2. The candidate models incorporated covariate combinations f1, f2 and f3, and were fitted within the Stan framework.

In Figure 4.3b, point referenced spatial datasets were generated using an increased sample size $n = 15$, spatial variance $\sigma_{\boldsymbol{\omega}}^2 = 3$, iid variance $\tau^2 = 3$ and smoothness parameter $\nu = 3/2$ for a Matérn covariance function. Similar to the previous experiment, the selected covariate combination was f2, and the candidate models considered covariate combinations f1, f2 and f3, all fitted within the Stan framework.

Lastly, Figure 4.3c corresponds an experiment where the point referenced spatial datasets were generated using $n = 10$, spatial variance $\sigma^2_{\boldsymbol{\omega}} = 2$, iid variance $\tau^2 = 2$, smoothness parameter $\nu = 3/2$ for a Matérn covariance function, and the chosen covariate combination was f1. Once again, the candidate models included covariate combinations f1, f2, and f3, and they were fitted within the Stan framework. Additionally, all simulation experiments in Figure 4.3 utilised the spatial decay parameter $\phi = 3/0.5$ in their data generation.

There are several noteworthy observations when examining the simulation experiment results from the selected candidate models. Figure 4.2 shows that the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, denoted as NF WAIC and NF PSIS-LOOIC respectively within the figures, consistently outperform the WAIC and PSIS-LOOIC computed using the log likelihoods extracted from Stan when tasked with identify the covariance function employed in generating the datasets. This distinction is particularly evident in Figures 4.2b and 4.2c, where the WAIC and PSIS-LOOIC computed using the log likelihoods extracted from Stan tend to favour the candidate model fitted using the Matérn covariance function and smoothness parameter $\nu = 1/2$, when the datasets were originally generated using the Matérn covariance function and smoothness parameters $\nu = 3/2$ (Figure 4.2b) and $\nu = 5/2$ (Figure 4.2c).

However, we observe from Figure 4.3 that neither the WAIC_{NF}, the $\text{PSIS-LOOIC}_{\text{NF}}$, nor the WAIC and PSIS-LOOIC computed from the log likelihoods extracted from Stan perform well in identifying the covariate combination employed in data generation. An exception to this observation occurs when the datasets were generated using covariate combination f1. In such cases, both the WAIC and PSIS-LOOIC computed using the NF log likelihood from Algorithm 1, as detailed in Section 3.5, and the log likelihood extracted from Stan perform well and correctly identify the covariate combination in most datasets.

Overall, the WAIC and PSIS-LOOIC computed with the log likelihoods extracted from Stan consistently outperforms our WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ across all simulation experiments in context of variable selection task. The findings from Figures 4.2 and 4.3 suggest that our proposed WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ are more suitable for model selection task, specifically in the context of covariance function selection, as opposed to variable selection tasks.

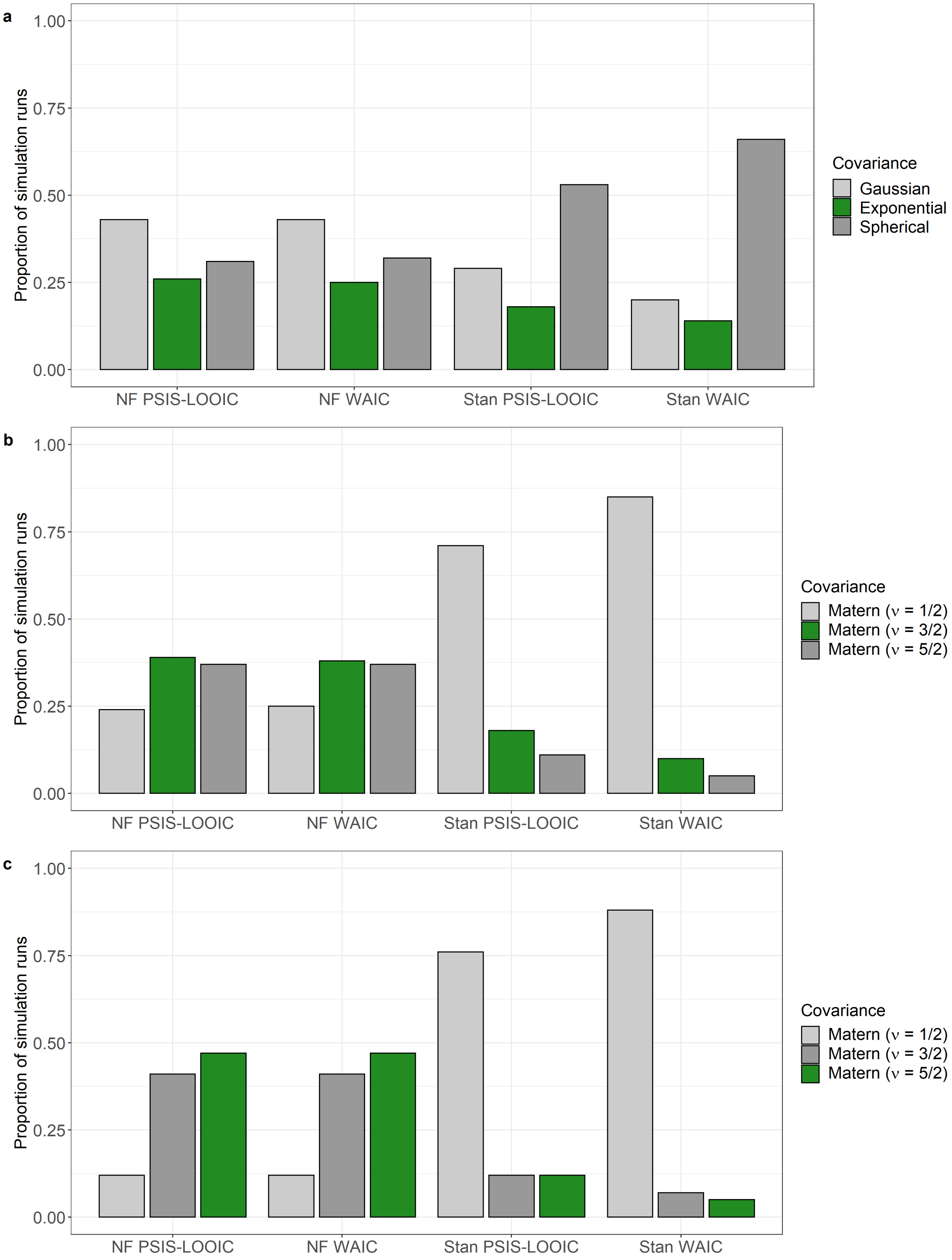

Figure 4.2. Proportion of simulation runs correctly identifying the generating configuration, with green bars indicating the correct configuration.

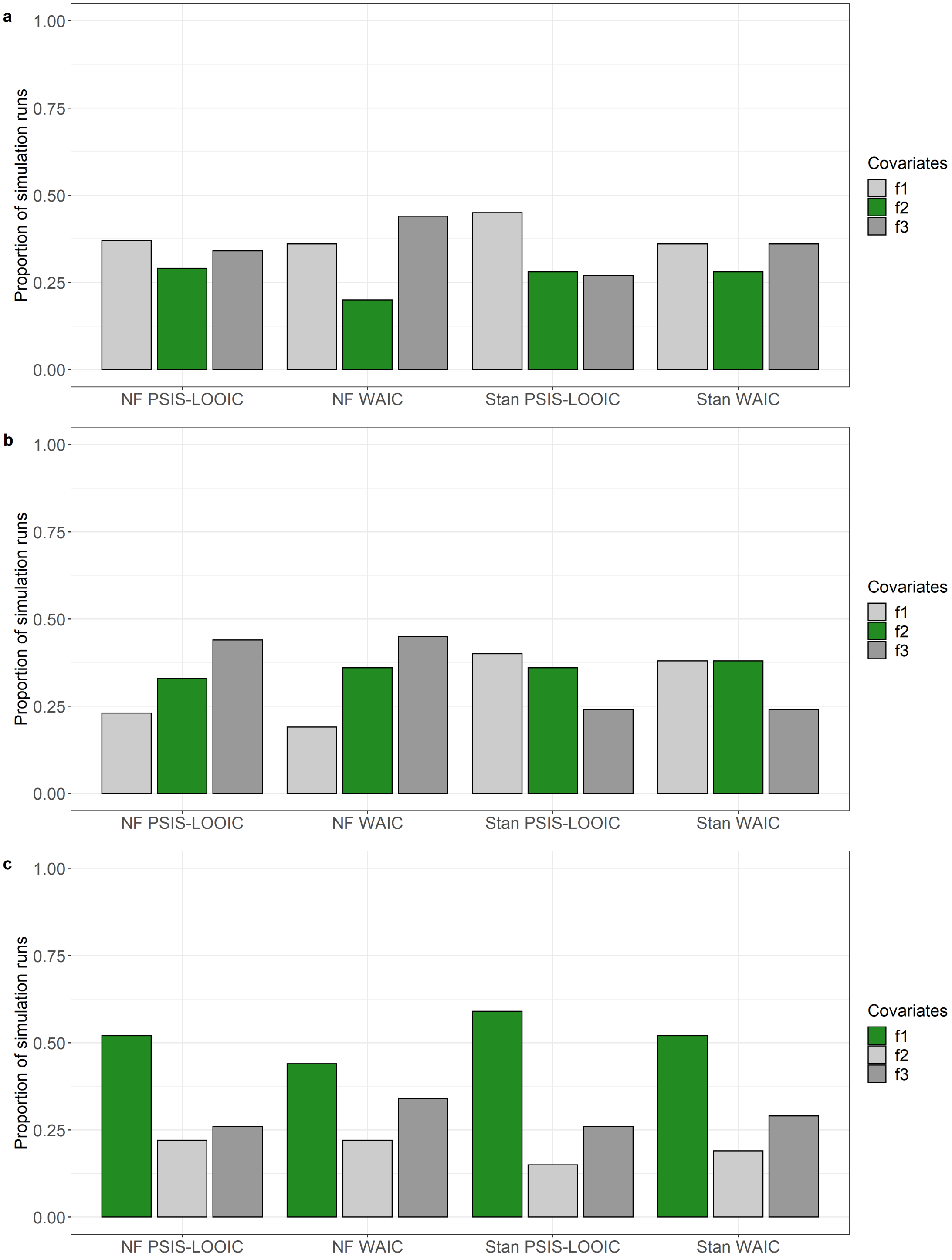

Figure 4.3. Proportion of simulation runs correctly identifying the generating configuration, with green bars indicating the correct configuration.

Figure 4.4 presents the results of experiments with configurations that are similar to those in Figure 4.2. However, in this case, the candidate models are fitted using the INLA method. In Figure 4.4a, point referenced spatial datasets were generated using $n = 50$, spatial variance $\sigma^2_{\boldsymbol{\omega}} = 2$, iid variance $\tau^2 = 2$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 1$ for a Matérn covariance function. The candidate models employed Matérn covariance functions and smoothness parameters, $\nu = 1/10$, $\nu = 1/2$ and $\nu = 1$, with the latter configuration aligning the configuration used to generate the datasets in this experiment.

Moving to Figure 4.4b, point referenced spatial datasets were generated using $n = 100$, while retaining the remaining parameters consistent with the setup in Figure 4.4a. Similarly, the candidate models utilised Matérn covariance functions and smoothness parameters, $\nu = 1/10$, $\nu = 1/2$ and $\nu = 1$, where the latter is the configuration used to generate the datasets in this experiment.

In Figure 4.4c, point referenced spatial datasets were generated using increased spatial variability, $\sigma^2_{\boldsymbol{\omega}} = 3$, and iid variance $\tau^2 = 3$. This experiment used $n = 50$ and $\phi = 3/0.5$. The smoothness parameter was set to $\nu = 1/2$ for a Matérn covariance function, implying the incorporation of an exponential covariance function.

Finally, the results in Figure 4.4d were derived from an experiment that generated datasets using $n = 100$, spatial variance $\sigma^2_{\boldsymbol{\omega}} = 3$, iid variance $\tau^2 = 3$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 1/10$ for a Matérn covariance function.

Figure 4.5 presents results similar to those in Figure 4.3, showcasing selected candidate models from variable selection tasks. Both Figures 4.5a and 4.5b are from experiments using similar data generating configurations employing spatial variance $\sigma^2_{\boldsymbol{\omega}} = 3$, iid variance $\tau^2 = 3$, spatial decay parameter $\phi = 3/0.5$, smoothness parameter $\nu = 1/2$ for a Matérn covariance function, and covariate combination f2. The distinguishing factor between these two experiments is in their sample sizes, with Figure 4.5a using $n = 50$, and Figure 4.5b using $n = 100$. In both instances, the candidate models incorporate the covariate combinations f1, f2 and f3, and were fitted using the INLA method.

Figures 4.5c and 4.5d, are also from experiments using similar data generating configurations. Both experiments used spatial variance $\sigma^2_{\boldsymbol{\omega}} = 2$, iid variance $\tau^2 = 2$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 1/2$, for a Matérn covariance function, implying the utilisation of an exponential covariance function.

Furthermore, both experiments generated datasets using the covariate combination f1. The simulation sample size is different between the two experiments, with $n = 50$ and $n = 100$ for Figures 4.5c and 4.5d respectively.

The results illustrated in Figures 4.4 and 4.5 yield several noteworthy observations. First, we observe that the WAIC_{NF} consistently outperforms the WAIC calculated by INLA in its ability to consistently and accurately identify the covariance function employed in dataset generation. Second, we observe that neither the WAIC_{NF} nor the INLA-calculated WAIC accurately demonstrates high accuracy in distinguishing the generating covariate combination.

Overall, the WAIC calculated by INLA outperform our proposed WAIC_{NF} across all experiments in the context of variable selection tasks. These findings provide valuable insights into the comparative strengths and limitations of these two approaches. In summary, the result from Figures 4.4 and 4.5 highlight the effectiveness of our proposed WAIC_{NF} in model selection tasks, especially concerning the identification of the underlying covariance function, while indicating its limitation in variable selection tasks.

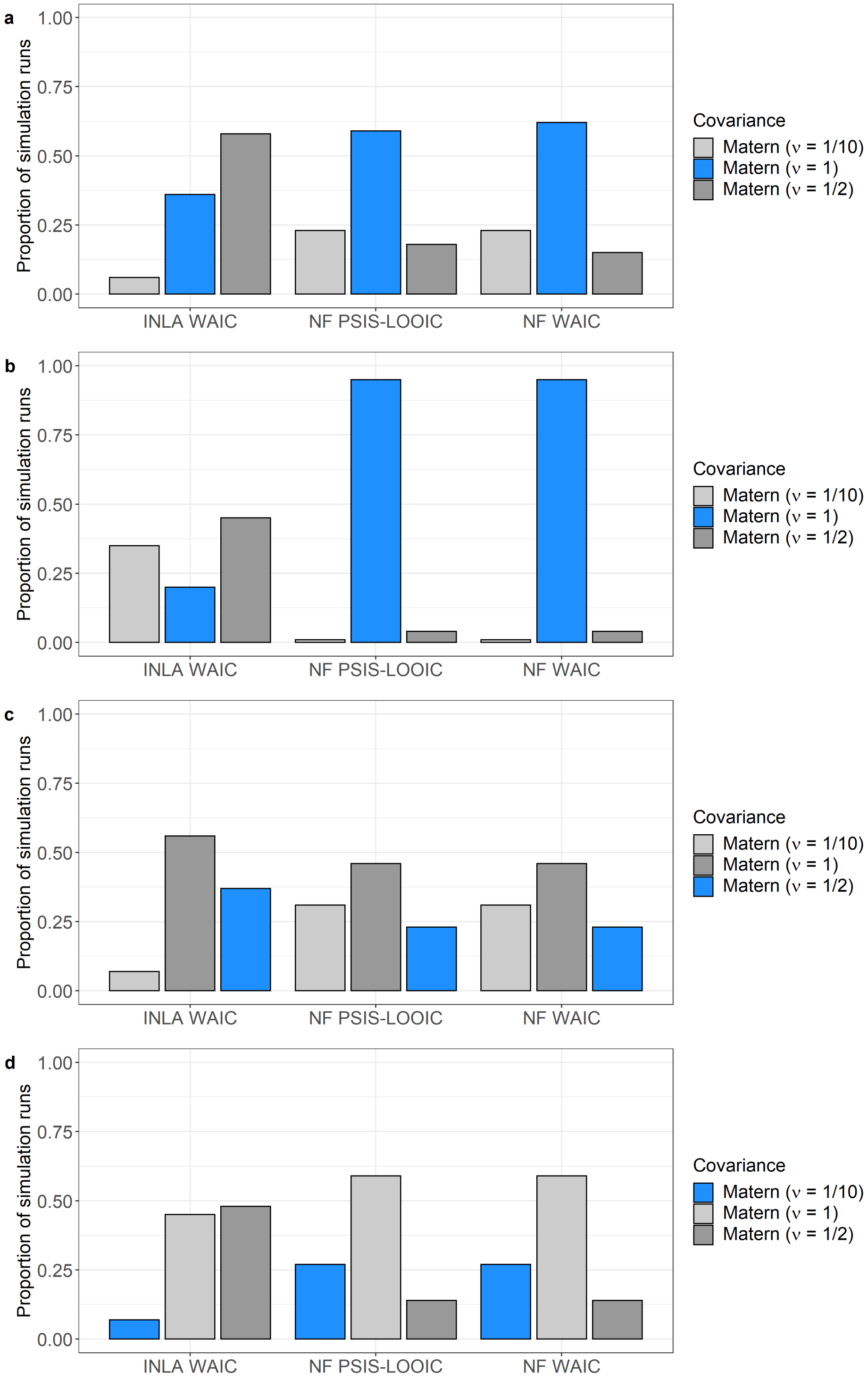

Figure 4.4. Proportion of simulation runs correctly identifying the generating configuration, with blue bars indicating the correct configuration.

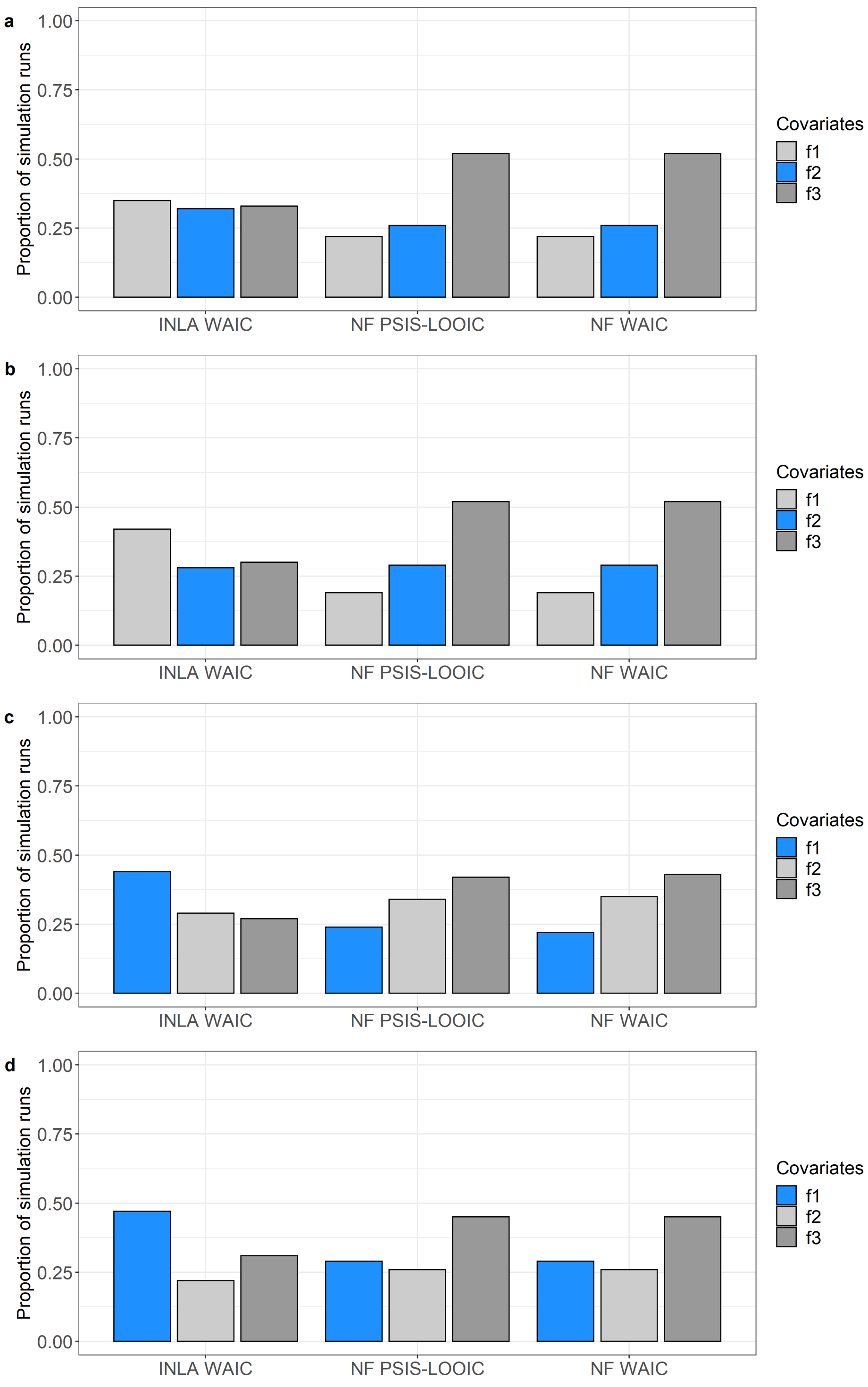

Figure 4.5. Proportion of simulation runs correctly identifying the generating configuration, with blue bars indicating the correct configuration.

While Figures 4.2, 4.3, 4.4 and 4.5 show the proportion of simulation runs in which the selected candidate model correctly identifies the configuration of the generating dataset, it is also valuable to examine the individual values of the selection criteria alongside their corresponding penalty components.

Table 4.1 provides the calculated selection criteria and their corresponding penalty components. The results presented in Table 4.1 do not involve candidate models. Instead, the selection criteria are computed from Bayesian models fitted using the Stan framework, and they are constructed using parameters according to the configuration of the generating dataset.

M1 represents an experiment where the point referenced spatial dataset was generated using $n = 15$, spatial variance $\sigma^2 = 2$, iid variance $\tau^2 = 2$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 3/2$ for a Matérn covariance function. Additionally, the dataset was generated using covariate combination f1.

In M2, the point reference spatial dataset was generated using $n = 20$, spatial variance $\sigma^2 = 3$, iid variance $\tau^2 = 3$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 1/2$ for a Matérn covariance function, implying an exponential covariance function. Furthermore the dataset was generated using covariate combination f1.

Finally, in M3, the point referenced spatial dataset was generated using $n = 10$, spatial variance $\sigma^2 = 3$, iid variance $\tau^2 = 3$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 5/2$ for a Matérn covariance function. The dataset was generated using covariate combination f2.

Table 4.1. Selection criteria and corresponding penalty components derived from Bayesian models fitted within the Stan framework.

	M1	M2	M3
WAIC_{NF} (p_{WAIC})	48.56 (0.19)	66.21 (0.31)	33.49 (0.22)
Stan WAIC (p_{WAIC})	66.99 (4.64)	91.98 (7.11)	50.20 (3.79)
$\text{PSIS-LOOIC}_{\text{NF}}$ ($p_{\text{PSIS-LOOIC}}$)	48.56 (0.19)	66.23 (0.32)	33.52 (0.23)
Stan PSIS-LOOIC ($p_{\text{PSIS-LOOIC}}$)	67.76 (5.02)	93.53 (7.88)	51.06 (4.22)

The selection criteria and their corresponding penalty components presented in Table 4.2 are computed from Bayesian models fitted using the INLA method, and they are constructed using parameters corresponding to the following data generating configurations.

In M4, the point referenced spatial dataset was generated using $n = 50$, spatial variance $\sigma^2_{\boldsymbol{\omega}} = 2$, iid variance $\tau^2 = 2$, spatial decay parameter $\phi = 3/0.5$, smoothness parameter $\nu = 1$ for a Matérn covariance function, along with covariate combination f1.

In M5, the point referenced spatial dataset was generated using an increased sample size $n = 100$, spatial variance $\sigma^2_{\boldsymbol{\omega}} = 3$, iid variance $\tau^2 = 3$, spatial decay parameter $\phi = 3/0.5$ and smoothness parameter $\nu = 1/2$ for a Matérn covariance function. Additionally, the dataset was generated using covariate combination f1.

In M6, the point referenced spatial dataset was generated using $n = 50$, spatial variance $\sigma^2_{\boldsymbol{\omega}} = 3$, iid variance $\tau^2 = 3$, spatial decay parameter $\phi = 3/0.5$, smoothness parameter $\nu = 1$ for a Matérn covariance function and covariate combination f2.

In M7, the point referenced spatial dataset was generated using $n = 100$, spatial variance $\sigma^2_{\boldsymbol{\omega}} = 2$, iid variance $\tau^2 = 2$, spatial decay parameter $\phi = 3/0.5$, smoothness parameter $\nu = 1/10$ for a Matérn covariance function and covariate combination f2.

Table 4.2. Selection criteria and corresponding penalty components derived from Bayesian models fitted using the INLA method.

	M4	M5	M6	M7
INLA WAIC (p_{WAIC})	174.88 (17.38)	413.37 (38.66)	209.19 (3.82)	407.97 (10.66)
WAIC$_{\text{NF}}$ (p_{WAIC})	144.76 (0.90)	336.57 (0.47)	159.66 (0.31)	320.17 (0.39)
PSIS-LOOIC$_{\text{NF}}$ ($p_{\text{PSIS-LOOIC}}$)	144.88 (0.95)	336.57 (0.47)	159.66 (0.32)	320.17 (0.39)

From Tables 4.1 and 4.2, we observe that the WAIC$_{\text{NF}}$ and PSIS-LOOIC$_{\text{NF}}$ values are consistently less than the WAIC values computed by INLA, and the WAIC and PSIS-LOOIC values calculated using the log likelihoods extracted from Stan, across all experiments. Similarly, the values of the corresponding penalty component for WAIC$_{\text{NF}}$ and PSIS-LOOIC$_{\text{NF}}$ are consistently less than the values of the penalty components of the WAIC computed by INLA, and the WAIC and PSIS-LOOIC calculated using the log likelihoods extracted from Stan. These relationships are further illustrated in Figures 4.6 and 4.7.

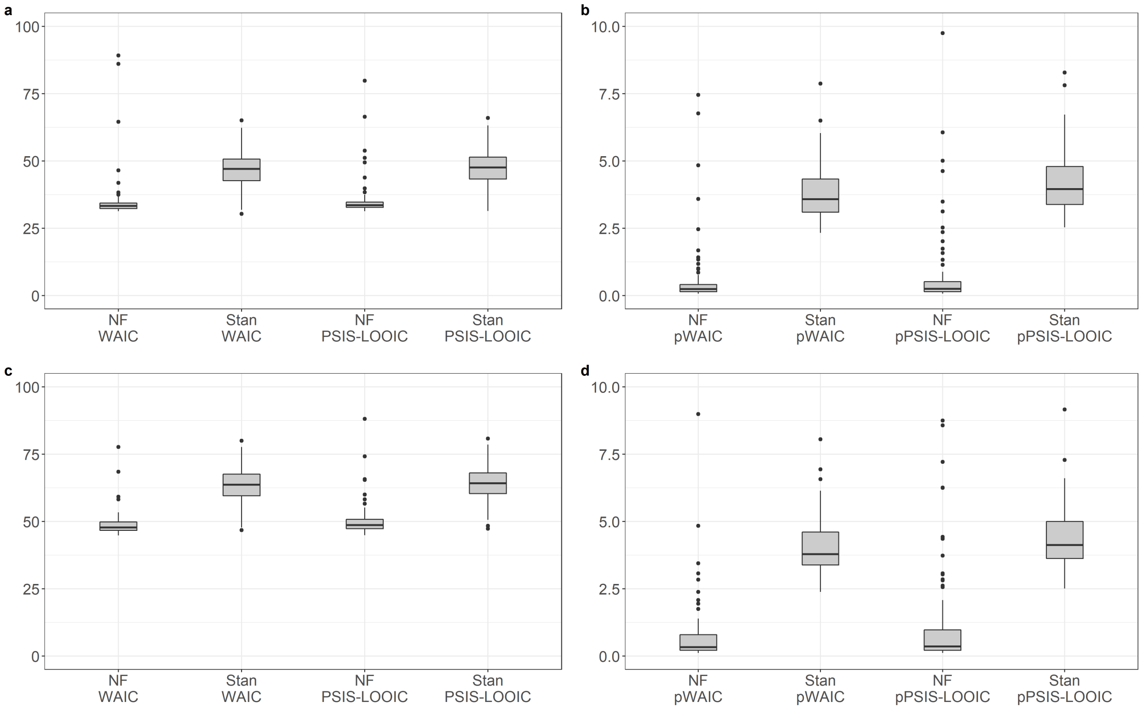

Figure 4.6. Selection criteria and associated penalty components calculated from Bayesian models fitted using the Stan framework across 100 simulation runs.

In Figure 4.6 we further explore the values of the selection criteria and their associated penalty components. We constructed Bayesian models based on the configurations used to generate the datasets for each experiment. The configurations for the experiments are as follows. In Figures 4.6a and 4.6b, the point referenced spatial dataset was generated using $n = 10$, $\sigma_{\boldsymbol{\omega}}^2 = 3$, $\tau^2 = 3$, $\phi = 3/0.5$, and $\nu = 5/2$ for a Matérn covariance function. In Figures 4.6c and 4.6d, the dataset was generated using $n = 15$, $\sigma_{\boldsymbol{\omega}}^2 = 2$, $\tau^2 = 2$, $\phi = 3/0.5$ and $\nu = 3/2$ for a Matérn covariance function.

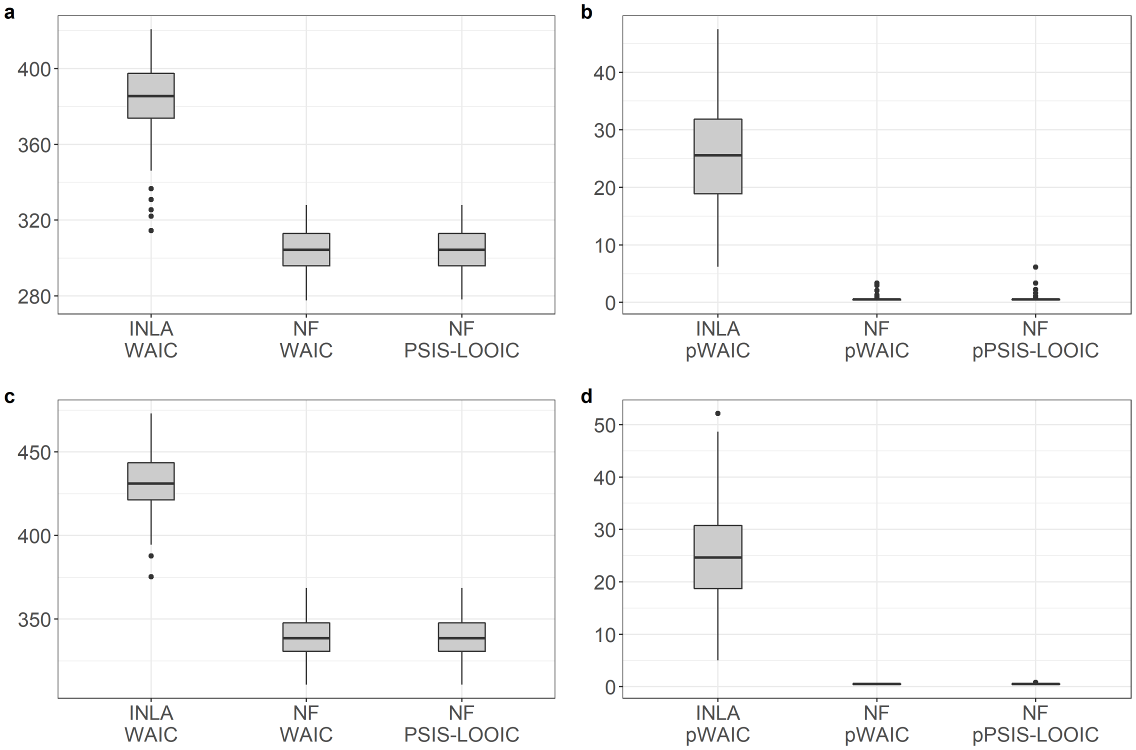

Figure 4.7. Selection criteria and associated penalty components calculated from Bayesian models fitted using the INLA method across 100 simulation runs.

The experiments behind the results presented in Figure 4.7 is similar to those in Figure 4.6, except the models are fitted using the INLA method. The data generating configurations for the experiments in Figure 4.7 is as follows. In Figures 4.7a and 4.7b, the point referenced spatial dataset was generated using $n = 100$, $\sigma^2_{\boldsymbol{\omega}} = 2$, $\tau^2 = 2$, $\phi = 3/0.5$ and $\nu = 1$ for a Matérn covariance function. In Figures 4.7c and 4.7d, the dataset was generated using $n = 100$, $\sigma^2_{\boldsymbol{\omega}} = 3$, $\tau^2 = 3$, $\phi = 3/0.5$ and $\nu = 1/2$ for a Matérn covariance function.

Figures 4.6 and 4.7 once again demonstrate that the values of WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, along with their corresponding penalty components, are less than the values of the WAIC and PSIS-LOOIC, as well as their associated penalty components, computed by both INLA and derived from the log likelihoods extracted from Stan.

Figures 4.8 and 4.9 show that the relationship holds, where the values of the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ is less than those of the WAIC and PSIS-LOOIC computed from INLA and derived from the log likelihoods of Stan, even when the spatial decay parameter varies. In both figures, the datasets were generated using $n = 15$, spatial variance $\sigma^2_{\omega} = 3$, iid variance $\tau^2 = 3$ and smoothness parameter $\nu = 1/2$ for a Matérn covariance function, implying an exponential covariance function. We considered three different spatial decay parameters for both data generating and model fitting. They are $\phi = 3/0.25$, $\phi = 3/0.50$ and $\phi = 3/0.75$.

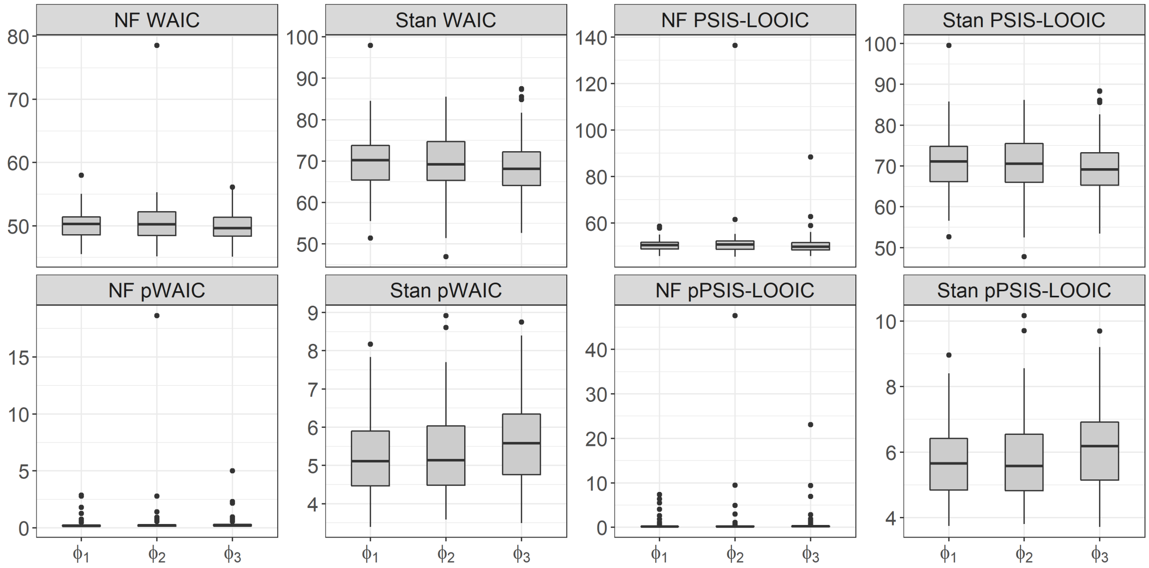

Figure 4.8. Selection criteria values and their corresponding penalty components, computed from models fitted within the Stan framework across 100 simulation runs, using $\phi_1 = 3/0.25$, $\phi_2 = 3/0.50$ and $\phi_3 = 3/0.75$.

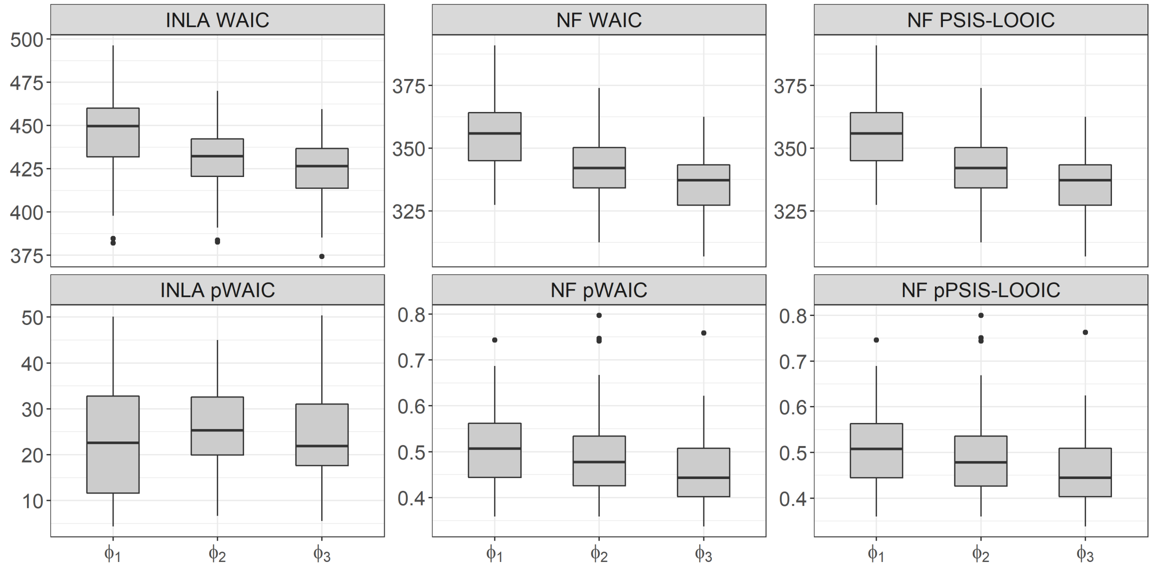

Figure 4.9. Selection criteria values and their corresponding penalty components, computed from models fitted using the INLA method across 100 simulation runs, with $\phi_1 = 3/0.25$, $\phi_2 = 3/0.50$ and $\phi_3 = 3/0.75$.

4.4 Summary of the simulation example results

In summary, the results from the simulation examples conducted in this chapter provided valuable insights into both the capabilities and limitations of our proposed approach for calculating the WAIC and PSIS-LOOIC using non-factorisable model log likelihoods. These findings suggest that the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ serve as adept selection criteria for model selection tasks. Specifically, they are suitable for identifying the optimal spatial model among candidate models, each defined by distinct covariance functions within the spatial random effect.

However, the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ exhibit limitations when applied to variable selection tasks, where the objective is to choose a subset of covariates from a larger pool of potential candidates. Our simulation results consistently indicate a tendency for both the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ to prefer models that include all available covariates, even when that configuration does not match the one we used to generate the datasets.

In conclusion, these findings offer a comprehensive perspective on the utility of the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, highlighting their effectiveness in context of model selection, while also acknowledging the need for alternative approaches for scenarios centered on variable selection.

Chapter 5

The WAIC$_{\text{NF}}$ in practice: Mapping MCV1 coverage in Nigeria

In this chapter, we illustrate the practical application of our novel criterion, the WAIC$_{\text{NF}}$, for models of point referenced spatial data, within the context of a model selection task. Specifically, our focus centres on the selection of the optimal covariance function for models constructed for a specific dataset concerning MCV1 coverage in Nigeria. The outline of this chapter is the following. We begin by introducing the MCV1 dataset. Subsequently, we provide descriptions of the constructed model. Following this, we present the results of two calculations: the Watanabe-Akaike information criterion (WAIC) as computed by INLA, and our WAIC$_{\text{NF}}$, which was introduced in Chapter 3. Finally, we discuss our findings and the associated interpretations and implications arising from the results.

5.1 Measles vaccine coverage in Nigeria

Measles stands as one of the primary contributors to mortality among children under-five in Nigeria that is preventable through vaccination (Ibrahim et al., 2019, Shorunke et al., 2019). Symptoms of this acute viral infection include fever, coughing, coryza (commonly referred to as a runny nose) and conjunctivitis (characterised by red eyes) (WHO, 2019, Wariri et al., 2021). Although industrialised countries have achieved effective control over measles, it remains endemic in Nigeria. Along with 45 other countries, Nigeria account for 94% of global measles-related deaths (WHO, 2006, Ibrahim et al., 2019). Incidences of measles escalates during the dry seasons (Shorunke et al., 2019, Ori et al., 2021) and whenever vaccination coverage is low (Muscat et al., 2009, Curtale et al., 2010, Nmor et al., 2011, Kabra and Lodha, 2013). In response to low measles immunisation, health organisations have adopted and implemented strategies to eliminate measles by 2020. These strategies involve increasing first-dose of measles-containing vaccine (MCV1) coverage at both the national and district level (WHO, 2011, Orenstein et al., 2018, WHO, 2019, Faruk et al., 2020, Ori et al., 2021). Continuous monitoring of MCV1 coverage is important.

Models for point referenced spatial data are valuable for understanding the heterogeneities in MCV1 coverage across the country by producing spatially detailed estimates and maps of vaccination coverage. The Demographic Health Survey (DHS) program collects and disseminates accurate, nationally representative data on fertility, family planning, maternal and child health, gender, HIV/AIDS, malaria, and nutrition for lower- to middle-income countries (Croft et al., 2018). The DHS program conducts a survey collection once every five years for Nigeria (National Population Commission Nigeria and ICF, 2019). Recognising the constraints of finite resources and workforce availability, complete collection of MCV1 coverage data in every single community across Nigeria proves to be impractical. Instead, constructing models facilitates efficient high-resolution mapping of MCV1 coverage to understand variation in coverage at the community level. Predictions with the constructed models can cover locations that were not surveyed and can provide information that is crucial for policy making. High-resolution prediction surfaces can be used to identify "cold spots" characterised by low coverage (Utazi et al., 2020). Policy makers can use this information to focus more resources and strategic efforts towards areas necessitating targeted interventions in prospective planning endeavours.

In this chapter, we use the MCV1 survey data within the 2018 Nigeria DHS dataset, which we will refer to as the "MCV1 dataset". The objectives of this chapter are to construct models for the MCV1 dataset, and to identify the optimal covariance function among the models using the $\mathrm{WAIC}_{\mathrm{NF}}$ and the WAIC computed from INLA.

5.2 Data

The following section provides an overview of the data employed in this chapter. This includes the MCV1 dataset, the geographical boundaries of Nigeria sourced from the Database of Global Administrative Areas and the geospatial covariates assembled for the analysis.

5.2.1 The MCV1 dataset

The MCV1 dataset contains vaccination records of MCV1 from the 2018 DHS dataset. The vaccination records of MCV1 have the responses "vaccination date on card", "vaccination marked on card", "reported by mother", "do not know" and "no". We applied a binary coding scheme to these responses, with "no" and "do not know" coded as zeroes, and the other responses coded as ones. The recoded survey responses were aggregated by the survey cluster ID such that for each survey cluster location, we have a total number of survey responses, and a count of how many children aged 12 to 23 months were vaccinated with MCV1. To prevent excessive proportions of zeroes and ones, survey cluster locations with fewer than two were excluded. Thus, the analysis is based on $n = 1319$ survey cluster locations. For the remainder of this chapter, we use "count" to denote MCV1-vaccinated children in a DHS cluster, "total" to represent the total survey responses in a DHS cluster and "proportion" (or "prop") to denote the proportion, calculated as the count divided by the total.

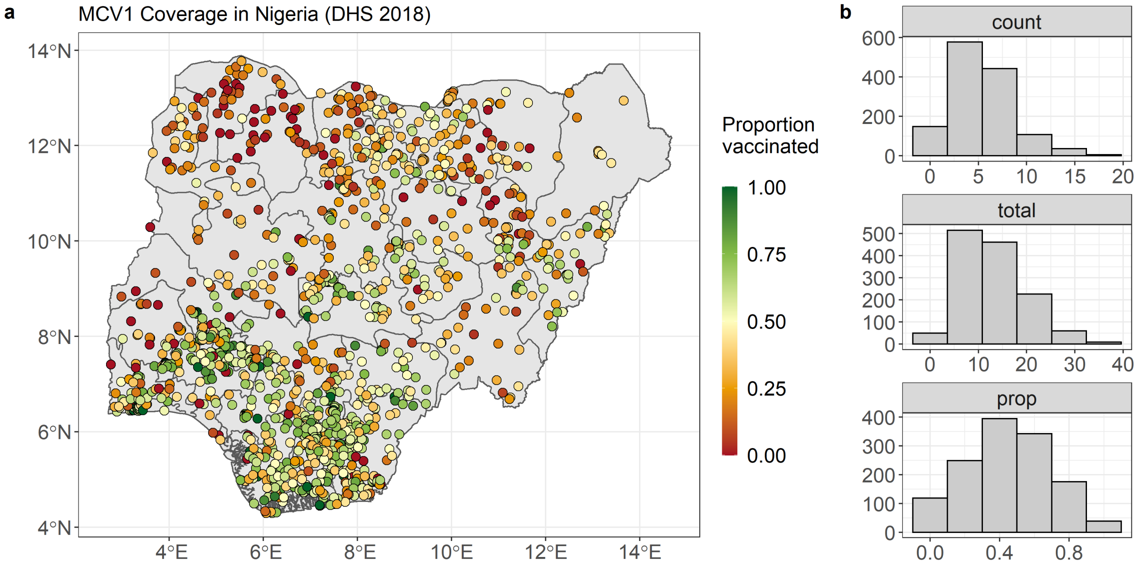

Figure 5.1. The 2018 DHS survey cluster locations across Nigeria and histograms of MCV1 survey data.

Figure 5.1a shows the cluster locations across Nigeria and the corresponding proportion of children vaccinated with MCV1. Notably, areas of lower MCV1 vaccination proportions are concentrated in the north-western region, while clusters in the southern part of the country exhibit higher proportions. Across the $n = 1319$ DHS survey clusters, the maximum count is 18 and the median count is 5. Similarly, the maximum total is 38 and the median total is 12. The mean proportion is approximately 0.46. The histograms in Figure 5.1b illustrate positive skewness of both count and total, whereas proportion is observed to be normally distributed.

5.2.2 The geographical boundaries of Nigeria

Shape files containing geographical boundaries were employed primarily for visualisation purposes. These boundaries were sourced from the Database of Global Administrative Areas (GADM) version 4.1 (Global Administrative Areas, 2018). The first-level sub-division provided by GADM encompasses Nigeria's 36 states along with the Federal Capital Territory, as depicted by the borders in Figure 5.1a. The second-level sub-division comprises the 774 local government areas (LGAs).

5.2.3 The geospatial covariates

We consider predictors that are known to influence MCV1 coverage in our model. Detail related to model construction will be further discussed in Section 5.3. In this study, we use the following to represent the environmental, geographical and socioeconomic factors that influence MCV1 coverage: "poverty", "temperature", "nightlights" and "traveltime" (Utazi et al., 2018, 2020). These geospatial covariates are given as 1×1km rasters, as visualised in Figure 5.2.

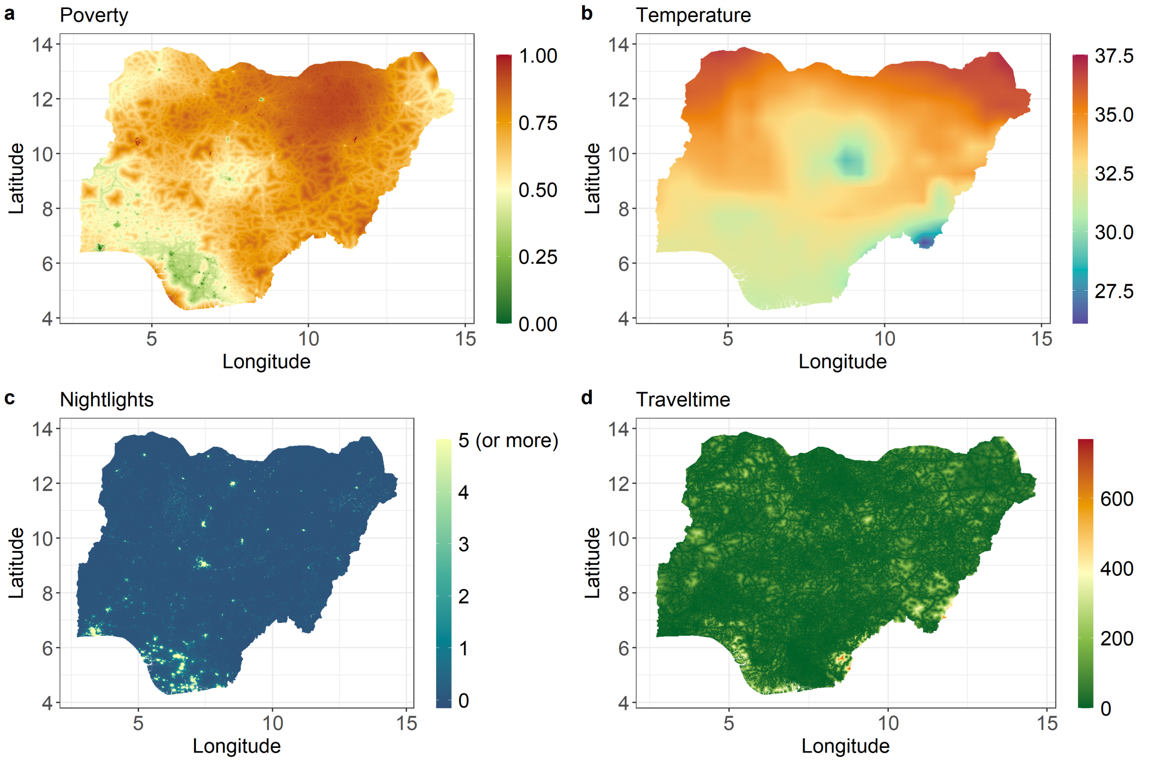

Figure 5.2. The geospatial covariates given as 1 × 1km rasters.

The geospatial covariates listed above and shown in Figure 5.2 are described as follows. "Poverty" denotes the 2010 estimates of proportion of individuals residing in poverty, per grid square, according to a \$1.25-a-day threshold (Tatem et al., 2013). "Temperature" denotes the average maximum temperature recorded between 2013 and 2018, and is measured in degrees Celsius (Wan et al., 2015). "Nightlights" denote the nocturnal luminosity in 2016 acquired from the Visible Infrared Imaging Radiometer Suite and quantified in nano-watts (NOAA, 2019). "Traveltime" denotes the travel duration to cities in 2015 in minutes (Weiss et al., 2018).

Recall, the MCV1 dataset is aggregated to the DHS cluster locations (Figure 5.1a) while the geospatial covariates are given at the grid level (Figure 5.2) — a higher spatial resolution. To harmonise these data, we employ an extraction and aggregation process on the geospatial covariates to match the MCV1 dataset. The following steps outline this process. First, we refer to the DHS urban-rural classification associated with each cluster ID of the MCV1 dataset. This information is important for the extraction process because the DHS program intentionally displaces the actual survey locations to ensure anonymity. This displacement is up to 2km for urban survey locations and 5km for rural survey locations (Burgert et al., 2013). Using the urban-rural classifications, we create "buffer zones" of 2km and 5km radii around the latitude and longitude coordinates of the DHS survey locations given in the MCV1 dataset on the raster. We extract the grids of the geospatial covariate within these buffer zones. Finally, we compute the mean values from these extracted grids to represent the geospatial covariate at the DHS survey cluster level which matches the MCV1 dataset. When this extraction and aggregation process is completed for all geospatial covariates, we have the full MCV1 dataset.

5.3 Constructing models for the MCV1 dataset

The models we construct for the MCV1 dataset have the following structure. For DHS survey cluster location $\mathbf{s}_i$, where $i = 1, \dots, n$ DHS survey clusters, let $Y(\mathbf{s}_i)$ denote children vaccinated with MCV1 in $\mathbf{s}_i$ and $N(\mathbf{s}_i)$ denote the total number surveyed in $\mathbf{s}_i$. Now, $Y(\mathbf{s}_i)$ follows a binomial distribution with $N(\mathbf{s}_i)$ and $p(\mathbf{s}_i)$, which denotes the probability of children vaccinated with MCV1,

$$Y(\mathbf{s}_i) \sim \text{Binomial}\big(N(\mathbf{s}_i), p(\mathbf{s}_i)\big). \tag{5.1}$$

The model further assumes that $p(\mathbf{s}_i)$ is linked to the geospatial covariates, which was described in Section 5.2.3, and the random effects through a logit link,

$$\text{logit}\big(p(\mathbf{s}_i)\big) = \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \omega(\mathbf{s}_i), \tag{5.2}$$

where $\boldsymbol{\beta}$ denotes a p-dimension column vector of regression coefficients, $\mathbf{x}(\mathbf{s}_i)'$ denotes a vector of geospatial covariates associated with $\mathbf{s}_i$ and $\omega(\mathbf{s}_i)$ denotes a vector of spatial random effects.

Let us first focus on $\mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta}$. We can express this component in a more compact manner with the matrix notation $X\boldsymbol{\beta}$, where X is an $n \times p$ design matrix, and the definition of $\boldsymbol{\beta}$ remains unchanged. At this stage of the model construction, we consider the appropriate covariate transformations and the combination of covariates to incorporate into our model. Transformations of the geospatial covariates are implemented to improve the linearity between these covariates and MCV1 coverage, specifically on the logit scale. To determine if covariate transformations are needed, we observe the histogram of the covariates. Figure 5.3 shows that nightlights and traveltime are heavily right skewed. Applying a log transformation on them transforms their histogram to become somewhat more symmetrical, relative to the original scale. The distribution of poverty is normally distributed to begin with, so transformation is not needed. Applying a log transformation on temperature does not change the overall shape of the distribution, so we will leave it at the original scale. In summary, we apply the log transformation only on nightlights and traveltime.

Next, we need to check that multicollinearity is not an issue amongst the geospatial covariates. In practice (Utazi et al. (2022), Pezzulo et al. (2023), Utazi et al. (2023) for example), multicollinearity is checked with both the Pearson's correlation matrix and variance inflation factors (VIF) (Fox and Monette, 1992, Kutner and Nachtsheim, 2004). Figure 5.4 shows the pairwise scatter plot of the empirical logit of MCV1 coverage and the geospatial covariates. We define the empirical logit as

$$\text{elogit}\big(p(\mathbf{s}_i)\big) = \log\left(\frac{y(\mathbf{s}_i) + 0.5}{m(\mathbf{s}_i) - y(\mathbf{s}_i) + 0.5}\right),$$

where $y(\mathbf{s}_i)$ denote the observed number of children vaccinated with MCV1 at survey cluster location $\mathbf{s}_i$, and $m(\mathbf{s}_i)$ denote the observed total number of surveys conducted at $\mathbf{s}_i$. We also included 0.5 within the calculation of the empirical logit to avoid computational issues related to the $\log(\cdot)$ operator. In Figure 5.4, we observe that there are no correlation coefficient values $|r| > 0.8$, where $|r|$ denotes the absolute value of the correlation coefficient value. To further assess potential multicollinearity, we employed a non-spatial model that incorporates all geospatial covariates. Subsequently, we computed the VIF, and found that all VIF values are less than four. Hence, we can confidently assert that multicollinearity is not an issue. In the situation where multicollinearity is present, a decision needs to be made by selecting one of the covariates between the problematic pair and omitting the other. We note that in this study, we only considered the linear relationship between the geospatial covariates and without any interactions for $\mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta}$.

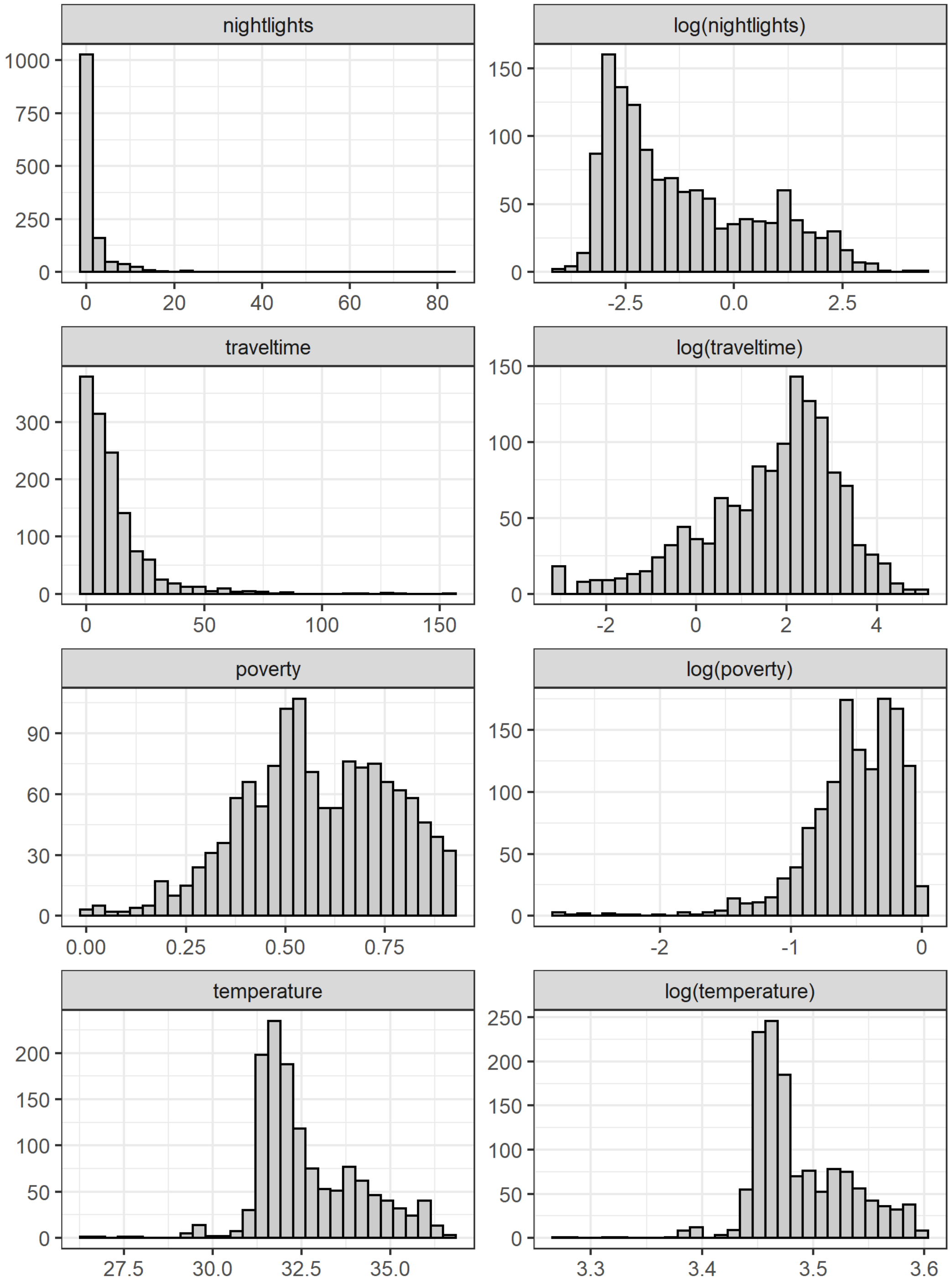

Figure 5.3. Histograms of the covariates. Panels on the left show the histogram of the covariates at the original scale. Panels on the right show the histograms of the covariates at the log scale.

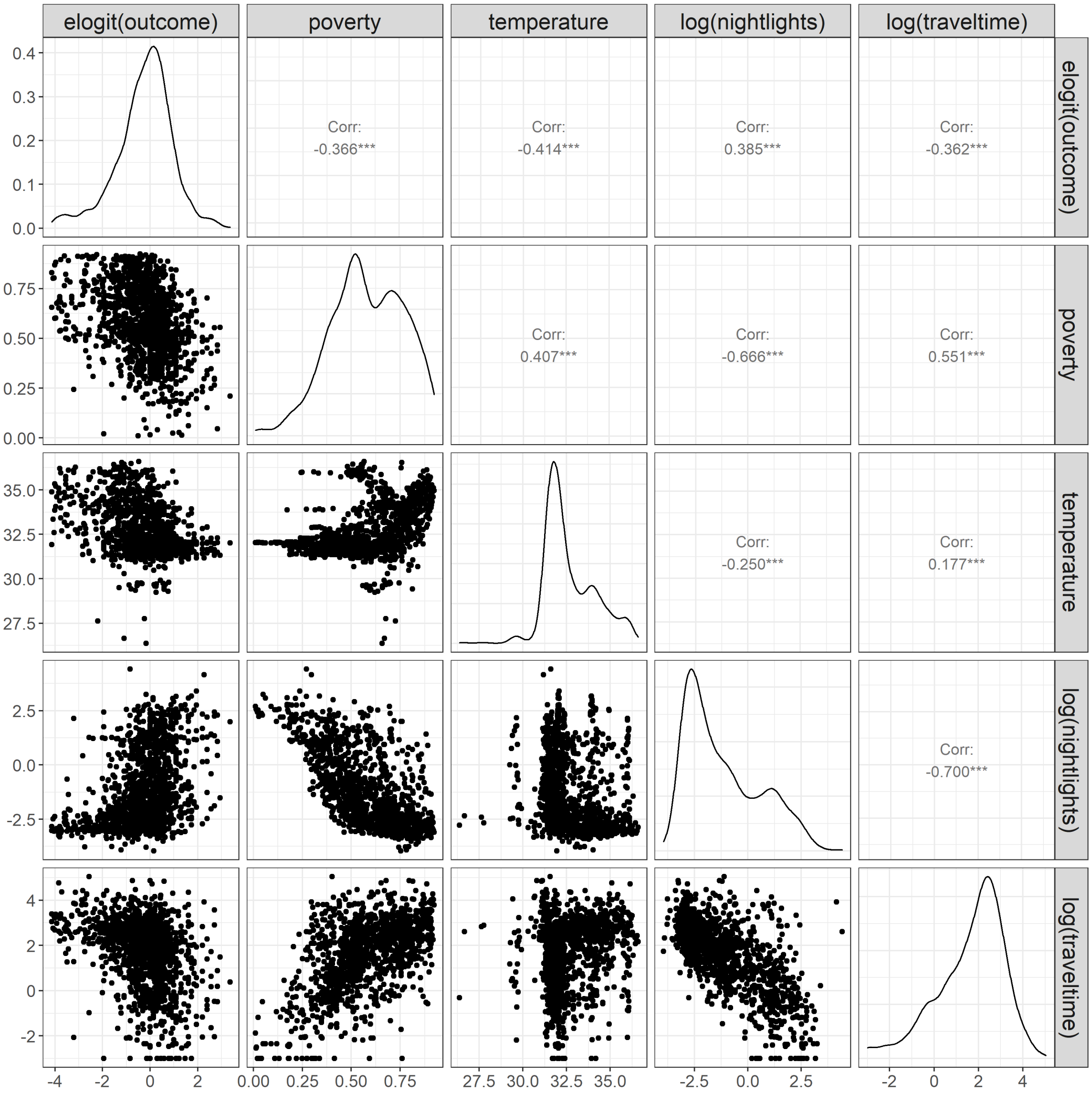

Figure 5.4. Pairwise scatter plot and Pearson's correlation coefficient matrix amongst the covariates and the empirical logit of the proportion of children vaccinated with MCV1.

Returning to the logit link (5.2), the spatial random effects $\boldsymbol{\omega} = \big(\omega(\mathbf{s}_1), \ldots, \omega(\mathbf{s}_n)\big)'$ are used to capture residual spatial correlation in the data and follow an n-dimensional multivariate normal distribution with mean zero and an $n \times n$ covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$,

$$\boldsymbol{\omega} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}).$$

We assume that the elements within the covariance matrix $\mathbf{\Sigma}_{\boldsymbol{\omega}}$ are calculated from the Matérn covariance function, which has the following definition within the INLA framework,

$$\text{Cov}(\mathbf{s}_i, \mathbf{s}_j) = \frac{\sigma^2_{\boldsymbol{\omega}}}{2^{\nu-1}\Gamma(\nu)}(\kappa||\mathbf{s}_i - \mathbf{s}_j||)^{\nu} K_{\nu}(\kappa||\mathbf{s}_i - \mathbf{s}_j||), \tag{5.3}$$

for $i,\ j = 1, \dots, n$, where $||\mathbf{s}_i - \mathbf{s}_j||$ denotes the Euclidean distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$. The $\Gamma(\cdot)$ is the mathematical Gamma function, K_ν is the modified Bessel function of the second kind and order ν, ν denotes the smoothness parameter, κ denotes a scaling parameter and $\sigma^2_{\boldsymbol{\omega}}$ denotes the spatial variance. The ν parameter is usually a fixed value, since it is poorly identified (Lindgren and Rue, 2015). It is given as $\nu = \alpha - d/2$, where α is a fractional operator and $d = 2$ when we are working in the spatial domain. The scaling parameter κ has an empirically derived definition of the range parameter, given as $r_{sp} = \sqrt{(8\nu)}/\kappa$, which corresponds to spatial correlation close to 0.1 (Lindgren et al., 2011). The range parameter r_{sp} is used in favour over κ for its interpretability.

The exponential covariance function is a popular choice for applied spatial models (Moraga, 2019) and is a special case of (5.3) of when $\nu = 1/2$. From the definitions, $\nu = 1/2$ when $\alpha = 3/2$ and $d = 2$, and is given as

$$\text{Cov}(\mathbf{s}_i, \mathbf{s}_j) = \sigma^2_{\boldsymbol{\omega}} \exp(-\kappa||\mathbf{s}_i - \mathbf{s}_j||). \tag{5.4}$$

To complete the setup for Bayesian modelling, we specified the following prior distributions to the unknown parameters of interest. We assign the regression coefficients $\boldsymbol{\beta} \sim N_p(\mathbf{0}, 2I_p)$, where $N_p(\cdot)$ denotes a p-dimension multivariate normal distribution, and I_p denotes a $p \times p$ identity matrix. We assign penalised complexity (PC) priors (Simpson et al., 2017) on the spatial variance $\sigma^2_{\boldsymbol{\omega}}$ and the spatial range r_{sp}, such that $P(\sigma^2_{\boldsymbol{\omega}} > 1) = 0.05$ and $P(r_{sp} < r_0) = 0.01$, respectively. Within the PC priors, r_0 denotes the 5% of the extent of Nigeria in the east-west direction.

5.4 Covariance function selection

To demonstrate the practical application of our WAIC$_{\text{NF}}$, we construct three models. The models follow (5.2) as described in Section 5.3. For the Matérn covariance function (5.3) in our models, we use three different ν parameters: $\nu = 1/2$, $\nu = 1$

and $\nu = 1/10$. These ν parameters translate to $\alpha = 3/2$, $\alpha = 2$ and $\alpha = 1.1$, respectively, since $d = 2$ and $\nu = \alpha - d/2$. These parameters are supplied as arguments in the `inla.spde2.pcmatern()` function within the `INLA` package (Rue et al., 2009, Martins et al., 2013) in R. It is noted that when using the `inla.spde2.pcmatern()`, only a limited range of α values are supported. Specifically, $\alpha \in (0, 2]$ are supported for $\nu = \alpha - d/2 > 0$. Since $d = 2$, then $\nu = \alpha > 1$, meaning that we can only use $\alpha \in (1, 2]$. The limitation on α was discussed in detail in Section 2.6.1.

After processing the three models in INLA, we calculate the WAIC_{NF}. Additionally, we also calculate the $\text{PSIS-LOOIC}_{\text{NF}}$, which was introduced in Section 3.5. We compare the WAIC_{NF} against the WAIC computed from INLA. As discussed in Chapter 2.7.2, we determine the model that returns the smallest WAIC value to be the best candidate model. Furthermore, we conduct K-fold cross-validations to validate model performance. Particularly, we calculate the out-of-sample root mean-squared error (RMSE), the mean absolute error (MAE) and the continuous ranked probability score (CRPS). The RMSE and MAE can be used to assess parameter estimation, and the CRPS can be used to assess predictive ability. They are given as

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big(\hat{p}(\mathbf{s}_i) - p(\mathbf{s}_i)\big)^2},$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\big|\hat{p}(\mathbf{s}_i) - p(\mathbf{s}_i)\big|,$$

where $\hat{p}(\mathbf{s}_i)$ is the predicted proportion and $p(\mathbf{s}_i)$ is the observed proportion at DHS cluster locations $\mathbf{s}_i$. The CRPS is defined by Gneiting and Raftery (2007) as

$$\text{CRPS}(F, y) = E_F|Y - y| - \frac{1}{2}E_F|Y - Y^*|,$$

where Y and Y^* are independent copies of a random variable with the cumulative distribution function $F(\cdot)$. The CRPS can be estimated with MCMC samples following the equation given by Sahu (2022)

$$\widehat{\text{CRPS}} = \frac{1}{S}\sum_{j=1}^{S}|y^{(j)} - y| - \frac{1}{2S^2}\sum_{j=1}^{S}\sum_{k=1}^{S}|y^{(j)} - y^{(k)}|,$$

where S is the total number of MCMC samples. We fitted the model with the `INLA` package (Rue et al., 2009, Martins et al., 2013) and conducted all analyses in R version 4.0.4 (R Core Team, 2021).

5.5 Results

The WAIC_{NF}, $\text{PSIS-LOOIC}_{\text{NF}}$ and the WAIC computed from INLA are presented in Table 5.1. The WAIC calculated by INLA show marginal differences between the the three candidate models with different ν parameters. However, determining the best model strictly from the calculated size of the WAIC value, the best model is the one fitted using a Matérn covariance function and smoothness parameter $\nu = 1/10$ (WAIC = 4067.733). On the other hand, the WAIC_{NF} values are very different for the three models. The best model amongst the three candidate models is the one fitted using the Matérn covariance function and smoothness parameter $\nu = 1$ ($\text{WAIC}_{\text{NF}} =$ 4149.58). While the PSIS-LOOIC is not calculated by INLA, our proposed non-factorisable model likelihood approach enables the calculation of the $\text{PSIS-LOOIC}_{\text{NF}}$. The $\text{PSIS-LOOIC}_{\text{NF}}$ show agreeable results with the WAIC_{NF}, where the model fitted using the Matérn covariance function and smoothness parameter $\nu = 1$ is the best model ($\text{PSIS-LOOIC}_{\text{NF}} = 4150.04$) out of the three candidate models.

Table 5.1. The WAIC_{NF}, $\text{PSIS-LOOIC}_{\text{NF}}$ and the WAIC computed by INLA for models fitted using Matérn covariance functions and different smoothness parameters

Matérn	INLA WAIC (p_{WAIC})	WAIC_{NF} ($p_{\text{WAIC}_{\text{NF}}}$)	$\text{PSIS-LOOIC}_{\text{NF}}$ ($p_{\text{PSIS-LOOIC}_{\text{NF}}}$)
$\nu = 1/2$	4067.866 (73.431)	4918.664 (74.164)	4919.146 (74.406)
$\nu = 1$	4067.967 (69.966)	4149.583 (32.433)	4150.036 (32.659)
$\nu = 1/10$	4067.733 (79.639)	5059.547 (92.563)	5060.459 (93.019)

We performed K-fold cross-validation ($K = 5$) for the three candidate models and observed no discernible differences in performance based on the validation statistics RMSE, MAE, and CRPS, as shown in Table 5.2. These results align with the WAIC results calculated by INLA in Table 5.1, indicating comparable model performance among the candidate models. This suggests a strength of the NF method in effectively discerning the most optimal candidate model from a pool of high-performing options. The consistent performance across different choices of the ν parameter in the covariance function suggests that all were valid candidates for this dataset. However, our NF model selection method successfully identified one model from among these candidates. This may imply that the WAIC_{NF} (and $\text{PSIS-LOOIC}_{\text{NF}}$) is more sensitive to the choice of covariance function, specifically the selection of ν for the `inla2.spde.pcmatern()` function within the INLA framework.

Table 5.2. 5-fold cross-validation statistics for models fitted using Matérn covariance functions and different smoothness parameters

Matérn	**RMSE**	**MAE**	**CRPS**
$\nu = 1/2$	0.197	0.158	0.112
$\nu = 1$	0.197	0.158	0.112
$\nu = 1/10$	0.198	0.158	0.112

Constructing models for the MCV1 dataset enable prediction in unobserved locations. Following the model fitted using the Matérn covariance function and smoothness parameter $\nu = 1$, the summary statistics from INLA are given in Table 5.3, and the high-resolution prediction and uncertainty surfaces are shown in Figure 5.5. From the model summary statistics, we see that poverty, temperature and traveltime all have a negative effect on MCV1 coverage whereas nightlights have a positive effect on MCV1 coverage. The spatial range is given in decimal degrees and translates to approximately 341 km; indicating the presence of residual spatial correlation in our model.

Table 5.3. Summary statistics of the model fitted with spatial random effects with covariance matrix that has elements derived from a Matérn covariance function and smoothness parameter $\nu = 1$.

	Mean	SD	2.5%	50.0%	97.5%
(Intercept)	-0.822	0.218	-1.254	-0.826	-0.367
Poverty	-0.104	0.070	-0.242	-0.103	0.034
Temperature	-0.198	0.111	-0.410	-0.201	0.028
log(Nightlights)	0.088	0.027	0.0360	0.088	0.140
log(Traveltime)	-0.199	0.061	-0.319	-0.199	-0.079
Spatial range ($\hat{r}_{sp}$)	3.079	0.950	1.803	2.874	5.468
Spatial variance ($\hat{\sigma}^2_\omega$)	0.433	0.136	0.246	0.404	0.774

From Figure 5.5a, the southern region exhibits higher MCV1 coverage, while the north-western part of Nigeria displays lower coverage. Moreover, the high-resolution prediction surface highlights pockets of high MCV1 coverage in the central, northern, and eastern regions of the country. Complementing this prediction surface, Figure 5.5b shows the high-resolution uncertainty surface, conveying corresponding posterior standard deviations. Interestingly, areas characterised by the lowest uncertainty align with predictions of lower MCV1 coverage, while regions of greater uncertainty correspond to our prediction of higher MCV1 coverage.

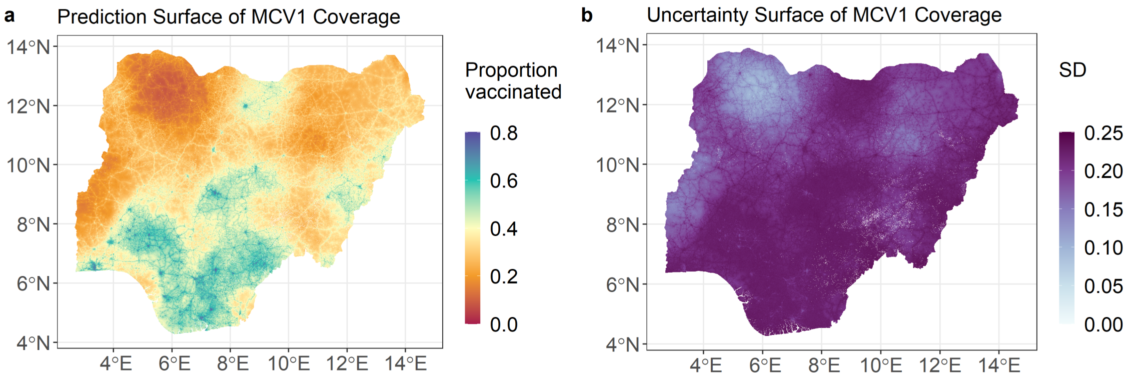

Figure 5.5. High-resolution 1 × 1km prediction (a) and uncertainty (b) surfaces for MCV1 coverage among children aged 12-23 months in Nigeria in 2018.

The high-resolution surfaces have been aggregated to a lower spatial resolution to enhance interpretability. Figure 5.6 illustrates this and depicts the surface obtained by aggregating the high-resolution data from the 1 × 1km grid level to the second-level sub-division district level provided by the GADM, which consists of the 774 local government areas (LGAs) within Nigeria.

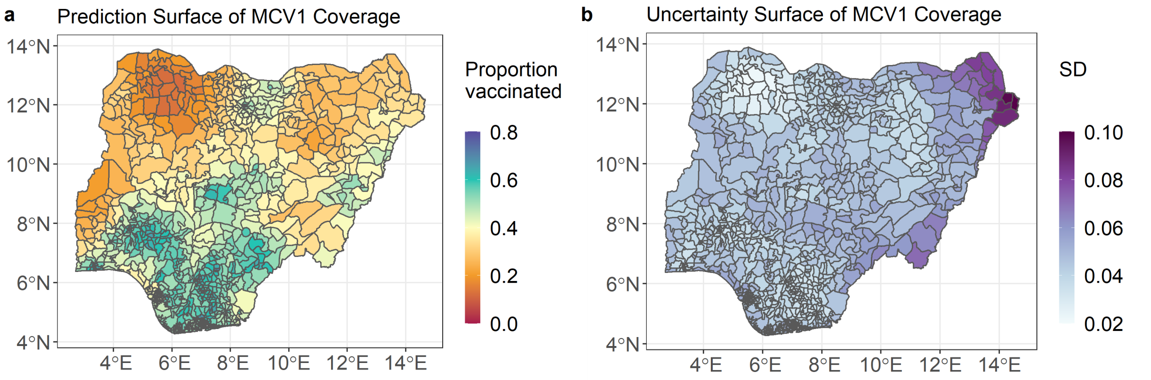

Figure 5.6. Proportion of MCV1-vaccinated children, calculated through population weighted aggregations from grid-level to district-level (local government areas).

5.6 Discussion

Our model (Table 5.3) showed the negative effects of traveltime to healthcare facilities and poverty on MCV1 coverage. Long travel times discourage families from making the trip to vaccinate their children and, in some cases, are simply impractical. This finding aligns with the work of Utazi et al. (2020). Similarly, poverty has a negative effect on MCV1 coverage. While this is intuitive, as poorer families often have less access to vaccines, the relationship is complex and involves various underlying factors. The poverty covariate used in this study is derived from a modelling work (Tatem et al., 2013), which encapsulates broader socioeconomic challenges that hinder vaccination uptake, including parental education and household sizes (Faruk et al., 2020, Ori et al., 2021).

Geospatial analysis at both the grid and district levels (Figures 5.5 and 5.6) corroborates existing literature, indicating higher measles severity in northern Nigeria. This region suffers from inadequate measles control strategies, including insufficient routine immunization efforts (Ameh et al., 2016, Saleh et al., 2016, Kagucia et al., 2018, Ibrahim et al., 2019). Addressing these challenges is crucial to improving MCV1 coverage.

Despite global and national increases in vaccination coverage, the goal of eliminating measles by 2020 remains unmet. Studies suggest that measles will persist as an endemic disease in Africa beyond this target year (Patel et al., 2019, Goodson, 2020, Gignoux et al., 2021). Challenges such as humanitarian crises, political instability (Ori et al., 2021), insufficient vaccine investments (Goodson, 2020, Gostin et al., 2020), and disruptions caused by the COVID-19 pandemic (Dixon et al., 2021, Gignoux et al., 2021) contribute to stagnant or declining vaccination rates in certain populations.

Addressing these inequities through targeted routine immunisation programs is crucial. Enhancing access to healthcare for vulnerable and impoverished populations, especially in cold spots of low MCV1 coverage, is essential to reduce measles-related mortality (Gignoux et al., 2021, Wariri et al., 2021). Strategies could include increasing public awareness about the benefits of vaccination and utilising innovative tools like measles rapid diagnostic tests(Grant et al., 2019, Goodson, 2020). As Durrheim (2020) asserted, measles outbreaks highlight the weaknesses in health systems, making them a critical measure of progress towards universal health coverage and public health accountability.

There are several limitations in this study related to the geostatistical models used to generate high spatial resolution prediction maps of MCV1 coverage. Firstly, the models were constructed using survey-based data, primarily from vaccine cards and maternal recall, which introduces the possibility of recall bias, a concern acknowledged in similar research (Wariri et al., 2021). Secondly, the selection of geospatial covariates is another limitation. A broader and more diverse set of covariates could provide additional insights into MCV1 coverage and should be considered in future investigations. However, this study's primary aim is to highlight the applicability of the WAIC_{NF} and the $\text{PSIS-LOOIC}_{\text{NF}}$ in determining the appropriate covariance function for our model. Despite these limitations, this study contributes to the methodology of model selection for point referenced data by demonstrating the practical application of the WAIC_{NF} and the $\text{PSIS-LOOIC}_{\text{NF}}$.

Regarding the choice of prior distribution, we specified $\beta \sim N(0, 2)$ for the regression coefficients in the geostatistical models. This choice was informed by examining the histograms of the standardised geospatial covariates, such as poverty, temperature, and the log transformations of travel time and nightlights, which closely resemble normal distributions with means close to zero (Figure 5.3). While a larger variance could have been chosen to make the prior less informative, we found that this specific choice of prior did not heavily influence our results. To verify this, we conducted additional tests using wider priors, specifically $N(0, 1000)$, and observed that the results remained consistent with those reported in this thesis. This indicates that our estimates are robust and not overly sensitive to the choice of prior distribution.

In this chapter, we demonstrated that, in a practical application setting, the WAIC computed by INLA failed to discern differences among various covariance functions. Conversely, our proposed WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ provided clear discernment for the optimal model among a pool of high-performing candidates. This suggests the efficacy of WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ for model selection tasks, particularly in choosing the appropriate covariance function for models of point referenced spatial data. Additionally, the smoothness parameter ν selected by our WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ aligns with the findings of Whittle (1954) and Lindgren and Rue (2015), specifically favouring $\alpha = 2$ (equivalent to $\nu = 1$) as a more suitable choice for models in the spatial domain $d = 2$.

Chapter 6

Calculating the WAIC for large spatial datasets using NNGP

Bayesian modelling of point referenced spatial data often require Markov chain Monte Carlo (MCMC) algorithms that repeatedly assess the full conditional density functions, and evaluate the likelihood functions, joint densities or conditional densities within the context of Gaussian process. However, computational operations involving matrices of dimensions $n \times n$ are susceptible to instability, increased computational cost or, in certain instances, become computationally infeasible. This computational challenge is commonly referred to as the “big n problem”.

To illustrate this challenge, we encountered difficulties when attempting to model the MCV1 dataset using the Stan framework in Chapter 5. Instead, we employed an approximation approach through the stochastic partial differential equation (SPDE) framework and the integrated nested Laplace approximation (INLA) method. Detailed discussion on these approaches can be found in Section 2.6. The strategies aimed at mitigating the big n problem encompass approximating the exact likelihood or devising models adept at handling large datasets.

Spatial covariance functions typically do not produce exploitable structures within their resulting matrices (Zhang et al., 2019). So, the general strategies for modelling large spatial datasets are to either exploit “low-rank” models or to leverage sparsity. Banerjee et al. (2014) have classified these strategies into three categories: approximate likelihood approaches, low-rank models and predictive process models.

A popular way of addressing the challenges posed by large spatial datasets is to devise models that bring about dimension reduction. Essentially, the spatial process is replaced with a dimension-reducing process that is constructed based on a representation in terms of the realisations of some latent process, over a smaller set of coordinates called "knots". This is called a low-rank model. The primary objective of low-rank models is to create spatial processes within a lower-dimensional subspace, thereby decreasing computational costs (Datta et al., 2016a).

Sparse methods include covariance tapering (Furrer et al., 2006, Kaufman et al., 2008), which introduces sparsity in the covariance matrix using compactly supported covariance functions. Covariance tapering is effective for parameter estimation and interpolation of the responses (Datta et al., 2016a). Furthermore, introducing sparsity to the precision matrix is a prevalent strategy in approximating the Gaussian process likelihood. Examples of this strategy include using Markov random fields (Rue and Held, 2005, Lindgren et al., 2011), products of lower-dimensional conditional distributions (Stein et al., 2004), or composite likelihoods (Bevilacqua and Gaetan, 2015).

We strongly recommend referring to the works of Sun et al. (2012), Bradley et al. (2016) and Heaton et al. (2019) for a comprehensive compilation of strategies pertaining to the handling of large spatial datasets. These references cover each strategy in great detail, provide practical implementation guidelines and conduct comparative analyses on the strategies, while highlighting their strengths and limitations.

In this chapter, we will provide detail on the three following strategies devised to address the big n problem: Gaussian predictive processes, stochastic partial differential equations and nearest-neighbour Gaussian processes. Amongst these three strategies, we will focus most on the latter.

6.1 Gaussian predictive processes

The idea behind Gaussian predictive process (GPP) is to consider some other locations, say $\mathbf{s}^*$, within the same study domain ($\mathbf{s}^* \in \mathcal{D}$) instead of the original observed locations $\mathbf{s}$. The key here is that the number of locations we choose for $\mathbf{s}^*$ is much smaller than the number of original locations $\mathbf{s}$. More formally, a set of "knots" is given by $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*\}$, such that $m \ll n$ in the entire collection of observed locations $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ (Banerjee et al., 2008).

Consider the following the model

$$Y(\mathbf{s}_i) = \mathbf{x}'(\mathbf{s}_i)\boldsymbol{\beta} + \omega(\mathbf{s}_i) + \epsilon(\mathbf{s}_i),$$

if $\boldsymbol{\omega} = \big(\omega(\mathbf{s}_1), \ldots, \omega(\mathbf{s}_n)\big)'$ is assumed to be $\boldsymbol{\omega} \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}})$, then $\boldsymbol{\omega}^* = \big(\omega(\mathbf{s}_1^*), \ldots, \omega(\mathbf{s}_m^*)\big)'$ is assumed to be $\boldsymbol{\omega}^* \sim N_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}^*})$, where $N_m(\cdot)$ is an m-dimensional multivariate normal distribution, and $\boldsymbol{\Sigma}_{\boldsymbol{\omega}^*}$ is an $m \times m$ covariance matrix. Now, the predictive process, denoted as $\tilde{\omega}(\mathbf{s}_i)$, is defined by Finley et al. (2009) as

$$\tilde{\omega}(\mathbf{s}_i) = E\big(\omega(\mathbf{s}_i)|\boldsymbol{\omega}^*\big) = \mathbf{c}'(\mathbf{s}_i, \mathbf{s}_j^*)\boldsymbol{\Sigma}_{\boldsymbol{\omega}^*}^{-1}\boldsymbol{\omega}^*. \tag{6.1}$$

Within Equation (6.1), there is the precision matrix, denoted as $\boldsymbol{\Sigma}_{\boldsymbol{\omega}^*}^{-1}$, which is the inverse of $\boldsymbol{\Sigma}_{\boldsymbol{\omega}^*}$. There are the spatial random effects of the knots, denoted as $\boldsymbol{\omega}^*$, as described above. There is the component $\mathbf{c}'(\mathbf{s}_i, \mathbf{s}_j^*)$, which is the covariance between $\mathbf{s}_i$ and knots $\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*$. Finley et al. further defined $\tilde{\boldsymbol{\omega}} = \big(\tilde{\omega}(\mathbf{s}_1), \ldots, \tilde{\omega}(\mathbf{s}_n)\big)'$, and

$$\begin{aligned} \tilde{\boldsymbol{\omega}} &\sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\omega}}}), \\ \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\omega}}} &= \mathbf{C}'\boldsymbol{\Sigma}_{\boldsymbol{\omega}^*}^{-1}\mathbf{C}, \end{aligned} \tag{6.2}$$

where $\mathbf{C}'$ is an $n \times m$ covariance matrix. The elements on the ith row of $\mathbf{C}'$ are calculated by $\mathbf{c}'(\mathbf{s}_i, \mathbf{s}_j^*)$, which denotes the covariance between $\mathbf{s}_i$ and knots $\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*$ (Finley et al., 2009). The predictive process approach exploits the Gaussian process assumption of the spatial random effects, and is implemented by replacing $\boldsymbol{\omega}(\mathbf{s})$ in our example model with $\tilde{\boldsymbol{\omega}}(\mathbf{s})$ from (6.1), such that

$$Y(\mathbf{s}_i) = \mathbf{x}'(\mathbf{s}_i)\boldsymbol{\beta} + \tilde{\omega}(\mathbf{s}_i) + \epsilon(\mathbf{s}_i).$$

We note that there is also a *modified predictive process*, proposed by Finley et al. (2009), to overcome the positive bias in the non-spatial error term of the model induced by the formulation of the predictive process. Finley et al. found that,

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\omega}}} \leq \boldsymbol{\Sigma}_{\boldsymbol{\omega}},$$

where $\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\omega}}}$ denotes the variance of $\tilde{\boldsymbol{\omega}}$ as provided in (6.2). The average bias underestimation is $E\big(\Sigma_{\boldsymbol{\omega}} - \mathbf{C}'\Sigma_{\boldsymbol{\omega}^*}^{-1}C\big)$, where $E(\cdot)$ denotes the element-wise expectation. The modified predictive process extends our example model above as

$$Y(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \tilde{\omega}(\mathbf{s}_i) + \tilde{\epsilon}(\mathbf{s}_i) + \epsilon(\mathbf{s}_i),$$

where $\tilde{\boldsymbol{\epsilon}} = \big(\tilde{\epsilon}(\mathbf{s}_1), \ldots, \tilde{\epsilon}(\mathbf{s}_n)\big)'$ are spatially independent random effects with the distribution

$$\tilde{\boldsymbol{\epsilon}} \sim N_n\big(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}} - \mathbf{C}'\boldsymbol{\Sigma}_{\boldsymbol{\omega}^*}^{-1}\mathbf{C}\big),$$

such that $\mathrm{Var}\big(\tilde{\boldsymbol{\omega}} + \tilde{\boldsymbol{\epsilon}}\big) = \boldsymbol{\Sigma}_{\boldsymbol{\omega}}$. To reiterate, the modified predictive process introduces an additional random effect to fix the positive bias in the non-spatial error term.

Selecting an appropriate $\mathcal{S}^*$ is important for the GPP strategy. This decision affects the computational cost and the estimates of the spatial range and variance components (Banerjee et al., 2008). However, selecting an appropriate $\mathcal{S}^*$ is a complicated process. $\mathcal{S}^*$ can be, but not necessarily, a subset of the original observed locations (Banerjee et al., 2008). $\mathcal{S}^*$ can be the centroids of the grids from equally dividing the spatial domain (Finley et al., 2009). We recommend Xia et al. (2006), Banerjee et al. (2008) and Finley et al. (2009) for further discussions on knot selection strategies.

6.2 Stochastic partial differential equations

The stochastic partial differential equations (SPDE) method represents another strategic approach developed to address the challenges posed by large spatial datasets. It should be noted that we have discussed this method in detail in Section 2.6.1. Therefore, we will refrain from providing further information on the SPDE approach and instead recommend readers to revisit the aforementioned section for a more detailed discussion of this method.

Conceptually, the SPDE method can be understood as projecting a continuous Gaussian field as a discrete Gaussian Markov random fields, represented by a mesh. Formally, the SPDE approach is based on the equivalence between Matérn covariance fields and the stochastic partial differential equations (Heaton et al., 2019).

Again, consider the model

$$Y(\mathbf{s}_i) = \mathbf{x}'(\mathbf{s}_i)\boldsymbol{\beta} + \omega(\mathbf{s}_i) + \epsilon(\mathbf{s}_i).$$

The SPDE approximates the spatial random effect $\omega(\mathbf{s}_i)$ as

$$\omega(\mathbf{s}_i) \approx \tilde{\omega}(\mathbf{s}_i) = \sum_{k=1}^{K} h_k(\mathbf{s}_i)\omega_k^*,$$

where the "basis functions", denoted as $h_k(\mathbf{s}_i)$, are chosen to be piece-wise linear on a triangulation (the mesh) of the domain, and ω_k^* denotes a coefficient (Rue and Held, 2005, Lindgren et al., 2011). The optimal joint distribution for ω_k^* is obtained through a finite element construction, which leads to a sparse inverse covariance matrix called the precision matrix (Heaton et al., 2019). The elements of the precision matrix are polynomials in the precision and inverse range parameters, with sparse matrix coefficients determined by the choice of triangulation. Computational and storage cost for the posterior predictions and multivariate Gaussian likelihood of a spatial Gaussian Markov random field is $O(K^{3/2})$ (Heaton et al., 2019).

6.3 Nearest-neighbour Gaussian process

The nearest-neighbour Gaussian process (NNGP) is a well-defined and highly scalable family of Gaussian processes that offers an efficient solution to the challenges posed by large spatial datasets (Datta et al., 2016a). The NNGP is defined from the conditional specification of the joint distribution of spatial random effects, and it achieves scalability by generating finite dimensional Gaussian densities with sparse matrices.

To illustrate the NNGP, let us consider the following

$$\begin{aligned} Y(\mathbf{s}_i) &= \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \omega(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \\ \boldsymbol{\omega} &\sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}), \\ \boldsymbol{\epsilon} &\sim N_n(\mathbf{0}, \tau^2 I_n), \end{aligned}$$

where the observations, denoted $\mathbf{y}(\mathbf{s})$, at locations $\mathbf{s} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is modelled by the covariates $\mathbf{x}(\mathbf{s})$ at locations $\mathbf{s}$, the spatial random effects $\boldsymbol{\omega} = \big(\omega(\mathbf{s}_1), \omega(\mathbf{s}_2), \ldots, \omega(\mathbf{s}_n)\big)'$ that follows a zero-mean Gaussian process (GP), and unstructured random effects $\boldsymbol{\epsilon} = \big(\epsilon(\mathbf{s}_1), \epsilon(\mathbf{s}_2), \ldots, \epsilon(\mathbf{s}_n)\big)'$, that capture random noise. The GP in the setup is an n-dimensional multivariate normal distribution, as notated by $N_n(\cdot)$. Here, I_n

denotes an $n \times n$ identity matrix, and $\mathbf{\Sigma}_{\boldsymbol{\omega}}$ is an $n \times n$ covariance matrix with elements calculated from a covariance function, such as the Matérn covariance function (2.6) or the exponential covariance function (2.5), as detailed in Section 2.3.3. Recall from Section 2.4.3 that the setup above can be equivalently expressed as

$$\begin{aligned}\mathbf{Y} &\sim N_n\big(X\boldsymbol{\beta} + \boldsymbol{\omega}, \tau^2 I_n\big),\\ \boldsymbol{\omega} &\sim N_n(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\omega}}).\end{aligned}$$

Finley et al. (2017) and Zhang (2018) express this hierarchical model as

$$N_n\big(\mathbf{Y}|X\boldsymbol{\beta} + \boldsymbol{\omega}, \tau^2 I\big)\ N_n\big(\boldsymbol{\omega}|\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\omega}}\big)\ p(\boldsymbol{\psi}), \tag{6.3}$$

where $p(\boldsymbol{\psi})$ are the prior distributions for the parameters $\boldsymbol{\psi} = (\sigma^2_{\boldsymbol{\omega}}, \phi, \nu, \boldsymbol{\beta}, \tau^2)'$. Furthermore, $\boldsymbol{\psi}$ is sampled through the posterior distribution,

$$p\big(\boldsymbol{\psi}|\mathbf{y}(\mathbf{s})\big) \propto N_n\big(\mathbf{y}(\mathbf{s})|\mathbf{x}(\mathbf{s})'\boldsymbol{\beta}, \mathbf{\Sigma}_{\boldsymbol{\omega}} + \tau^2 I_n\big)p(\boldsymbol{\psi}). \tag{6.4}$$

The primary challenge encountered when modelling (6.4) is the computational complexities associated with the dense covariance matrix, denoted as $\mathbf{\Sigma}_{\boldsymbol{\omega}}$. The Bayesian modelling of (6.4) involves the computation of the inverse of the covariance matrix, referred to as the precision matrix and denoted as $\mathbf{\Sigma}_{\boldsymbol{\omega}}^{-1}$, as well as the determinant of the covariance matrix, denoted as $\det(\mathbf{\Sigma}_{\boldsymbol{\omega}})$. Both operations require $O(n^3)$ floating point operations (flops). Moreover, the storage requirements for $\mathbf{\Sigma}_{\boldsymbol{\omega}}$ encompass a dynamic memory allocation of $O(n^2)$, as documented in prior works (Datta et al., 2016a, Finley et al., 2017). Evidently, as the number of sites n increases within the spatial dataset, handling $\mathbf{\Sigma}_{\boldsymbol{\omega}}$ becomes increasingly computationally expensive.

The NNGP, introduced in the works of Datta et al. (2016a) and Datta et al. (2016b), provides a solution for the computational challenges mentioned above. The NNGP accomplishes this by approximating the spatial covariance matrix through the utilisation of the covariance structure among neighbouring points. In contrast to evaluating the entire dataset, the NNGP identifies a subset of nearest neighbours for each site, based on their spatial proximity. More explicitly, a model employing the the NNGP can be expressed as

$$N_n\big(\mathbf{y}(\mathbf{s})|\mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + \boldsymbol{\omega}(\mathbf{s}), \tau^2 I_n\big)\ N_n(\boldsymbol{\omega}(\mathbf{s})|\mathbf{0}, \tilde{\mathbf{\Sigma}}_{\boldsymbol{\omega}})\ p(\boldsymbol{\psi}). \tag{6.5}$$

Notice that Equation (6.5) is similar to Equation (6.3). The distinguishing factor between the two equations is in their covariance matrix. The NNGP (6.5) denotes the covariance matrix as $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}$, and it is elaborated as follows.

The precision matrix of $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}$ is defined by Finley et al. as

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1} = (I - \tilde{A})'\tilde{D}^{-1}(I - \tilde{A}). \tag{6.6}$$

The likelihood function of $\boldsymbol{\omega}(\mathbf{s})$ based on $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1}$ is a good approximation and a computationally efficient approximation to the likelihood function of $\boldsymbol{\omega}(\mathbf{s})$ with $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}^{-1}$. Furthermore, the computation of the density $N_n(\boldsymbol{\omega}(\mathbf{s})|\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}})$ requires $O(n)$ flops (Finley et al., 2017, 2020).

To derive (6.6), we need to compute the matrices $\tilde{A}$ and $\tilde{D}$. Here, $\tilde{A}$ denotes a sparse and strictly lower triangular matrix, characterised by a maximum of m non-zero entries in each row, where $m \ll n$. The component $\tilde{D}$ denotes a diagonal matrix (Finley et al., 2017, Zhang et al., 2019). They are given as

$$\begin{aligned}\tilde{A}\big(i, N(\mathbf{s}_i)\big) &= C_{\boldsymbol{\theta}}\big(\mathbf{s}_i, N(\mathbf{s}_i)\big)\big(C_{\boldsymbol{\theta}}(\mathbf{s}_i, N(\mathbf{s}_i)\big) + \tau^2 I_n)^{-1},\\ \tilde{D}(i,i) &= C_{\boldsymbol{\theta}}(\mathbf{s}_i, \mathbf{s}_i) + \tau^2 - C_{\boldsymbol{\theta}}\big(\mathbf{s}_i, N(\mathbf{s}_i)\big)\big(C_{\boldsymbol{\theta}}(\mathbf{s}_i, N(\mathbf{s}_i)\big) + \tau^2 I_n)^{-1} C_{\boldsymbol{\theta}}\big(N(\mathbf{s}_i), \mathbf{s}_i\big),\end{aligned}$$

where $N(\mathbf{s}_i)$ denotes the m closest points to $\mathbf{s}_i$ among the locations indexed less than i. In the expressions for matrices $\tilde{A}$ and $\tilde{D}$ above, we follow the convention used in the works of Zhang (2018) and Zhang et al. (2019) for the covariance function, denoted as $C_{\boldsymbol{\theta}}(\cdot,\cdot)$. Note that $C_{\boldsymbol{\theta}}(\cdot,\cdot)$ expresses a covariance function with parameters $\boldsymbol{\theta} = (\sigma^2_{\boldsymbol{\omega}}, \phi, \nu)'$, and is equivalent to $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$. However, expressing the covariance in this manner offers the advantage of explicitly depicting the sites supplied as arguments within the covariance function.

Who are our neighbours? Formally, the neighbours are defined by Finley et al. (2020) as follows,

$$\begin{aligned} N(\mathbf{s}_1) &= \{\},\\ N(\mathbf{s}_i) &= \min(m, i-1) \text{ nearest neighbours of } \mathbf{s}_i,\\ N(\mathbf{s}) &= m \text{ nearest neighbours of } \mathbf{s} \in \mathbb{R}.\end{aligned}$$

Conceptually, the neighbours can be understood as follows. Consider $n = 20$ sites generated from a unit square, as illustrated in Figure 6.1. We first index the points based on coordinate 1, as denoted alongside the points within the figure. Now, suppose we wish to identify the 5 nearest neighbours of the observation located at index 15. Within the subset of sites that are indexed lower than 15, we determine the 5 closest sites based on the Euclidean distance. In Figure 6.1, they are sites corresponding to indices 7, 10, 11, 12 and 13.

Notice that the choice of neighbours depended on the ordering and indexing of the sites. We could have indexed the points based on coordinate 2 instead of coordinate 1. However, Datta et al. (2016a) suggested that the model is not sensitive to the choice of ordering in most empirical findings.

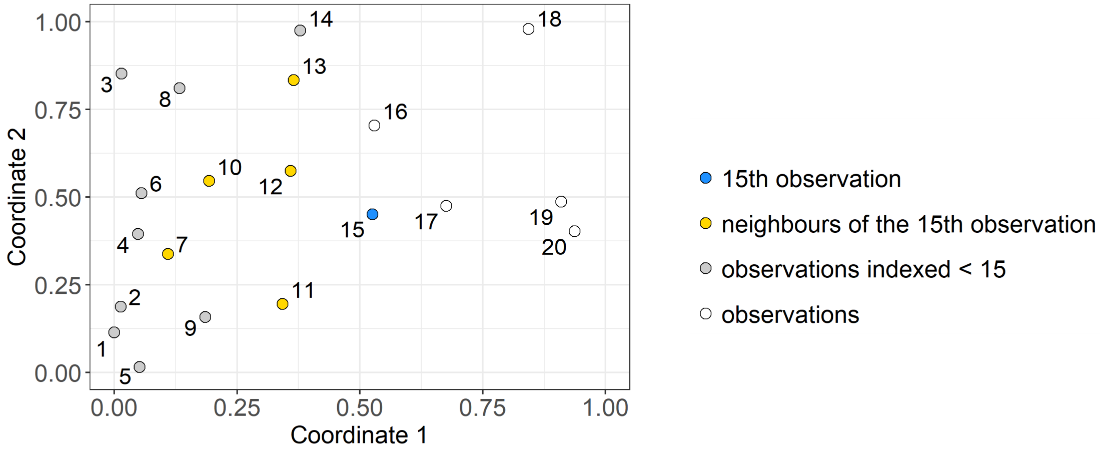

Figure 6.1. Neighbours in NNGP

How many neighbours do we need? Increasing the number of neighbours may lead to an increase in the computational run time for NNGP models. Consequently, the choice of the number of neighbours for NNGP models is an important decision. While specifying fewer neighbours can effectively reduce computational time, the NNGP model may become inaccurate, and may underestimate the spatial random effects (Quiroz et al., 2023). Conversely, Quiroz et al. found no clear patterns indicating that a greater number of neighbours lead to an improved model. Hence, the decision regarding the number of neighbours is a delicate balance between selecting as few neighbours as possible, while ensuring a reasonably accurate approximation of the true spatial process.

It is noteworthy that within the existing literature, the explicit reasoning behind the choice of the number of neighbours is frequently omitted (Finley et al., 2017, Zhang, 2018, Zhang et al., 2019, Finley et al., 2020). However, we observe that the chosen number of neighbours are often very small, when compared to the entire dataset. Datta et al. have suggested that a range of 10 to 15 neighbours can be sufficient, and may produce performance similar to using the all available sites; although a larger number of neighbours may be still be beneficial for exceptionally extensive datasets.

It should be noted that all current implementations of the NNGP model assume that the response is Gaussian. The response and the conjugate NNGP model explicitly rely on Gaussian distributions to derive the marginal distribution for the response. Finley et al. (2020) noted that although closed form marginal distributions are not available for non-Gaussian responses, the latent NNGP model can be conceptually extended to non-Gaussian setting.

6.4 Modelling MCV1 coverage using NNGP

In Chapter 5, we employed the stochastic partial differential equations (SPDE) approach (Lindgren et al., 2011) and the integrated nested Laplace approximation (INLA) method (Rue et al., 2009, Martins et al., 2013), as detailed in Section 2.6, for Bayesian modelling for the MCV1 dataset. It would not have been possible to use the Stan framework for Bayesian modelling, even though the dataset consist of a moderately sample size of $n = 1319$. However, using the NNGP approach, we are poised to reattempt Bayesian modelling within the Stan framework.

First, recall the model for the MCV1 dataset is the following,

$$\begin{aligned} Y(\mathbf{s}_i) &\sim \text{Binomial}\big(n(\mathbf{s}_i), p(\mathbf{s}_i)\big), \\ \text{logit}\big(p(\mathbf{s}_i)\big) &= \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \omega(\mathbf{s}_i), \end{aligned}$$

for $i = 1, \ldots, n$. As before, $\omega(\mathbf{s}_i)$ denotes the spatial random effect, which captures the spatial autocorrelation and describes the spatial residual structure amongst the data. In Section 5.3, the spatial random effects follows an n-dimensional multivariate normal distribution with zero-mean and an $n \times n$ covariance matrix that has elements calculated from a covariance function, as described in Section 2.3.3. However, as noted in Section 6.3, implementation of the the NNGP model necessitate the assumption of a Gaussian response. In order to align with this prerequisite, we introduce an adjustment to our model for the MCV1 dataset as follows,

$$\text{logit}\big(p(\mathbf{s}_i)\big) \sim N(\mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + \omega(\mathbf{s}_i), \tau^2). \tag{6.7}$$

Furthermore, the spatial random effects $\boldsymbol{\omega} = \big(\omega(\mathbf{s}_1), \ldots, \omega(\mathbf{s}_n)\big)'$ now follows an NNGP, such that

$$\boldsymbol{\omega} \sim \text{NNGP}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}), \tag{6.8}$$

where $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\omega}}^{-1}$ denotes the precision matrix, as defined in Equation (6.6).

To implement the model for the MCV1 dataset and (6.8) within the Stan framework, we follow the code developed by Zhang (2018), and used $m = 10$ neighbours. We also specify the following prior distributions for the parameters: a $N(0, 2)$ for the spatial variance $\sigma^2_{\boldsymbol{\omega}}$, the iid variance τ^2 and the regression coefficients, and a $\Gamma(2, 2)$ for the spatial decay parameter ϕ. We fit the model in Stan using 2 chains and 5000 iterations per chain, of which 2500 are burn-in iterations. The target average acceptance probability and the maximum tree-depth are adjusted appropriately to avoid problems with convergence or mixing. We run the model on a Windows machine with 4-core Intel process and 8GB random-access memory (RAM), and using R version 4.0.4 (R Core Team, 2021).

Table 6.1. Summary statistics for the Bayesian model fitted to the MCV1 dataset using the Stan framework.

	Mean	SD	2.5%	50.0%	97.5%	ESS
(Intercept)	-0.36	0.09	-0.54	-0.36	-0.18	93
Poverty	-0.10	0.08	-0.25	-0.10	0.04	477
Temperature	-0.27	0.07	-0.40	-0.28	-0.13	282
log(Nightlights)	0.19	0.05	0.08	0.19	0.29	2340
log(Traveltime)	-0.14	0.05	-0.24	-0.14	-0.05	1894
Spatial decay (ϕ)	0.03	0.00	0.02	0.03	0.04	150
Spatial variance ($\sigma^2_{\boldsymbol{\omega}}$)	0.66	0.08	0.52	0.65	0.82	701

Figure 6.2 shows stability in the MCMC chains and demonstrates good mixing between chains. Additionally, Table 6.1 presents summary statistics that are within our expectation, based on the results observed from the model fitted using the INLA

method and the reasoning provided in Chapter 5. Specifically, the variables poverty, temperature and traveltime exhibit a negative effect on MCV1 coverage, whereas nightlights have a positive effect on MCV1 coverage.

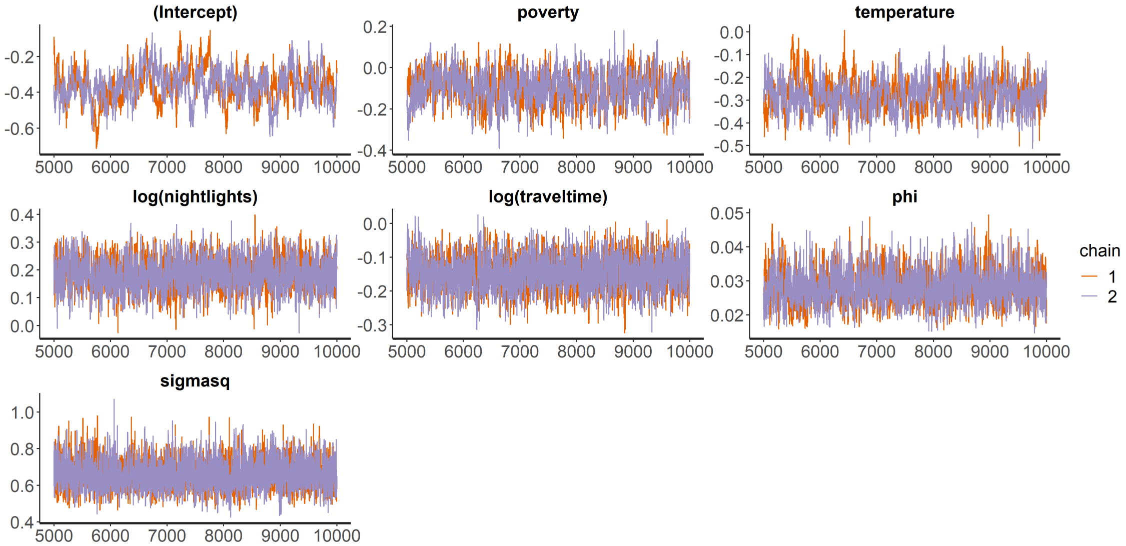

Figure 6.2. Trace plots of the Bayesian model for the MCV1 dataset from Stan

After fitting the Bayesian model within the Stan framework, we proceed to compute the Watanabe-Akaike information criteria (WAIC) and the Pareto smoothed importance sampling leave-one-out cross validation information criterion (PSIS-LOOIC). A detailed exposition of these selection criteria are provided in Sections 2.7.2 and 2.7.3, respectively. Specifically, their computation can utilise both the log likelihoods extracted from Stan, and the non-factorisable model log likelihoods following Algorithm 1, as detailed in Section 3.5.

Our calculations yielded the following results: WAIC $= 4065.479$ ($p_{\text{WAIC}} = 111.098$) and PSIS-LOOIC $= 4068.376$ ($p_{\text{PSIS}-\text{LOOIC}} = 112.546$). Additionally, the WAIC$_{\text{NF}}$ $= 3715.508$ ($p_{\text{WAIC}_{\text{NF}}} = 17.104$) and the PSIS-LOOIC$_{\text{NF}}$ $= 3715.614$ ($p_{\text{PSIS-LOOIC}_{\text{NF}}} = 17.156$) for the model fitted using the Matérn covariance function with smoothness parameter $\nu = 1$. These values hold particular significance in scenarios necessitating the comparison of candidate models characterised by different covariance functions, as discussed in Section 4.4.

Chapter 7

Conclusions and future work

7.1 Conclusions

In this thesis, we have introduced an alternative approach for the computation of the Watanabe-Akaike information criterion (WAIC). The computation utilises the log likelihood function tailored for non-factorisable models. We call this the WAIC_{NF}. The WAIC_{NF} is based on log likelihoods designed to accommodate outcomes that exhibit conditional dependencies. Such dependencies are commonly encountered in point referenced spatial data. Furthermore, we have expanded the application of these non-factorisable model log likelihoods to facilitate the calculation of the Pareto smoothed importance sampling leave-one-out cross-validation information criterion (PSIS-LOOIC), which we refer to as the $\text{PSIS-LOOIC}_{\text{NF}}$.

The motivation behind introducing the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ stems from the limitation of the conventional approach to formulating the WAIC, which relies on a log likelihood that assumes conditional independence among the outcomes of the dataset. However, this assumption does not hold for spatial models for point referenced datasets. In the context of point referenced spatial dataset, we assume that the outcomes located across the spatial domain of interest exhibit spatial dependence, a fundamental characteristic of models for point referenced spatial data. Consequently, models for point referenced spatial data violate the conditional independence assumption that the conventional WAIC calculation relies upon.

In Chapter 3, we provided Algorithm 1 to compute the non-factorisable model log likelihoods, which can be used within Equations (2.26) and (2.31) to compute the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, respectively. In Chapter 4, we investigated the utility of our proposed WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ in model selection tasks and variable selection tasks. We found that our proposed WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ are more suitable for model selection tasks, specifically involving covariance function selection. However, our proposed WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ is not as suitable for variable selection tasks as our results indicate that the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ tend to prefer models that include all available covariates, even when it is not appropriate to do so.

Following this, we demonstrated the practical implementation of our proposed WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ in a spatial modelling scenario using real-world data sourced from the 2018 Nigeria Demographic Health Survey program, focusing on first-dose measles-containing vaccine (MCV1) coverage. In this context, we constructed three distinct models, each characterised by a different covariance function. Subsequently, we computed the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, and extracted the WAIC calculated by INLA. Our analysis revealed notable discrepancies in the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ values across the different models, in contrast to the marginal variations observed in the WAIC values computed by INLA. Leveraging the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, we successfully identified a spatial model with optimally specified covariance functions for the MCV1 dataset. Our findings from this chapter affirm the effectiveness of our proposed criteria in facilitating informed model selection, particularly in spatial modeling applications involving varying covariance specifications.

As an extension to the practical application of the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, we implemented the nearest-neighbour Gaussian process (NNGP) approach within the Stan framework as an alternative to address computational challenges of fitting spatial models in the Bayesian context. By leveraging NNGP, we achieved a computationally efficient mean of fitting the spatial model for the MCV1 dataset that would have otherwise remained unfeasible. Subsequently, we computed the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, providing a comprehensive evaluation of the suitability for our real-world spatial model.

This thesis contributes to existing knowledge on model selection methodologies for Bayesian models of point referenced spatial data with the proposed calculation of the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$. We demonstrated that these criteria excel in model selection tasks, especially when choosing among models with different covariance functions. We also demonstrated the integration of these criteria in popular Bayesian spatial model fitting frameworks, including INLA and the NNGP approach in Stan. This incorporation enhances the relevance of our proposed selection criteria while addressing computational challenges, thereby offering a robust model selection framework tailored for spatial models for point referenced data.

7.2 Future work

Future work can focus on extending Algorithm 1, as presented in Section 3.5, to incorporate more flexible covariance structures. This enhancement would allow for a broader selection of forms for WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, including the capability to handle point referenced spatiotemporal data. Integrating temporal dimensions into point referenced spatiotemporal datasets introduces an additional layer of complexity to the covariance function, as it requires consideration of both spatial and temporal dynamics. Selecting an appropriate covariance function is crucial for accurately modeling point referenced spatiotemporal data, given its role in effectively capturing the underlying dependencies.

Consider the following model,

$$Y(\mathbf{s}_i, t) = \mathbf{x}(\mathbf{s}_i, t)'\boldsymbol{\beta} + \epsilon(\mathbf{s}_i, t), \tag{7.1}$$

where the random variables located at site $\mathbf{s}_i$ and time point t, for $i = 1, \dots, n$ and $t = 1, \dots, T$, are modelled using covariates $\mathbf{x}(\mathbf{s}_i, t)$ and an error term $\epsilon(\mathbf{s}_i, t)$. The error term $\epsilon(\mathbf{s}_i, t)$ is assumed to follow a zero-mean spatiotemporal Gaussian process with a separable covariance structure,

$$\text{Cov}\big(\epsilon(\mathbf{s}_i, t_k), \epsilon(\mathbf{s}_j, t_l)\big) = \sigma^2 \rho_s(\cdot)\rho_t(\cdot), \tag{7.2}$$

where $\rho_s(\cdot)$ denotes the correlation function within the spatial domain, and $\rho_t(\cdot)$ denotes the correlation function within the temporal domain. However, the separable covariance structure (7.2) does not account for the interaction between space and time in the dependence structure. Despite this limitation, Sahu (2022) suggested that the separable covariance structure may still be useful for model comparison purposes.

Alternatively, we may assume a temporally independent Gaussian process for the spatiotemporal process. We can rewrite equation (7.1) as follows,

$$Y(\mathbf{s}_i, t) = \mathbf{x}(\mathbf{s}_i, t)'\boldsymbol{\beta} + \omega(\mathbf{s}_i, t) + \epsilon(\mathbf{s}_i, t), \tag{7.3}$$

where,

$$\mathrm{Cov}\big(\omega(\mathbf{s}_i, t), \omega(\mathbf{s}_j, t)\big) = \sigma_\omega^2 \rho(\cdot), \tag{7.4}$$

and $\rho(\cdot)$ denotes the correlation function within the spatial domain. Furthermore, we can express (7.3) using the following hierarchical model setup,

$$\mathbf{Y}_t | \boldsymbol{\omega}_t \sim N_n(\boldsymbol{\mu}_t + \boldsymbol{\omega}_t, \sigma_\epsilon^2 I_n),$$
$$\boldsymbol{\omega}_t \sim N_n(\mathbf{0}, \sigma_{\boldsymbol{\omega}}^2 \boldsymbol{\Sigma}_{\boldsymbol{\omega}}),$$

independently for $t = 1, \ldots, T$.

The challenge in implementing Algorithm 1 within the context of Bayesian models for point referenced spatiotemporal data lies in specifying the covariance matrix. While the algorithm detailed in Chapter 3 can handle basic covariance structures, extending it to incorporate more complex and flexible structures is essential for future work. At that stage, careful consideration is required to select the most suitable covariance structure for this specific task.

Another important area for future work involves improving the $\mathrm{WAIC_{NF}}$ penalty term. Currently, the $\mathrm{WAIC_{NF}}$ shows strong performance in covariance function selection but tends to include all available covariates and produces sub-optimal results for variable selection tasks, as shown in Section 4.4. This suggests room for improvement in our proposed $\mathrm{WAIC_{NF}}$. Utilising the structure of WAIC penalty functions p_{WAIC1} (Watanabe and Opper, 2010) and p_{WAIC2} (Gelman et al., 2014), our proposed $\mathrm{WAIC_{NF}}$ tends to include all covariates, indicating the need for a stricter penalty function. One potential solution is to develop a new, more stringent penalty function for the $\mathrm{WAIC_{NF}}$ that better balances model complexity and variable selection. By adjusting the penalty term, we can potentially reduce the tendency to include all covariates, thereby enhancing the model's ability to select the most relevant variables.

Extending the algorithm to incorporate a more general covariance function and developing a new, more stringent penalty term represent valuable avenues for further research and development. These enhancements highlight a promising direction for significantly improving the performance and robustness of the $\mathrm{WAIC_{NF}}$ and $\mathrm{PSIS\text{-}LOOIC_{NF}}$.

Bibliography

M. Abramowitz and I. A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

H. Akaike. Information theory and an extention of the maximum likelihood principle. In *2nd International Symposium on Information Theory, 1973*, pages 267–281. Akademiai Kiado, 1973.

C. A. Ameh, M. B. Sufiyan, M. Jacob, N. E. Waziri, and A. T. Olayinka. Evaluation of the measles surveillance system in kaduna state, nigeria (2010-2012). *Online journal of public health informatics*, 8(3), 2016.

K. A. Ayalew, S. Manda, and B. Cai. A comparison of bayesian spatial models for hiv mapping in south africa. *International Journal of Environmental Research and Public Health*, 18(21):11215, 2021.

A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):579–602, 1999.

G. J. Babu and E. D. Feigelson. Spatial point processes in astronomy. *Journal of statistical planning and inference*, 50(3):311–326, 1996.

S. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848, 2008.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC press, 2014.

M. Bass. *Efficient parameterisation of hierarchical Bayesian models for spatially correlated data*. PhD thesis, University of Southampton, 2015.

J. O. Berger and L. R. Pericchi. The intrinsic bayes factor for linear models. *Bayesian statistics*, 5:25–44, 1996.

L. M. Berliner. Hierarchical bayesian time series models. In *Maximum Entropy and Bayesian Methods: Santa Fe, New Mexico, USA, 1995 Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods*, pages 15–22. Springer, 1996.

J. M. Bernardo. Expected information as expected utility. *the Annals of Statistics*, pages 686–690, 1979.

J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20, 1991.

M. Bevilacqua and C. Gaetan. Comparing composite likelihood methods based on pairs for spatial gaussian random fields. *Statistics and Computing*, 25:877–892, 2015.

M. Blangiardo and M. Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

G. E. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

J. R. Bradley, N. Cressie, and T. Shi. A comparison of spatial predictors when datasets could be very large. *Statistics Survey*, 10:100–131, 2016.

K. Buraham and D. Anderson. Model selection and inference: An information-theoretic approach, 1998.

C. R. Burgert, J. Colston, T. Roy, and B. Zachary. Geographic displacement procedure and georeferenced data release policy for the demographic and health surveys, 2013.

P.-C. Bürkner, J. Gabry, and A. Vehtari. Efficient leave-one-out cross-validation for bayesian non-factorized normal and student-t models. *Computational Statistics*, 36 (2):1243–1261, 2021.

B. Cai, A. B. Lawson, M. M. Hossain, J. Choi, R. S. Kirby, and J. Liu. Bayesian semiparametric model with spatially–temporally varying coefficients selection. *Statistics in medicine*, 32(21):3670–3685, 2013.

G. Casella and E. I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

W. Cheney and D. Kincaid. Linear algebra: Theory and applications. *The Australian Mathematical Society*, 110:544–550, 2009.

N. Cressie. *Statistics for spatial data.* John Wiley & Sons, 2015.

N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1): 209–226, 2008.

T. N. Croft, A. M. J. Marshall, and C. K. Allen. Guide to dhs statistics, 2018.

F. Curtale, F. Perrelli, J. Mantovani, M. C. d. Atti, A. Filia, L. Nicoletti, F. Magurano, P. Borgia, and D. Di Lallo. Description of two measles outbreaks in the lazio region, italy (2006-2007). importance of pockets of low vaccine coverage in sustaining the infection. *BMC Infectious Diseases*, 10(1):1–9, 2010.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812, 2016a.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. On nearest-neighbor gaussian process models for massive spatial data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(5):162–171, 2016b.

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

M. G. Dixon, M. Ferrari, S. Antoni, X. Li, A. Portnoy, B. Lambert, S. Hauryski, C. Hatcher, Y. Nedelec, M. Patel, et al. Progress toward regional measles elimination—worldwide, 2000–2020. *Morbidity and Mortality Weekly Report*, 70(45):1563, 2021.

D. Draper and M. Krnjajic. Bayesian model specification. In *Presentation at eighth Valencia/ISBA world meeting on Bayesian statistics*, volume 171, 2006.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

E. W. Duncan and K. L. Mengersen. Comparing bayesian spatial models: Goodness-of-smoothing criteria for assessing under-and over-smoothing. *PloS one*, 15(5): e0233019, 2020.

W. L. Dunn and J. K. Shultis. *Exploring monte carlo methods*. Elsevier, 2022.

D. N. Durrheim. Measles eradication—retreating is not an option. *The Lancet Infectious Diseases*, 20(6):e138–e141, 2020.

M. L. Eaton. *Multivariate statistics: a vector space approach*. John Wiley and Sons, 1983.

A. S. Faruk, A. S. Adebowale, M. S. Balogun, L. Taiwo, O. Adeoye, S. Mamuda, and N. E. Waziri. Temporal trend of measles cases and impact of vaccination on mortality in jigawa state, nigeria, 2013-2017: a secondary data analysis. *The Pan African Medical Journal*, 35(Suppl 1), 2020.

A. O. Finley, H. Sang, S. Banerjee, and A. E. Gelfand. Improving the performance of predictive process modeling for large datasets. *Computational statistics & data analysis*, 53(8):2873–2884, 2009.

A. O. Finley, A. Datta, B. C. Cook, D. C. Morton, H. E. Andersen, and S. Banerjee. Applying nearest neighbor gaussian processes to massive spatial data sets forest canopy height prediction across tanana valley alaska. *arXiv preprint arXiv:1702.00434*, 7, 2017.

A. O. Finley, A. Datta, and S. Banerjee. spnngp r package for nearest neighbor gaussian process models. *arXiv preprint arXiv:2001.09111*, 2020.

J. Fox and G. Monette. Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417):178–183, 1992.

R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3): 502–523, 2006.

A. C. Gatrell, T. C. Bailey, P. J. Diggle, and B. S. Rowlingson. Spatial point pattern analysis and its application in geographical epidemiology. *Transactions of the Institute of British geographers*, pages 256–274, 1996.

S. Geisser. An introduction to predictive inference, 1993.

A. E. Gelfand. Hierarchical modeling for spatial data problems. *Spatial statistics*, 1: 30–39, 2012.

A. E. Gelfand and D. K. Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3): 501–514, 1994.

A. E. Gelfand and S. K. Ghosh. Model choice: a minimum posterior predictive loss approach. *Biometrika*, 85(1):1–11, 1998.

A. E. Gelfand, D. K. Dey, and H. Chang. Model determination using predictive distributions with implementation via sampling-based methods. Technical report, Stanford Univ CA Dept of Statistics, 1992.

A. E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. *Handbook of spatial statistics*. CRC press, 2010.

A. Gelman, G. O. Roberts, and W. R. Gilks. Efficient metropolis jumping rules. *Bayesian statistics 5*, 5:599–608, 1996.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2003. ISBN 9781420057294. URL `https://books.google.co.uk/books?id=TNYhnkXQSjAC`.

A. Gelman, J. Hwang, and A. Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.

E. Gignoux, L. Esso, and Y. Boum. Measles: the long walk to elimination drawn out by covid-19. *The Lancet Global Health*, 9(3):e223–e224, 2021.

W. R. Gilks, A. Thomas, and D. J. Spiegelhalter. A language and program for complex bayesian modelling. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 43(1):169–177, 1994.

Global Administrative Areas. Gadm database of global administrative areas, 2018. URL `www.gadm.org`. version 4.1.

T. Gneiting. Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600, 2002.

T. Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011.

T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

V. Gómez-Rubio. *Bayesian inference with INLA*. CRC Press, 2020.

J. L. Goodson. Recent setbacks in measles elimination: the importance of investing in innovations for immunizations. *The Pan African Medical Journal*, 35(Suppl 1), 2020.

L. O. Gostin, J. G. Hodge Jr, B. R. Bloom, A. El-Mohandes, J. Fielding, P. Hotez, A. Kurth, H. J. Larson, W. A. Orenstein, K. Rabin, et al. The public health crisis of underimmunisation: a global plan of action. *The Lancet Infectious Diseases*, 20 (1):e11–e16, 2020.

G. B. Grant, B. G. Masresha, W. J. Moss, M. N. Mulders, P. A. Rota, S. B. Omer, A. Shefer, J. L. Kriss, M. Hanson, D. N. Durrheim, et al. Accelerating measles and rubella elimination through research and innovation–findings from the measles & rubella initiative research prioritization process, 2016. *Vaccine*, 37(38):5754–5761, 2019.

E. J. Green, A. O. Finley, and W. E. Strawderman. *Introduction to Bayesian Methods in Ecology and Natural Resources*. Springer Nature, 2020.

R. P. Haining and G. Li. *Regression Modelling With Spatial and Spatial-Temporal Data: A Bayesian Approach.* CRC Press, 2020.

M. L. Hammond, C. Beaulieu, S. A. Henson, and S. K. Sahu. Regional surface chlorophyll trends and uncertainties in the global ocean. *Scientific reports*, 10(1): 1–9, 2020.

I. Harbers. Spatial effects and party nationalization: The geography of partisan support in mexico. *Electoral Studies*, 47:55–66, 2017.

W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

M. J. Heaton, A. Datta, A. O. Finley, R. Furrer, J. Guinness, R. Guhaniyogi, F. Gerber, R. B. Gramacy, D. Hammerling, M. Katzfuss, et al. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24:398–425, 2019.

L. Held, B. Schrödle, and H. Rue. Posterior and cross-validatory predictive checks: a comparison of mcmc and inla. *Statistical modelling and regression structures: Festschrift in honour of ludwig fahrmeir*, pages 91–110, 2010.

J. A. Hoeting, R. A. Davis, A. A. Merton, and S. E. Thompson. Model selection for geostatistical models. *Ecological Applications*, 16(1):87–98, 2006.

M. D. Hoffman, A. Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

M. B. Hooten and N. T. Hobbs. A guide to bayesian model selection for ecologists. *Ecological monographs*, 85(1):3–28, 2015.

B. S. Ibrahim, R. Usman, Z. D. Yahaya Mohammed, O. Okunromade, A. A. Abubakar, and P. M. Nguku. Burden of measles in nigeria: a five-year review of casebased surveillance data, 2012-2016. *The Pan African Medical Journal*, 32 (Suppl 1), 2019.

E. L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

H. Jeffreys. Theory of probability (3rd edt.) oxford university press. *MR0187257*, 432, 1961.

J. B. Johnson and K. S. Omland. Model selection in ecology and evolution. *Trends in ecology & evolution*, 19(2):101–108, 2004.

S. K. Kabra and R. Lodha. Antibiotics for preventing complications in children with measles. *Cochrane database of systematic reviews*, 8, 2013.

J. B. Kadane and N. A. Lazar. Methods and criteria for model selection. *Journal of the American statistical Association*, 99(465):279–290, 2004.

E. W. Kagucia et al. *Health Interventions to Improve Measles Vaccination Coverage and Timeliness: An Assessment of the Immediate and Long-term Impact on Vaccine-Seeking in Rural Kenya*. PhD thesis, Johns Hopkins University, 2018.

R. E. Kass and A. E. Raftery. Bayes factors. *Journal of the american statistical association*, 90(430):773–795, 1995.

C. G. Kaufman, M. J. Schervish, and D. W. Nychka. Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555, 2008.

E. T. Krainski, V. Gómez-Rubio, H. Bakka, A. Lenzi, D. Castro-Camilo, D. Simpson, F. Lindgren, and H. Rue. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. CRC press, 2018.

S. Kullback and R. A. Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

N. Kutner and C. Nachtsheim. *Applied Linear Regression Models*. McGraw-Hill Irwin, 2004.

B. Lambert. *A student's guide to Bayesian statistics*. Sage, 2018.

T. Laurent and P. Margaretic. Predictions in spatial econometric models: Application to unemployment data. In *Advances in Contemporary Statistics and Econometrics*, pages 409–426. Springer, 2021.

K. Le Rest, D. Pinaud, P. Monestiez, J. Chadoeuf, and V. Bretagnolle. Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global ecology and biogeography*, 23(7):811–820, 2014.

D. Lee, C. Robertson, C. Ramsay, and K. Pyper. Quantifying the impact of the modifiable areal unit problem when estimating the health effects of air pollution. *Environmetrics*, 31(8):e2643, 2020.

M. D. Lee and E.-J. Wagenmakers. *Bayesian cognitive modeling: A practical course.* Cambridge university press, 2014.

S. M. Lewis and A. E. Raftery. Estimating bayes factors via posterior simulation with the laplace—metropolis estimator. *Journal of the American Statistical Association*, 92(438):648–655, 1997.

J. Liang. Mapping large-scale forest dynamics: a geospatial approach. *Landscape ecology*, 27:1091–1108, 2012.

F. Lindgren and H. Rue. Bayesian spatial modelling with r-inla. *Journal of statistical software*, 63(19), 2015.

F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498, 2011.

D. J. Lunn, A. Thomas, N. Best, and D. Spiegelhalter. Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10: 325–337, 2000.

C. Mallows. Choosing variables in a linear regression: A graphical aid. In *Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, KS, 1964*, 1964.

C. L. Mallows. Some comments on cp. *Technometrics*, 42(1):87–94, 2000.

J.-M. Marin and C. P. Robert. *Bayesian essentials with R*, volume 48. Springer, 2014.

S. Martino and A. Riebler. Integrated nested laplace approximations (inla). *arXiv preprint arXiv:1907.01248*, 2019.

S. Martino and H. Rue. Implementing approximate bayesian inference using integrated nested laplace approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU, Norway*, 2009.

T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with inla: new features. *Computational Statistics & Data Analysis*, 67:68–83, 2013.

A. M. Mathai and S. B. Provost. *Quadratic forms in random variables: theory and applications.* Dekker, 1992.

G. Matheron. Principles of geostatistics. *Economic geology*, 58(8):1246–1266, 1963.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

D. L. Miller, R. Glennie, and A. E. Seaton. Understanding the stochastic partial differential equation approach to smoothing. *Journal of Agricultural, Biological and Environmental Statistics*, 25(1):1–16, 2020.

J. Møller and R. P. Waagepetersen. *Statistical inference and simulation for spatial point processes.* CRC press, 2003.

J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

J. Moody. The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. *Advances in neural information processing systems*, 4, 1991.

P. Moraga. *Geospatial health data: Modeling and visualization with R-INLA and shiny.* CRC Press, 2019.

M. Muscat, H. Bang, J. Wohlfahrt, S. Glismann, K. Mølbak, et al. Measles in europe: an epidemiological assessment. *The Lancet*, 373(9661):383–389, 2009.

National Population Commission Nigeria and ICF. Nigeria demographic and health survey 2018, 2019.

R. Neal. Bayesian learning for neural networks springer. *New York*, 1996.

R. M. Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

J. C. Nmor, H. T. Thanh, and K. Goto. Recurring measles epidemic in vietnam 2005-2009: implication for strengthened control strategies. *International Journal of Biological Sciences*, 7(2):138, 2011.

NOAA. Visible infrared imaging radiometer suite, 2019. URL `https://ngdc.noaa.gov/eog/viirs/index.html`.

A. O'Hagan. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):99–118, 1995.

T. O'Hagan. Bayes factor. *Significance*, 3:184–186, 2006.

W. A. Orenstein, L. Cairns, A. Hinman, B. Nkowane, J.-M. Olivé, and A. L. Reingold. Measles and rubella global strategic plan 2012–2020 midterm review report: Background and summary. *Vaccine*, 36:A35–A42, 2018.

P. U. Ori, A. Adebowale, C. D. Umeokonkwo, U. Osigwe, and M. S. Balogun. Descriptive epidemiology of measles cases in bauchi state, 2013–2018. *BMC Public Health*, 21(1):1–11, 2021.

M. K. Patel, L. Dumolard, Y. Nedelec, S. V. Sodha, C. Steulet, M. Gacic-Dobo, K. Kretsinger, J. McFarland, P. A. Rota, and J. L. Goodson. Progress toward regional measles elimination—worldwide, 2000–2018. *Morbidity and Mortality Weekly Report*, 68(48):1105, 2019.

G. L. Perry, B. P. Miller, and N. J. Enright. A comparison of methods for the statistical analysis of spatial point patterns in plant ecology. *Plant ecology*, 187(1):59–82, 2006.

L. Pettit. The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):175–184, 1990.

C. Pezzulo, N. Tejedor-Garavito, H. M. T. Chan, I. Dreoni, D. Kerr, S. Ghosh, A. Bonnie, M. Bondarenko, M. Salasibew, and A. J. Tatem. A subnational reproductive, maternal, newborn, child, and adolescent health and development atlas of india. *Scientific Data*, 10(1):86, 2023.

W. Pineda-Ríos, R. Giraldo, and E. Porcu. Functional sar models: With application to spatial econometrics. *Spatial statistics*, 29:145–159, 2019.

M. Plummer et al. Jags: A program for analysis of bayesian graphical models using gibbs sampling. In *Proceedings of the 3rd international workshop on distributed statistical computing*, volume 124, pages 1–10. Vienna, Austria, 2003.

J. Pohjankukka, T. Pahikkala, P. Nevalainen, and J. Heikkonen. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019, 2017.

S. Portet. A primer on model selection using the akaike information criterion. *Infectious Disease Modelling*, 5:111–128, 2020.

Z. C. Quiroz, M. O. Prates, D. K. Dey, and H. å. Rue. Fast bayesian inference of block nearest neighbor gaussian models for large data. *Statistics and Computing*, 33(2):54, 2023.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL `https://www.R-project.org/`.

A. E. Raftery. Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika*, 83(2):251–266, 1996.

A. J. Righetto, C. Faes, Y. Vandendijck, and P. J. Ribeiro Jr. On the choice of the mesh for the analysis of geostatistical data using r-inla. *Communications in Statistics-Theory and Methods*, 49(1):203–220, 2020.

J. Rissanen. Stochastic complexity in statistical inquiry. *World scientific series in computer science*, 15:79–93, 1989.

C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 2004.

C. P. Robert, G. Casella, and G. Casella. *Introducing monte carlo methods with r*, volume 18. Springer, 2010.

G. O. Roberts and S. K. Sahu. Updating schemes, correlation structure, blocking and parameterization for the gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):291–317, 1997.

H. Rue and L. Held. *Gaussian Markov random fields: theory and applications.* CRC press, 2005.

H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392, 2009.

H. Rue, F. Lindgren, J. van Niekerk, E. Krainski, and E. Abdul Fattah. R-INLA project: FAQ, 2013. URL `https://www.r-inla.org/faq`.

H. Rue, S. Martino, F. Lindgren, D. Simpson, A. Riebler, E. Krainski, et al. Functions which allow to perform full bayesian analysis of latent gaussian models using integrated nested laplace approximaxion [www document]. *URL¡ http://inla. googlecode. com/hg-history/default/rinla/DESCRIPTION*, 2014.

S. K. Sahu. *Bayesian modeling of spatio-temporal data with R.* Chapman and Hall/CRC, 2022.

S. K. Sahu, A. E. Gelfand, and D. M. Holland. Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1):61–86, 2006.

S. K. Sahu, K. S. Bakar, J. Zhan, J. L. Campbell, and R. D. Yanai. Spatio-temporal bayesian modeling of precipitation using rain gauge data from the hubbard brook experimental forest, new hampshire, usa. In *Joint Statistical Meetings Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association: 77-92.*, pages 77–92, 2020.

J.-E. A. Saleh et al. Trends of measles in nigeria: A systematic review. *Sahel Medical Journal*, 19(1):5, 2016.

P. D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992.

A. M. Schmidt and A. O'Hagan. Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(3):743–758, 2003.

G. Schwarz. The bayesian information criterion. *Ann. Statist*, 6:461–464, 1978.

R. Shibata. Statistical aspects of model selection. In *From data to model*, pages 215–240. Springer, 1989.

F. O. Shorunke, O. Adeola-Musa, A. Usman, C. Ameh, E. Waziri, and S. A. Adebowale. Descriptive epidemiology of measles surveillance data, osun state, nigeria, 2016–2018. *BMC Public Health*, 19(1):1–8, 2019.

D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32, 2017.

D. Spiegelhalter, A. Thomas, N. Best, and W. Gilks. Bugs 0.5: Bayesian inference using gibbs sampling manual (version ii). *MRC Biostatistics Unit, Institute of Public Health, Cambridge, UK*, pages 1–59, 1996.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde. Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical report, Citeseer, 1998.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639, 2002.

Stan Development Team. RStan: the R interface to Stan, 2020. URL `http://mc-stan.org/`. R package version 2.21.2.

Stan Development Team. Stan modeling language users guide and reference manual, 2023. URL `http://mc-stan.org/`. version 2.33.

M. L. Stein, Z. Chi, and L. J. Welty. Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66 (2):275–296, 2004.

D. Stoyan and A. Penttinen. Recent applications of point process methods in forestry statistics. *Statistical science*, pages 61–78, 2000.

N. Sugiura. Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-theory and Methods*, 7(1):13–26, 1978.

Y. Sun, B. Li, and M. G. Genton. Geostatistics for large datasets. In *Advances and challenges in space-time modelling of natural events*, pages 55–77. Springer, 2012.

S. Sundarajan and S. Keerthi. Predictive approaches for choosing hyper-parameters in gaussian process. *Neural Computation*, 13:11031118, 2001.

K. Takeuchi. Distribution of informational statistics and a criterion of model fitting. suri-kagaku (mathematical sciences) 153 12-18, 1976.

A. J. Tatem, P. W. Gething, S. Bhatt, D. Weiss, and C. Pezzulo. Pilot high resolution poverty maps, 2013. University of Southampton / Oxford.

W. R. Tobler. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240, 1970.

C. Utazi, H. Chan, I. Olowe, A. Wigley, N. Tejedor-Garavito, A. Cunningham, M. Bondarenko, J. Lorin, D. Boyda, D. Hogan, et al. A zero-dose vulnerability index for equity assessment and spatial prioritization in low-and middle-income countries. *Spatial Statistics*, page 100772, 2023.

C. E. Utazi, J. Thorley, V. A. Alegana, M. J. Ferrari, S. Takahashi, C. J. E. Metcalf, J. Lessler, and A. J. Tatem. High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine*, 36(12):1583–1591, 2018.

C. E. Utazi, J. Wagai, O. Pannell, F. T. Cutts, D. A. Rhoda, M. J. Ferrari, B. Dieng, J. Oteri, M. C. Danovaro-Holliday, A. Adeniran, et al. Geospatial variation in measles vaccine coverage through routine and campaign strategies in nigeria: Analysis of recent household surveys. *Vaccine*, 38(14):3062–3071, 2020.

C. E. Utazi, J. M. K. Aheto, H. M. T. Chan, A. J. Tatem, and S. K. Sahu. Conditional probability and ratio-based approaches for mapping the coverage of multi-dose vaccines. *Statistics in Medicine*, 41(29):5662–5678, 2022.

M. Vahedi Saheli and M. Effati. Segment-based count regression geospatial modeling of the effect of roadside land uses on pedestrian crash frequency in rural roads. *International journal of intelligent transportation systems research*, 19:347–365, 2021.

A. Vehtari and J. Ojanen. A survey of bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

A. Vehtari, D. Simpson, A. Gelman, Y. Yao, and J. Gabry. Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*, 2015.

A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27:1413–1432, 2017a. doi: 10.1007/s11222-016-9696-4.

A. Vehtari, A. Gelman, and J. Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27(5):1413–1432, 2017b.

A. Vehtari, P. Buerkner, and J. Gabry. Leave-one-out cross-validation for non-factorizable models. Technical report, Technical report. URL http://mc-stan.org/loo/articles/loo2-non-factorizable . . . , 2018.

Z. Wan, S. Hook, and G. Hulley. Mod11c3 modis/terra land surface temperature/emissivity monthly l3 global 0.05deg cmg voo6, 2015. NASA EOSDIS Land Processes DAAC.

C. Wang, M. A. Puhan, R. Furrer, S. S. Group, et al. Generalized spatial fusion model framework for joint analysis of point and areal data. *Spatial Statistics*, 23: 72–90, 2018.

O. Wariri, E. Nkereuwem, N. A. Erondu, B. Edem, O. O. Nkereuwem, O. T. Idoko, E. Agogo, J. E. Enegela, T. Sesay, I. S. Conde, et al. A scorecard of progress towards measles elimination in 15 west african countries, 2001–19: a retrospective, multicountry analysis of national immunisation coverage and surveillance data. *The Lancet Global Health*, 9(3):e280–e290, 2021.

S. Watanabe. Equations of states in singular statistical estimation. *Neural Networks*, 23(1):20–34, 2010.

S. Watanabe and M. Opper. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12), 2010.

D. J. Weiss, A. Nelson, H. Gibson, W. Temperley, S. Peedell, A. Lieber, M. Hancher, E. Poyart, S. Belchior, N. Fullman, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553(7688):333–336, 2018.

D. C. Wheeler, D. A. Hickson, and L. A. Waller. Assessing local model adequacy in bayesian hierarchical models using the partitioned deviance information criterion. *Computational statistics & data analysis*, 54(6):1657–1671, 2010.

P. Whittle. On stationary processes in the plane. *Biometrika*, pages 434–449, 1954.

WHO. The health of the people: the african regional health report, 2006.

WHO. Measles elimination by 2020: a strategy for the african region report of the secretariat executive summary, 2011.

WHO. Measles, 2019. URL `https://www.who.int/news-room/fact-sheets/detail/measles`.

T. Wiegand and K. A. Moloney. *Handbook of spatial point-pattern analysis in ecology.* CRC press, 2013.

G. Xia, M. L. Miranda, and A. E. Gelfand. Approximately optimal spatial design approaches for environmental health data. *Environmetrics: The official journal of the International Environmetrics Society*, 17(4):363–385, 2006.

C. V. y Perdomo. The local context and the spatial diffusion of multiparty competition in urban mexico, 1994–2000. *Political Geography*, 23(4):403–423, 2004.

B. Yu. Minimum description length principle: a review. In *Proceeding of International Symposium of Information Theory and Its Application*, pages 432–435, 1996.

J. Zhang and M. A. Stephens. A new and efficient estimation method for the generalized pareto distribution. *Technometrics*, 51(3):316–325, 2009.

L. Zhang. Nearest neighbor gaussian processes (nngp) based models in stan. *Stan case study https://mc-stan. org/users/documentation/case-studies/nngp. html*, 2018.

L. Zhang, A. Datta, and S. Banerjee. Practical bayesian modeling and inference for massive spatial data sets on modest computing environments. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):197–209, 2019.

L. Zhu and B. P. Carlin. Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. *Statistics in Medicine*, 19 (17-18):2265–2278, 2000.

Appendix A

Deriving covariance functions from semivariograms

Assuming an isotropic and stationary spatial process, the relationship between the semivariogram $\gamma(\mathbf{h})$ and the covariance function $C(\mathbf{h})$ is

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}).$$

In the derivations below, $\mathbf{h}$ and $||\mathbf{h}||$ are used interchangeably, since only the length of the separation vector is concerned under the isotropic assumption. Now, further assuming that the spatial process is ergodic, $C(\mathbf{h}) \to 0$ as $||\mathbf{h}|| \to \infty$. Taking the limit from this expression gives

$$\begin{aligned}\lim_{||\mathbf{h}||\to\infty} \gamma(\mathbf{h}) &= C(\mathbf{0}) - \lim_{||\mathbf{h}||\to\infty} C(\mathbf{h}),\\ &= C(\mathbf{0}) - 0,\\ &= C(\mathbf{0}).\end{aligned}$$

Rewriting the relationship in terms of the semivariogram,

$$\begin{aligned}\gamma(\mathbf{h}) &= C(\mathbf{0}) - C(\mathbf{h}),\\ C(\mathbf{h}) &= C(\mathbf{0}) - \gamma(\mathbf{h}),\\ &= \lim_{||\mathbf{u}||\to\infty} \gamma(\mathbf{u}) - \gamma(\mathbf{h}),\end{aligned}$$

where $\mathbf{u}$ denotes $\mathbf{h}$ to avoid confusion with the other components in the expression.

As an example, consider the exponential semivariogram,

$$\gamma(\mathbf{h}) = \begin{cases} \tau^2 + \sigma^2\big(1 - \exp(-\phi||\mathbf{h}||)\big) & \text{if } ||\mathbf{h}|| > 0, \\ 0 & \text{if } ||\mathbf{h}|| = 0. \end{cases}$$

To simplify the notation, let d denote $\mathbf{h}$. Since the spatial process is assumed to be isotropic, d also represents $||\mathbf{h}||$. To derive the covariance function, take into account the two separate cases: when $d > 0$ and when $d = 0$. First, consider the case when $d > 0$. Following the relationship given above,

$$\begin{aligned} C(d) &= \lim_{||\mathbf{u}||\to\infty} \gamma(\mathbf{u}) - \gamma(d), \\ &= \lim_{||\mathbf{u}||\to\infty} \tau^2 + \sigma^2\big(1 - \exp(-\phi||\mathbf{u}||)\big) - \gamma(d), \\ &= \tau^2 + \sigma^2(1-0) - \gamma(d), \\ &= \tau^2 + \sigma^2 - \gamma(d), \\ &= \tau^2 + \sigma^2 - [\tau^2 + \sigma^2(1 - \exp(-\phi d))], \\ &= \tau^2 + \sigma^2 - \tau^2 - \sigma^2 + \sigma^2\exp(-\phi d), \\ &= \sigma^2\exp(-\phi d). \end{aligned}$$

The notation $\mathbf{u}$ was used to indicate that the limit should only be taken for the first semivariogram on the right-hand side of the first line of the derivation. Within the second and third lines of the derivation, the exponential of a large negative number approaches zero. Hence, $C(d) = \sigma^2\exp(-\phi d)$ when $d > 0$. Now consider the other case when $d = 0$,

$$\begin{aligned} C(d) &= \lim_{||\mathbf{u}||\to\infty} \gamma(\mathbf{u}) - \gamma(d), \\ &= \lim_{||\mathbf{u}||\to\infty} \tau^2 + \sigma^2\big(1 - \exp(-\phi||\mathbf{u}||)\big) - \gamma(d), \\ &= \tau^2 + \sigma^2(1-0) - \gamma(d), \\ &= \tau^2 + \sigma^2 - \gamma(d), \\ &= \tau^2 + \sigma^2 - 0, \\ &= \tau^2 + \sigma^2. \end{aligned}$$

Summarising the derivations, the covariance function of the exponential semivariogram is

$$C(d) = \begin{cases} \sigma^2 \exp(-\phi d) & \text{if } d > 0, \\ \sigma^2 + \tau^2 & \text{if } d = 0. \end{cases}$$

Consider the Matérn semivariogram as another example. The Matérn semivariogram is given as follows,

$$\gamma(d) = \begin{cases} \tau^2 + \sigma^2\big(1 - \frac{(\sqrt{2\nu}\phi d)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu}\phi d)\big) & \text{if } d > 0, \\ 0 & \text{if } d = 0. \end{cases}$$

First, consider when $d > 0$. Once again, the covariance function is given as

$$\begin{aligned} C(d) &= \lim_{||\mathbf{u}|| \to \infty} \gamma(\mathbf{u}) - \gamma(d), \\ &= \lim_{||\mathbf{u}|| \to \infty} \tau^2 + \sigma^2 \left(1 - \frac{(\sqrt{2\nu}\phi||\mathbf{u}||)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu}\phi||\mathbf{u}||)\right) - \gamma(d), \\ &= \tau^2 + \sigma^2(1 - 0) - \gamma(d), \\ &= \tau^2 + \sigma^2 - \gamma(d), \\ &= \tau^2 + \sigma^2 - \tau^2 - \sigma^2 \left(1 - \frac{(\sqrt{2\nu}\phi d)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu}\phi d)\right), \\ &= \tau^2 + \sigma^2 - \tau^2 - \sigma^2 + \sigma^2 \frac{(\sqrt{2\nu}\phi d)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu}\phi d), \\ &= \sigma^2 \frac{(\sqrt{2\nu}\phi d)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu}\phi d). \end{aligned}$$

As previously mentioned, the notation $\mathbf{u}$ is used to indicate that the limit should only be taken for first semivariogram on the right-hand side of the first line of the derivation. Within the second and third lines of the derivation, since $K_\nu(\cdot)$ denotes the modified Bessel function of the second kind of order ν, when the terms inside the function becomes large, the function tends towards zero. Now, consider the other case when $d = 0$,

$$\begin{aligned}
C(d) &= \lim_{||\mathbf{u}|| \to \infty} \gamma(\mathbf{u}) - \gamma(d), \\
&= \lim_{||\mathbf{u}|| \to \infty} \tau^2 + \sigma^2 \left[1 - \frac{(\sqrt{2\nu}\phi||\mathbf{u}||)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu}\phi||\mathbf{u}||) \right] - \gamma(d), \\
&= \tau^2 + \sigma^2(1 - 0) - \gamma(d), \\
&= \tau^2 + \sigma^2 - \gamma(d), \\
&= \tau^2 + \sigma^2 - 0, \\
&= \tau^2 + \sigma^2.
\end{aligned}$$

Summarising the derivations above, the covariance function for the Matérn semivariogram is

$$C(d) = \begin{cases} \sigma^2 \frac{(\sqrt{2\nu}\phi d)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(\sqrt{2\nu}\phi d) & \text{if } d > 0, \\ \tau^2 + \sigma^2 & \text{if } d = 0. \end{cases}$$

Appendix B

Posterior distribution for a multivariate normal model

Consider the following model,

$$\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 H),$$

where the $n \times 1$ column vector of responses, denoted as $\mathbf{Y}$, follows an n-dimensional multivariate normal distribution $N_n(\cdot)$ with mean structure $X\boldsymbol{\beta}$ and covariance matrix $\sigma^2 H$. The mean structure comprise an $n \times p$ design matrix, denoted as X, and a $p \times 1$ column vector of regression coefficients, denoted as $\boldsymbol{\beta}$. The covariance matrix is given by some known value σ^2 and an $n \times n$ matrix H, which is assumed to be the identity matrix. The prior distribution of $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, \sigma^2 M^{-1}),$$

is assumed to follow a p-dimensional multivariate normal distribution $N_p(\cdot)$ with mean structure $\boldsymbol{\beta}_0$ and covariance matrix $\sigma^2 M^{-1}$. The mean structure $\boldsymbol{\beta}_0$ is a $p \times 1$ column vector of known values. The covariance matrix is given by some known value σ^2 and a $p \times p$ matrix M, which is assumed to be the identity matrix. Since M is assumed to be the identity matrix, it implies that $M = M^{-1} = I_p$, where I_p denotes a $p \times p$ identity matrix.

To assist the following derivations, we define $\lambda^2 = 1/\sigma^2$. The probability density functions (PDFs) are given as follows,

$$
\begin{aligned}
p(\mathbf{y}|\boldsymbol{\beta},\lambda^2) &= \left(\frac{\lambda^2}{2\pi}\right)^{\frac{n}{2}} \det(H)^{-\frac{1}{2}} \exp\left(-\frac{\lambda^2}{2}(\mathbf{y}-X\boldsymbol{\beta})'H^{-1}(\mathbf{y}-X\boldsymbol{\beta})\right),\\
p(\boldsymbol{\beta}|\lambda^2) &= \left(\frac{\lambda^2}{2\pi}\right)^{\frac{p}{2}} \det(M)^{\frac{1}{2}} \exp\left(-\frac{\lambda^2}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'M(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right),
\end{aligned}
$$

where det(·) denotes the determinant. The posterior distribution can be calculated as posterior $\propto$ likelihood × prior. Using the PDFs given above, the full conditional posterior distribution, denoted as $p(\boldsymbol{\beta}|\mathbf{y},\lambda^2)$, is calculated as follows,

$$
\begin{aligned}
p(\boldsymbol{\beta}|\mathbf{y},\lambda^2) &\propto p(\mathbf{y}|\boldsymbol{\beta},\lambda^2) \times p(\boldsymbol{\beta}|\lambda^2),\\
&\propto \exp\left(-\frac{\lambda^2}{2}\Big[(\mathbf{y}-X\boldsymbol{\beta})'H^{-1}(\mathbf{y}-X\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'M(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\Big]\right).
\end{aligned}
$$

Notice that the above only comprise the components pertaining to $\boldsymbol{\beta}$. The other components from $p(\mathbf{y}|\boldsymbol{\beta},\lambda^2)$ and $p(\boldsymbol{\beta}|\lambda^2)$ are absorbed to the proportionality constant. Focusing on the components within the square brackets, the terms can be expanded and rearranged as follows

$$
\begin{aligned}
&(\mathbf{y}-X\boldsymbol{\beta})'H^{-1}(\mathbf{y}-X\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'M(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\
&= \mathbf{y}'H^{-1}\mathbf{y} - \mathbf{y}'H^{-1}(X\boldsymbol{\beta}) - (X\boldsymbol{\beta})'H^{-1}\mathbf{y} + (X\boldsymbol{\beta})'H^{-1}(X\boldsymbol{\beta}) + \boldsymbol{\beta}'M\boldsymbol{\beta} - \boldsymbol{\beta}'M\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0'M\boldsymbol{\beta} + \boldsymbol{\beta}_0'M\boldsymbol{\beta}_0,\\
&= \mathbf{y}'H^{-1}\mathbf{y} - \mathbf{y}'H^{-1}(X\boldsymbol{\beta}) - \boldsymbol{\beta}'X'H^{-1}\mathbf{y} + \boldsymbol{\beta}'X'H^{-1}(X\boldsymbol{\beta}) + \boldsymbol{\beta}'M\boldsymbol{\beta} - \boldsymbol{\beta}'M\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0'M\boldsymbol{\beta} + \boldsymbol{\beta}_0'M\boldsymbol{\beta}_0,\\
&= \boldsymbol{\beta}'(X'H^{-1}X + M)\boldsymbol{\beta} - \boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0) - (\mathbf{y}'H^{-1}X + \boldsymbol{\beta}_0'M)\boldsymbol{\beta} + \mathbf{y}'H^{-1}\mathbf{y} + \boldsymbol{\beta}_0'M\boldsymbol{\beta}_0.
\end{aligned}
$$

Recall that since H and M are both assumed to be identity matrix, $H = H^{-1}$ and $M = M^{-1}$. Furthermore, $H = H'$ and $M = M'$. Therefore, the derivation can be further simplified,

$$
\begin{aligned}
&(\mathbf{y}-X\boldsymbol{\beta})'H^{-1}(\mathbf{y}-X\boldsymbol{\beta}) + (\boldsymbol{\beta}-\boldsymbol{\beta}_0)'M(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\\
&= \boldsymbol{\beta}'(X'H^{-1}X + M)\boldsymbol{\beta} - \boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0) - (\mathbf{y}'H^{-1}X + \boldsymbol{\beta}_0'M)\boldsymbol{\beta} + \mathbf{y}'H^{-1}\mathbf{y} + \boldsymbol{\beta}_0'M\boldsymbol{\beta}_0,\\
&= \boldsymbol{\beta}'(X'H^{-1}X + M)\boldsymbol{\beta} - \boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0) - \boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0) + \mathbf{y}'H^{-1}\mathbf{y} + \boldsymbol{\beta}_0'M\boldsymbol{\beta}_0,\\
&= \boldsymbol{\beta}'(X'H^{-1}X + M)\boldsymbol{\beta} - 2\boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0) + \mathbf{y}'H^{-1}\mathbf{y} + \boldsymbol{\beta}_0'M\boldsymbol{\beta}_0.
\end{aligned}
$$

Since the components $\mathbf{y}'H^{-1}\mathbf{y}+\boldsymbol{\beta}_0'M\boldsymbol{\beta}_0$ above do not involve $\boldsymbol{\beta}$, they can be omitted from further deviations, and will be accounted for in the proportionality constant. For notation simplification, we define $M^* = X'H^{-1}X + M$. Returning to the posterior distribution, the current progress gives

$$p(\boldsymbol{\beta}|\mathbf{y},\lambda^2) \propto \exp\bigg(-\frac{\lambda^2}{2}\Big[\boldsymbol{\beta}'M^*\boldsymbol{\beta} - 2\boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)\Big]\bigg).$$

The simplification is incomplete. Refocusing on the terms within the square brackets, further simplification requires completing the square,

$$\begin{aligned}\boldsymbol{\beta}'M^*\boldsymbol{\beta} - 2\boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0) &= \boldsymbol{\beta}'M^*\boldsymbol{\beta} - 2\boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)\\ &\quad + (X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)'(M^*)^{-2}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)\\ &\quad - (X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)'(M^*)^{-2}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0).\end{aligned}$$

Notice that the last term $-(X'H^{-1}\mathbf{y}+M\boldsymbol{\beta}_0)'(M^*)^{-2}(X'H^{-1}\mathbf{y}+M\boldsymbol{\beta}_0)$ does not involve $\boldsymbol{\beta}$. This term can be omitted and will be absorbed into the proportionality constant. Using the remaining components, the following simplification can take place,

$$\begin{aligned}&\boldsymbol{\beta}'M^*\boldsymbol{\beta} - 2\boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0) + (X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)'(M^*)^{-2}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)\\ &= \big(\boldsymbol{\beta} - (M^*)^{-1}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)\big)'M^*\big(\boldsymbol{\beta} - (M^*)^{-1}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)\big).\end{aligned}$$

Define $\boldsymbol{\beta}^* = (M^*)^{-1}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)$.

$$\begin{aligned}&\boldsymbol{\beta}'M^*\boldsymbol{\beta} - 2\boldsymbol{\beta}'(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0) + (X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)'(M^*)^{-2}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)\\ &= \big(\boldsymbol{\beta} - (M^*)^{-1}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)\big)'M^*\big(\boldsymbol{\beta} - (M^*)^{-1}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0)\big),\\ &= (\boldsymbol{\beta} - \boldsymbol{\beta}^*)'M^*(\boldsymbol{\beta} - \boldsymbol{\beta}^*).\end{aligned}$$

Finally, returning to the full conditional posterior distribution,

$$p(\boldsymbol{\beta}|\mathbf{y},\lambda^2) \propto \exp\bigg(-\frac{\lambda^2}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)'M^*(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\bigg).$$

From the expression above, the posterior distribution is identified to be a p-dimensional multivariate normal distribution with mean structure $\boldsymbol{\beta}^*$ and covariance matrix $(\lambda^2 M^*)^{-1}$.

More formally,

$$\boldsymbol{\beta}|\mathbf{y}, \lambda^2 \sim N_p\biggl(\boldsymbol{\beta}^*, \frac{1}{\lambda^2}(M^*)^{-1}\biggr),$$

where

$$\begin{aligned}\boldsymbol{\beta}^* &= (M^*)^{-1}(X'H^{-1}\mathbf{y} + M\boldsymbol{\beta}_0),\\ M^* &= X'H^{-1}X + M.\end{aligned}$$

This implies that the posterior mean $E_{\boldsymbol{\beta}|\mathbf{y}}(\cdot)$ and posterior variance $\mathrm{Var}_{\boldsymbol{\beta}|\mathbf{y}}(\cdot)$ are

$$\begin{aligned}E(\boldsymbol{\beta}|\mathbf{y}, \lambda^2) &= \boldsymbol{\beta}^*,\\ \mathrm{Var}(\boldsymbol{\beta}|\mathbf{y}, \lambda^2) &= \frac{1}{\lambda^2}(M^*)^{-1}.\end{aligned}$$

Appendix C

Posterior predictive distribution for a multivariate normal model

Consider the following model,

$$\mathbf{Y} \sim N_n(X\boldsymbol{\beta}, \sigma^2 H),$$

where the $n \times 1$ column vector of responses, denoted as $\mathbf{Y}$, follows an n-dimensional multivariate normal distribution $N_n(\cdot)$ with mean structure $X\boldsymbol{\beta}$ and covariance matrix $\sigma^2 H$. The mean structure comprise an $n \times p$ design matrix, denoted as X, and a $p \times 1$ column vector of regression coefficients, denoted as $\boldsymbol{\beta}$. The covariance matrix is given by some known value σ^2 and an $n \times n$ matrix H, which is assumed to be the identity matrix. The prior distribution of $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} \sim N_p(\boldsymbol{\beta}_0, \sigma^2 M^{-1}),$$

is assumed to follow a p-dimensional multivariate normal distribution $N_p(\cdot)$ with mean structure $\boldsymbol{\beta}_0$ and covariance matrix $\sigma^2 M^{-1}$. The mean structure $\boldsymbol{\beta}_0$ is a $p \times 1$ column vector of known values. The covariance matrix is given by some known value σ^2 and a $p \times p$ matrix M, which is assumed to be the identity matrix. Since M is assumed to be the identity matrix, that implies $M = M^{-1} = I_p$, where I_p denotes a $p \times p$ identity matrix. To assist the following derivations, we define $\lambda^2 = 1/\sigma^2$.

The derived full conditional posterior distribution is

$$p(\boldsymbol{\beta}|\mathbf{y},\lambda^2) \propto \exp\left(-\frac{\lambda^2}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^*)'M^*(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\right),$$

where

$$\begin{aligned}\boldsymbol{\beta}^* &= (M^*)^{-1}(X'H^{-1}\mathbf{y}+M\boldsymbol{\beta}_0),\\ M^* &= X'H^{-1}X+M,\end{aligned}$$

which implies that the posterior mean $E_{\boldsymbol{\beta}|\mathbf{y}}(\cdot)$ and posterior variance $\mathrm{Var}_{\boldsymbol{\beta}|\mathbf{y}}(\cdot)$ are

$$\begin{aligned}E(\boldsymbol{\beta}|\mathbf{y},\lambda^2) &= \boldsymbol{\beta}^*,\\ \mathrm{Var}(\boldsymbol{\beta}|\mathbf{y},\lambda^2) &= \frac{1}{\lambda^2}(M^*)^{-1}.\end{aligned}$$

Now, suppose there is a new observation, denoted as Y_0, that follows a normal distribution with mean $\mathbf{x}_0'\boldsymbol{\beta}$ and variance σ^2, where $\mathbf{x}_0$ is a p-dimensional column vector of covariates, $\boldsymbol{\beta}$ are the regression coefficients and σ^2 denotes a known value,

$$Y_0 \sim N(\mathbf{x}_0'\boldsymbol{\beta},\sigma^2).$$

To derive the posterior predictive distribution, first start with the joint distribution of Y_0 and $\mathbf{Y}$, which is given as follows,

$$\begin{pmatrix} Y_0\\ \mathbf{Y}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{x}_0'\boldsymbol{\beta}\\ X\boldsymbol{\beta}\end{pmatrix}, \sigma^2\begin{pmatrix}1 & \Sigma_{12}\\ \Sigma_{21} & H\end{pmatrix}\right),$$

where $\Sigma_{12}=\Sigma_{21}'$, and Σ_{12} is a $1\times n$ vector with elements $\mathrm{Cor}(Y_0,Y_i)$, for $i=1,\ldots,n$. The notation $\mathrm{Cor}(\cdot)$ denotes the correlation function. From multivariate normal distribution theory, the conditional distribution for $Y_1|Y_2$ is given as a normal distribution with mean $\bar{\boldsymbol{\mu}}$ and covariance $\bar{\boldsymbol{\Sigma}}$, where

$$\begin{aligned}\bar{\boldsymbol{\mu}} &= \boldsymbol{\mu}_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}-\boldsymbol{\mu}_2),\\ \bar{\boldsymbol{\Sigma}} &= \boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.\end{aligned}$$

In our context, the conditional distribution is given as

$$Y_0|\mathbf{y}, \boldsymbol{\beta}, \lambda^2 \sim N\Bigg(\mathbf{x}_0'\boldsymbol{\beta} + \boldsymbol{\Sigma}_{12}H^{-1}(\mathbf{y} - X\boldsymbol{\beta}), \frac{1}{\lambda^2}(1 - \boldsymbol{\Sigma}_{12}H^{-1}\boldsymbol{\Sigma}_{21})\Bigg).$$

To simplify the notation above, we define

$$\begin{aligned} \delta^2 &= (1 - \boldsymbol{\Sigma}_{12}H^{-1}\boldsymbol{\Sigma}_{21}), \\ \mu_0 &= \mathbf{x}_0'\boldsymbol{\beta} + \boldsymbol{\Sigma}_{12}H^{-1}(\mathbf{y} - X\boldsymbol{\beta}), \end{aligned}$$

to reduce the expression as,

$$Y_0|\mathbf{y}, \boldsymbol{\beta}, \lambda^2 \sim N\Bigg(\mu_0, \frac{\delta^2}{\lambda^2}\Bigg).$$

Now,

$$\begin{aligned} y_0 - \mu_0 &= y_0 - \big(\mathbf{x}_0'\boldsymbol{\beta} + \boldsymbol{\Sigma}_{12}H^{-1}(\mathbf{y} - X\boldsymbol{\beta})\big), \\ &= y_0 - \mathbf{x}_0'\boldsymbol{\beta} - \big(\boldsymbol{\Sigma}_{12}H^{-1}(\mathbf{y} - X\boldsymbol{\beta})\big), \\ &= y_0 - \mathbf{x}_0'\boldsymbol{\beta} - \boldsymbol{\Sigma}_{12}H^{-1}\mathbf{y} + \boldsymbol{\Sigma}_{12}H^{-1}X\boldsymbol{\beta}, \\ &= y_0 - \boldsymbol{\Sigma}_{12}H^{-1}\mathbf{y} - (\mathbf{x}_0' - \boldsymbol{\Sigma}_{12}H^{-1}X)\boldsymbol{\beta}, \\ &= \tilde{y}_0 - \mathbf{g}'\boldsymbol{\beta}, \end{aligned}$$

where

$$\begin{aligned} \tilde{y}_0 &= y_0 - \boldsymbol{\Sigma}_{12}H^{-1}\mathbf{y}, \\ \mathbf{g}' &= (\mathbf{x}_0' - \boldsymbol{\Sigma}_{12}H^{-1}X). \end{aligned}$$

Using this information, the conditional distribution can be expressed as,

$$p(\tilde{Y}_0|\mathbf{y}, \boldsymbol{\beta}, \lambda^2) \propto \exp\left(\frac{\lambda^2}{2\delta^2}(\tilde{y}_0 - \mathbf{g}'\boldsymbol{\beta})^2\right),$$

which implies

$$\tilde{Y}_0|\mathbf{y}, \boldsymbol{\beta}, \lambda^2 \sim N\Bigg(\mathbf{g}'\boldsymbol{\beta}, \frac{\delta^2}{\lambda^2}\Bigg).$$

Deriving $\tilde{Y}_0|\mathbf{y}, \lambda^2$ from $\tilde{Y}_0|\mathbf{y}, \boldsymbol{\beta}, \lambda^2$ requires an integration with respect to $\boldsymbol{\beta}$. A shortcut is to use the derivation results of the posterior mean and variance. That is,

$$\begin{aligned}
E(\tilde{Y}_0|\mathbf{y}, \lambda^2) &= E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(E(\tilde{Y}_0|\mathbf{y}, \boldsymbol{\beta}, \lambda^2)\big), \\
&= E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(\mathbf{g}'\boldsymbol{\beta}\big), \\
&= \mathbf{g}'\boldsymbol{\beta}^*,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(\tilde{Y}_0|\mathbf{y}, \lambda^2) &= E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(\mathrm{Var}(\tilde{Y}_0|\mathbf{y}, \boldsymbol{\beta}, \lambda^2)\big) + \mathrm{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\big(E(\tilde{Y}_0|\mathbf{y}, \boldsymbol{\beta}, \lambda^2)\big), \\
&= E_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}\left(\frac{\delta^2}{\lambda^2}\right) + \mathrm{Var}_{\boldsymbol{\beta}|\mathbf{y},\lambda^2}(\mathbf{g}'\boldsymbol{\beta}), \\
&= \frac{\delta^2}{\lambda^2} + \frac{1}{\lambda^2}\mathbf{g}'(M^*)^{-1}\mathbf{g}, \\
&= \frac{1}{\lambda^2}\left(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g}\right).
\end{aligned}$$

In other words, the conditional posterior predictive distribution is

$$\tilde{Y}_0|\mathbf{y}, \lambda^2 \sim N\left(\mathbf{g}'\boldsymbol{\beta}^*, \frac{1}{\lambda^2}\big(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g}\big)\right),$$

with

$$\begin{aligned}
E(\tilde{Y}_0|\mathbf{y}, \lambda^2) &= \mathbf{g}'\boldsymbol{\beta}^*, \\
\mathrm{Var}(\tilde{Y}_0|\mathbf{y}, \lambda^2) &= \frac{1}{\lambda^2}\big(\delta^2 + \mathbf{g}'(M^*)^{-1}\mathbf{g}\big),
\end{aligned}$$

where

$$\begin{aligned}
\mathbf{g}' &= (\mathbf{x}_0' - \boldsymbol{\Sigma}_{12}H^{-1}X), \\
\delta^2 &= (1 - \boldsymbol{\Sigma}_{12}H^{-1}\boldsymbol{\Sigma}_{21}).
\end{aligned}$$

Appendix D

Gram-Schmidt orthogonalisation process

The Gram-Schmidt orthogonalisation process was applied to the generated covariates, in Chapter 4, to ensure independence. The covariate are generated from the following,

$$\begin{aligned} x_1(\mathbf{s}_i) &= 1, \\ x_2(\mathbf{s}_i) &\sim N(0,1), \\ x_3(\mathbf{s}_i) &\sim N(0,1), \\ x_4(\mathbf{s}_i) &\sim N(0,2), \end{aligned}$$

where $N(\cdot)$ denotes the Normal distribution. First, we define $\mathbf{x}_p = \big(x_p(\mathbf{s}_1), \ldots, x_p(\mathbf{s}_n)\big)'$, where $p = 1,2,3,4$. Then, we calculate $\mathbf{v}_p$ as follows,

$$\begin{aligned} \mathbf{v}_1 &= \mathbf{x}_1, \\ \mathbf{v}_2 &= \mathbf{x}_2 - \frac{\langle \mathbf{x}_2, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1, \\ \mathbf{v}_3 &= \mathbf{x}_3 - \frac{\langle \mathbf{x}_3, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1 - \frac{\langle \mathbf{x}_3, \mathbf{v}_2 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle}\mathbf{v}_2, \\ \mathbf{v}_4 &= \mathbf{x}_4 - \frac{\langle \mathbf{x}_4, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle}\mathbf{v}_1 - \frac{\langle \mathbf{x}_4, \mathbf{v}_2 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle}\mathbf{v}_2 - \frac{\langle \mathbf{x}_4, \mathbf{v}_3 \rangle}{\langle \mathbf{v}_3, \mathbf{v}_3 \rangle}\mathbf{v}_3, \end{aligned}$$

where $\langle \cdot \rangle$ denotes the dot product between the two vectors within the braces. For example, if $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_n)$ are row vectors, then $\langle \mathbf{a}, \mathbf{b} \rangle = a_1b_1 + a_2b_2 + \cdots + a_nb_n$, or in the case where $\mathbf{a}$ and $\mathbf{b}$ are column vectors $\langle \mathbf{a}, \mathbf{b} \rangle = \mathbf{a}'\mathbf{b}$.

Following this, we normalise $\mathbf{v}_p$ to get $\mathbf{w}_p$ as follows,

$$\mathbf{w}_p = \frac{\mathbf{v}_p}{\sqrt{\langle \mathbf{v}_p, \mathbf{v}_p \rangle}},$$

where $p = 1, 2, 3, 4$. We then use $\mathbf{w}_p$ as our covariates in place of $\mathbf{x}_p$ in the simulation examples, as we have now ensured independence among the covariates with $\mathbf{w}_p$.

Appendix E

Additional simulation designs 1

Following the simulation designs detailed in Chapter 4, we conducted additional experiments with point referenced spatial data generated from other structures. Specifically, they are the lattice design and the augmented lattice design. Examples of point referenced spatial data generated from these designs are shown in Figures E.1 and E.4, respectively.

In the lattice design, the point referenced data are equally spaced within a unit square. For this design, we generated data with sample size n that has an integer square-root. Figure E.1a shows a dataset generated using $n = 36$, so the points are equally spaced within the unit square in a 6×6 manner. In the augmented lattice design, the point referenced data are also equally spaced within a unit square. For this design, we chose three additional points and generated 3×3 lattices. Figure E.4a shows a dataset generated using $n = 25$, so the points are equally spaced within the unit square in a 5×5 manner. We then arbitrarily chose three points and created 3×3 lattices, including the selected points. This gives a total of $n = 49$ in this generated dataset.

Although the locations of the data are simulated differently, the responses (Figures E.1b and E.4b) are generated following the procedure detailed in Chapter 4, that is

$$\begin{aligned} \mathbf{Y} &\sim N_n(X\boldsymbol{\beta} + \boldsymbol{\omega}, \tau^2 I_n), \\ \boldsymbol{\omega} &\sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}), \end{aligned}$$

where $N_n(\cdot)$ denotes an n-dimensional multivariate normal distribution, I_n denotes an $n \times n$ identity matrix, $\boldsymbol{\beta}$ denotes a vector of regression coefficients, X denotes an $n \times p$

design matrix that contains the covariates and $\mathbf{\Sigma}_{\boldsymbol{\omega}}$ denotes an $n \times n$ covariance matrix, with elements calculated from some covariance functions, including those described in Sections 2.3.3 and 4.1.1. Furthermore, we considered the following combinations of covariates

$$
\begin{aligned}
\text{f1}: &\quad \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} = x_1(\mathbf{s}_i)\beta_1 + x_2(\mathbf{s}_i)\beta_2, \\
\text{f2}: &\quad \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} = x_1(\mathbf{s}_i)\beta_1 + x_2(\mathbf{s}_i)\beta_2 + x_3(\mathbf{s}_i)\beta_3, \\
\text{f3}: &\quad \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} = x_1(\mathbf{s}_i)\beta_1 + x_2(\mathbf{s}_i)\beta_2 + x_3(\mathbf{s}_i)\beta_3 + x_4(\mathbf{s}_i)\beta_4,
\end{aligned}
$$

where $\beta_1 = 1,\ \beta_2 = 2,\ \beta_3 = 2,\ \beta_4 = 2$ and

$$
\begin{aligned}
x_1(\mathbf{s}_i) &= 1, \\
x_2(\mathbf{s}_i) &\sim N(0,1), \\
x_3(\mathbf{s}_i) &\sim N(0,1), \\
x_4(\mathbf{s}_i) &\sim N(0,2).
\end{aligned}
$$

When fitting the model using the Stan framework and the INLA method, we specified the following prior distributions for the parameters,

$$
\begin{aligned}
\boldsymbol{\beta} &\sim N_p(\mathbf{0}, 2I_p), \\
\phi &\sim \Gamma(2,2), \\
\sigma^2_{\boldsymbol{\omega}} &\sim N(0,2), \\
\tau^2 &\sim N(0,2),
\end{aligned}
$$

where $N_p(\cdot)$ denotes an p-dimensional multivariate normal distribution, and I_p denotes a $p \times p$ identity matrix.

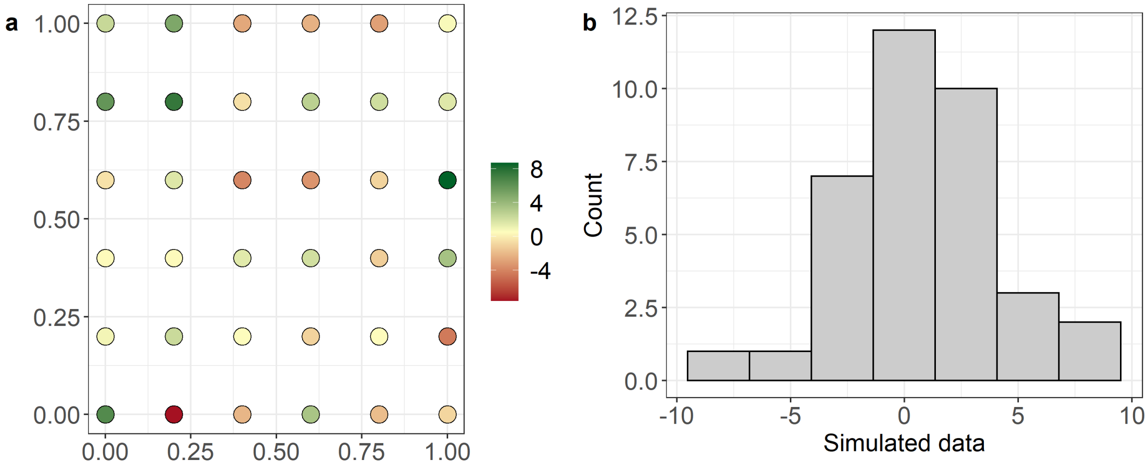

Figure E.1. Example of point referenced data simulated from (a) a lattice design and (b) a histogram depicting the simulated response.

Figure E.2 shows the results from four simulation experiments. In Figure E.2a, the dataset was generated with $n = 25,\ \sigma_{\boldsymbol{\omega}}^2 = 3,\ \tau^2 = 3,\ \phi = 3/0.5,\ \nu = 3/2$. In Figure E.2b, the dataset was generated using $n = 25,\ \sigma_{\boldsymbol{\omega}}^2 = 2,\ \tau^2 = 2,\ \phi = 3/0.5,\ \nu = 5/2$. In both cases, we first fitted the model with candidate parameters $\nu = 1/2,\ \nu = 3/2$ and $\nu = 5/2$ in the Stan framework, then we calculated the WAIC and PSIS-LOOIC using both the log likelihood extracted from Stan and the non-factorisable (NF) model log likelihood. In Figure E.2c, the dataset was generated using $n = 16,\ \sigma_{\boldsymbol{\omega}}^2 = 3,\ \tau^2 = 3,\ \phi = 3/0.5,\ \nu = 1/2$ and covariates combination f2. In Figure E.2d, the dataset was generated using $n = 16,\ \sigma_{\boldsymbol{\omega}}^2 = 2,\ \tau^2 = 2,\ \phi = 3/0.5,\ \nu = 5/2$ and covariates combination f1. In both cases, we first fitted the model using candidate covariates combinations f1, f2 and f3 in the Stan framework, then calculated the WAIC and PSIS-LOOIC using the log likelihoods extracted from Stan and the NF model log likelihoods.

Figures E.2a and E.2b show that the WAIC$_{\text{NF}}$ and PSIS-LOOIC$_{\text{NF}}$ outperforms the Stan WAIC and PSIS-LOOIC in correctly identifying the ν parameters used to generate the datasets. Figure E.2d shows that the Stan-computed WAIC and PSIS-LOOIC outperforms the WAIC$_{\text{NF}}$ and PSIS-LOOIC$_{\text{NF}}$ in correctly identifying the covariates combination used to generate the datasets. The results from Figure E.2c is unexpected, as we anticipated either marginal performance between the selection criteria calculated using the log likelihoods extracted from Stan and the NF model log likelihoods. Instead, we see that the WAIC$_{\text{NF}}$ and PSIS-LOOIC$_{\text{NF}}$ outperforms the Stan WAIC and PSIS-LOOIC in correctly identifying the covariates combination used to generate the datasets. Our explanation for this is due to the small n used in this experiment for data generation. Although we used different approaches to generate the data locations, the results shown in here are agreeable with what we have shown in Section 4.3.

Figure E.3 shows the results from four simulation experiments. In panel (a), the dataset was simulated using $n = 100,\ \sigma^2_{\boldsymbol{\omega}} = 2,\ \tau^2 = 2,\ \phi = 3/0.5,\ \nu = 1$; in panel (b), the dataset was simulated using $n = 100,\ \sigma^2_{\boldsymbol{\omega}} = 3,\ \tau^2 = 3,\ \phi = 3/0.5,\ \nu = 1$. In both cases, we first fitted the model with candidate parameters $\nu = 1/10,\ \nu = 1$ and $\nu = 1/2$ in the INLA method, then we calculated the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, as well as the WAIC using INLA. In panel (c), the dataset was simulated using $n = 100,\ \sigma^2_{\boldsymbol{\omega}} = 2,\ \tau^2 = 2,\ \phi = 3/0.5,\ \nu = 1$ and covariates combination f2; in panel (d), the dataset was simulated using $n = 100,\ \sigma^2_{\boldsymbol{\omega}} = 3,\ \tau^2 = 3,\ \phi = 3/0.5,\ \nu = 1$ and covariates combination f2. In both cases, we first fitted the model with candidate covariates combinations f1, f2 and f3 in the INLA method, then we calculated the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, as well as the WAIC using INLA.

Figures E.3a and E.3b show that our WAIC_{NF} outperforms the INLA-computed WAIC in correctly identifying the ν parameters we used to generate the datasets. Figures E.3c and E.3d show marginal performance between the WAIC_{NF} and the INLA-computed WAIC when tasked to correctly identify the covariates combination that we used to generate the datasets. Furthermore, both approaches failed to make the correct identification in majority of the generated datasets. However, these results are what we expect, and are agreeable with what was shown in Section 4.3.

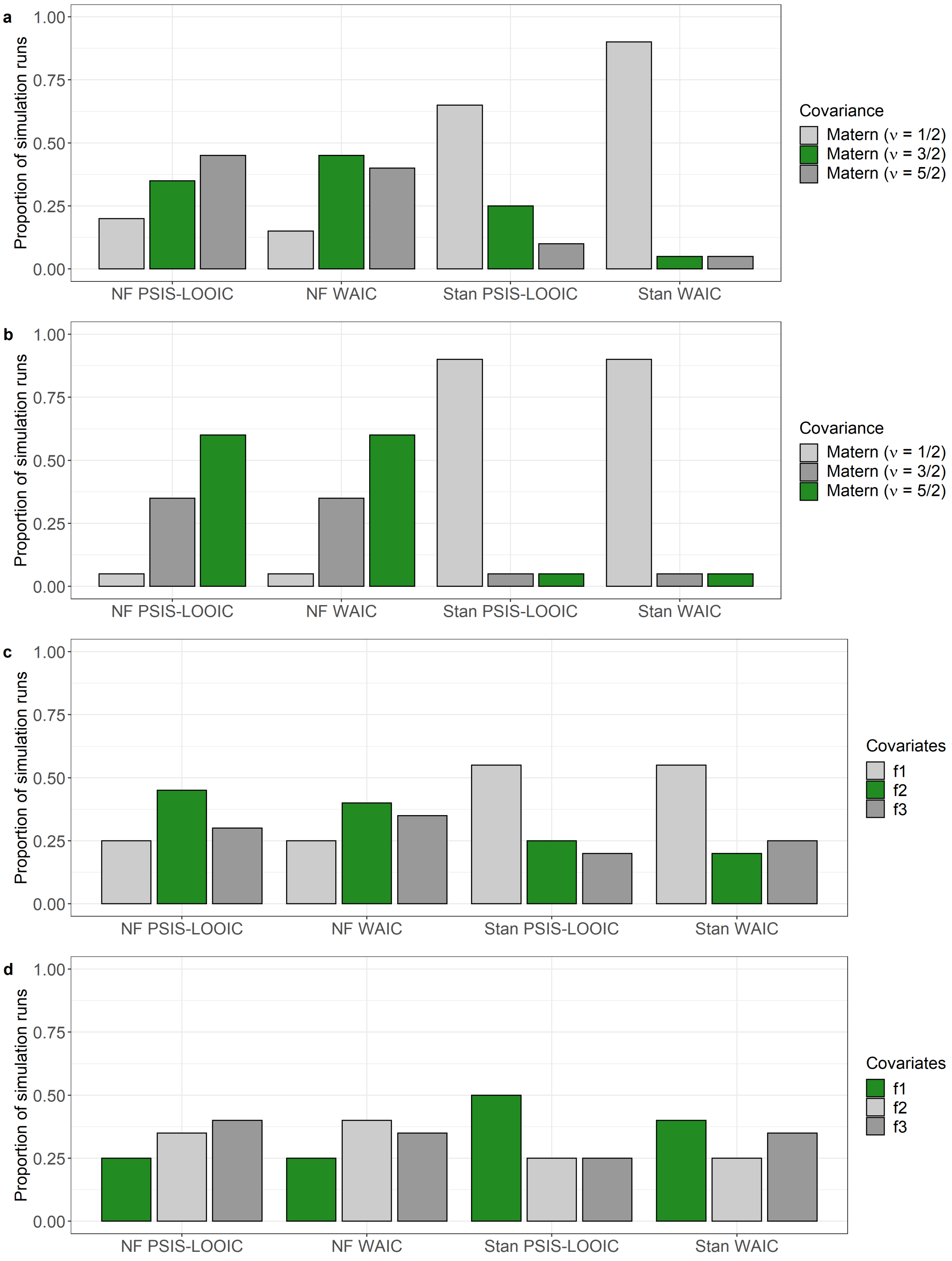

Figure E.2. The proportion of simulation runs where the selection criteria correctly identify the generating configuration from the lattice design. Green bars represent the generating configuration.

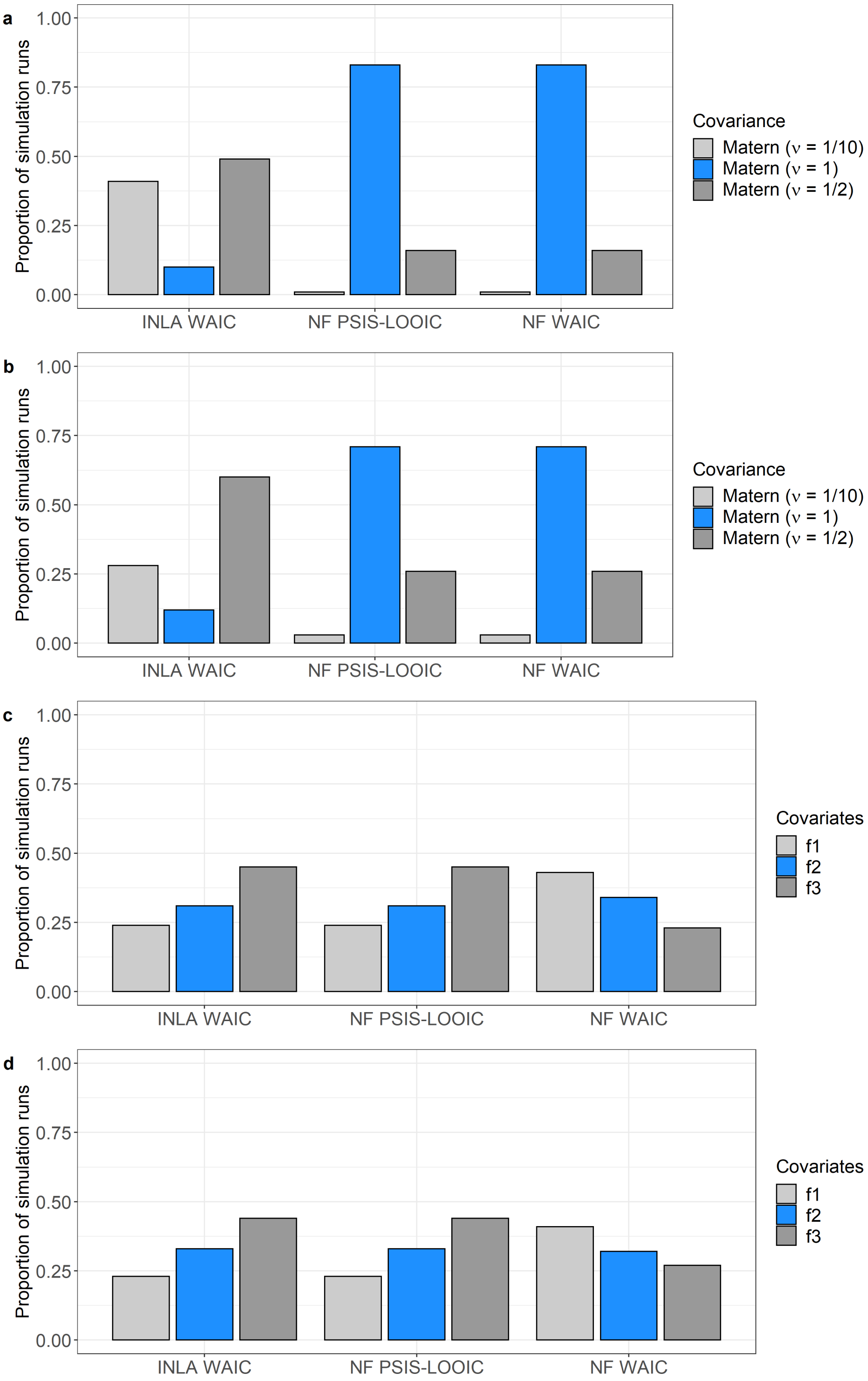

Figure E.3. The proportion of simulation runs where the selection criteria correctly identify the generating configuration from the lattice design. Blue bars represent the generating configuration.

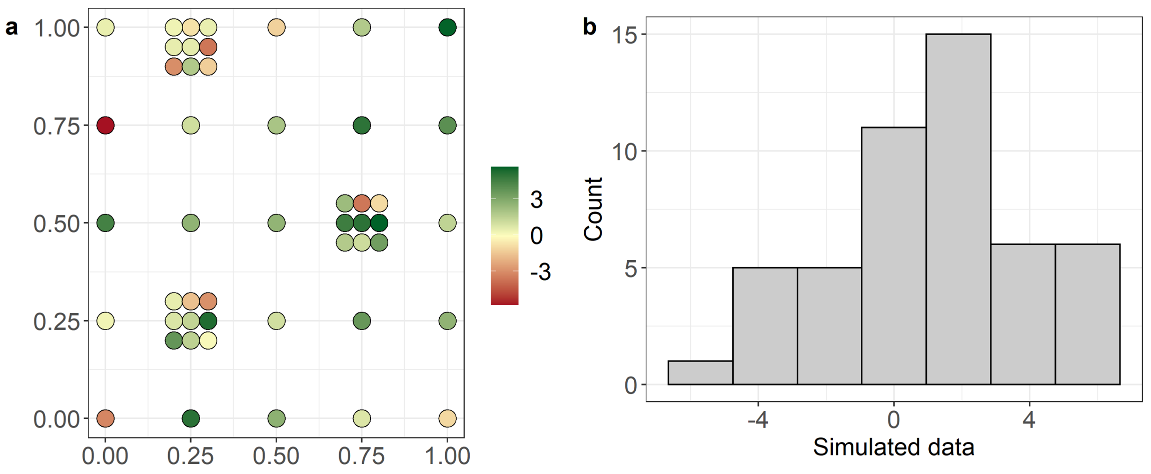

Figure E.4. Example of point referenced data generated from (a) an augmented lattice design and (b) a histogram depicting the simulated response.

Figure E.5 shows the results from four simulation experiments. In panel (a), the dataset was generated using $n = 49$, $\sigma^2_{\omega} = 3, \tau^2 = 3, \phi = 3/0.5$ and $\nu = 3/2$; in panel (b), the dataset was generated using $n = 49$, $\sigma^2_{\omega} = 2, \tau^2 = 2, \phi = 3/0.5$ and $\nu = 5/2$. In both cases, we first fitted the model with candidate parameters $\nu = 1/2$, $\nu = 3/2$ and $\nu = 5/2$ in the Stan framework, then we calculated the WAIC$_{\text{NF}}$ and PSIS-LOOIC$_{\text{NF}}$, as well as the WAIC and PSIS-LOOIC using the log likelihoods extracted from Stan. In panel (c), the dataset was generated using $n = 49$, $\sigma^2_{\omega} = 3$, $\tau^2 = 3$, $\phi = 3/0.5$, $\nu = 1/2$ and covariates combination f2; in panel (d), the dataset was generated using $n = 49$, $\sigma^2_{\omega} = 2$, $\tau^2 = 2$, $\phi = 3/0.5$, $\nu = 5/2$ and covariates combination f1.

Figures E.5a and E.5b show that our WAIC$_{\text{NF}}$ and PSIS-LOOIC$_{\text{NF}}$ outperforms the WAIC and PSIS-LOOIC computed using the log likelihoods extracted from Stan in correctly identifying the ν parameters that we used to generate the dataset. When tasked to identify the covariates combination used to generate the datasets, Figure E.5c shows that performance is marginal between the WAIC$_{\text{NF}}$ and PSIS-LOOIC$_{\text{NF}}$, and the WAIC and PSIS-LOOIC. Furthermore, Figure E.5d shows that the Stan WAIC and PSIS-LOOIC outperforms the WAIC$_{\text{NF}}$ and PSIS-LOOIC$_{\text{NF}}$. These results are what we expect, and are agreeable with what was shown in Section 4.3.

Figure E.6 shows the results from four simulation experiments. In panel (a), the dataset was generated using $n = 49$, $\sigma^2_{\omega} = 2$, $\tau^2 = 2$, $\phi = 3/0.5$ and $\nu = 1$; in panel (b), the dataset was generated using $n = 49$, $\sigma^2_{\omega} = 3$, $\tau^2 = 3$, $\phi = 3/0.5$ and $\nu = 1$. In both cases, we first fitted the model with candidate parameters $\nu = 1/10,\ \nu = 1$ and $\nu = 1/2$ in the INLA method, then we calculated the $\mathrm{WAIC_{NF}}$ and PSIS-$\mathrm{LOOIC_{NF}}$ and the the WAIC using the INLA method. In panel (c), the dataset was generated using $n = 49,\ \sigma^2_{\boldsymbol{\omega}} = 2,\ \tau^2 = 2,\ \phi = 3/0.5,\ \nu = 1$ and covariates combination f2; in panel (d), the dataset was generated using $n = 49,\ \sigma^2_{\boldsymbol{\omega}} = 3,\ \tau^2 = 3,\ \phi = 3/0.5,\ \nu = 1/2$ and covariates combination f2. Figures E.6a and E.6b show that our $\mathrm{WAIC_{NF}}$ outperforms the INLA-computed WAIC in correctly identifying the ν parameters we used to generate the datasets. However, the performance is marginal between the selection criteria when tasked to correctly identify the covariates combination we used to generate the datasets. In fact, the INLA-computed WAIC slightly outperforms the $\mathrm{WAIC_{NF}}$, as shown in Figures E.6c and E.6d. These results are what we expect and are agreeable with what was shown in Section 4.3.

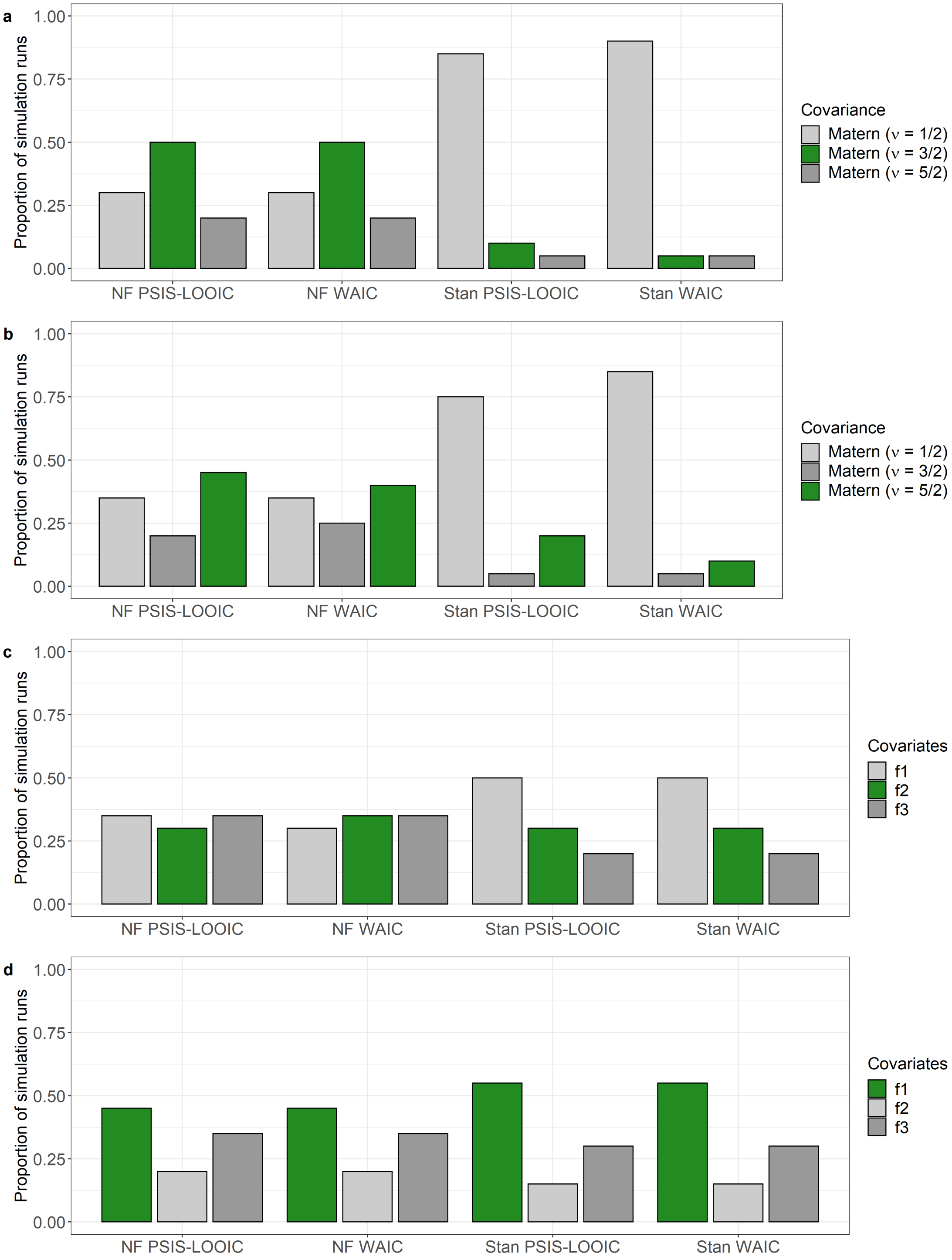

Figure E.5. The proportion of simulation runs where the selection criteria correctly identify the generating configuration from the augmented lattice design. Green bars represent the generating configuration.

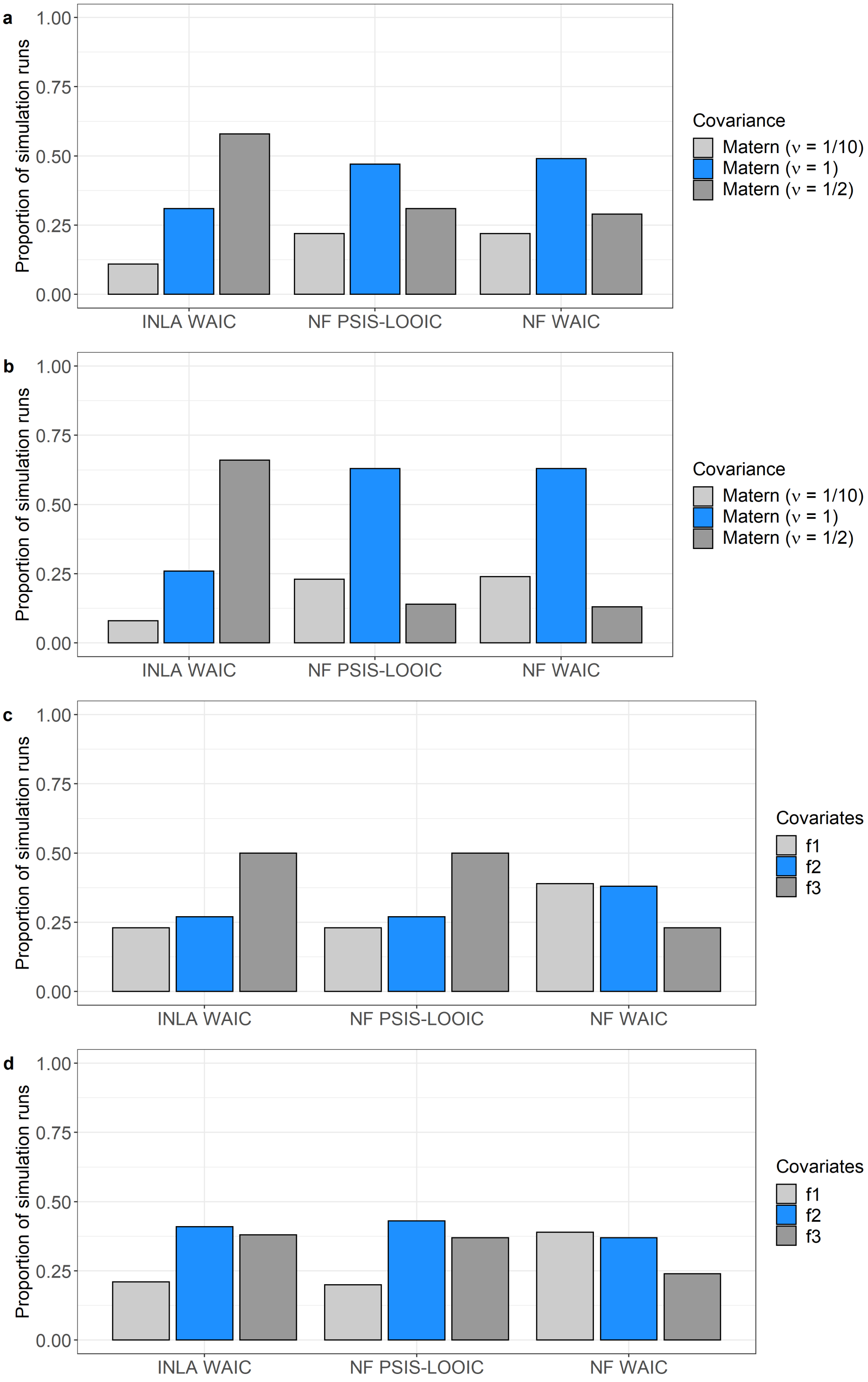

Figure E.6. The proportion of simulation runs where the selection criteria correctly identify the generating configuration from the augmented lattice design. Blue bars represent the generating configuration.

The results from these additional simulation experiments suggest that our proposed WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ are suitable for model selection tasks, particularly for determining the optimal covariance function for models of point referenced spatial data, but are not ideal for variable selection tasks.

Appendix F

Additional simulation designs 2

We conducted additional simulation experiments to explore the effects of the parameters on the WAIC and the PSIS-LOOIC. Specifically, we focus on the spatial decay parameter ϕ, the spatial variance $\sigma^2_{\boldsymbol{\omega}}$, and the iid variance τ^2. Recall from Chapter 4, the dataset generation configuration is given as follows,

$$\begin{aligned}\mathbf{Y} &\sim N_n(X\boldsymbol{\beta} + \boldsymbol{\omega}, \tau^2 I_n),\\ \boldsymbol{\omega} &\sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\omega}}),\end{aligned}$$

where $N_n(\cdot)$ denotes an n-dimensional multivariate normal distribution, I_n denotes an $n \times n$ identity matrix, $\boldsymbol{\beta}$ denotes a vector of regression coefficients, X denotes an $n \times p$ design matrix that contains the covariates and $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$ denotes an $n \times n$ covariance matrix, with elements calculated from the Matérn covariance function described in Section 2.3.3. We used the following covariates,

$$\mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} = x_1(\mathbf{s}_i)\beta_1 + x_2(\mathbf{s}_i)\beta_2,$$

where $\beta_1 = 1$ and $\beta_2 = 2$, and

$$\begin{aligned}x_1(\mathbf{s}_i) &= 1,\\ x_2(\mathbf{s}_i) &\sim N(0, 1).\end{aligned}$$

When fitting the model using the Stan framework and INLA method, we specified the following prior distributions for the parameters

$$\begin{aligned}
\beta_1 &\sim N(0,2),\\
\beta_2 &\sim N(0,2),\\
\phi &\sim \Gamma(2,2),\\
\sigma^2_{\boldsymbol{\omega}} &\sim N(0,2),\\
\tau^2 &\sim N(0,2).
\end{aligned}$$

In each of the following experiments, we first generated the dataset from the set up described above, then fitted the model using a range of fixed parameters and calculated the selection criteria, alongside their corresponding penalty components. Let us illustrate this through the following.

In the first experiment, we generated $n = 15$ points from a unit square and their responses following the set up above using the parameters $\sigma^2_{\boldsymbol{\omega}} = 3,\ \tau^2 = 3,\ \phi = 3/0.5,\ \nu = 1/2$. Using the simulated dataset, we fitted a model in the Stan framework using the prior distributions, as described above, for the parameters except for ϕ. For ϕ, we used the sequence generating function `seq()` in R to generate 10 values from 3/1.41 to 3/0.01, and fitted these values in the model. The choice of ϕ from 3/1.41 to 3/0.01 is because 1.41 is approximately the maximum distance between two points on a unit square and 0.01 is the minimum distance between two points on a unit square without being the exact same location. Finally, we calculated the WAIC, PSIS-LOOIC, WAIC_{NF}, $\text{PSIS-LOOIC}_{\text{NF}}$ of the model for each of these fitted values. We also calculated the corresponding penalty components. This was repeated five times; that is, we simulated five datasets from this set up. Figure F.1 shows the results from this experiment.

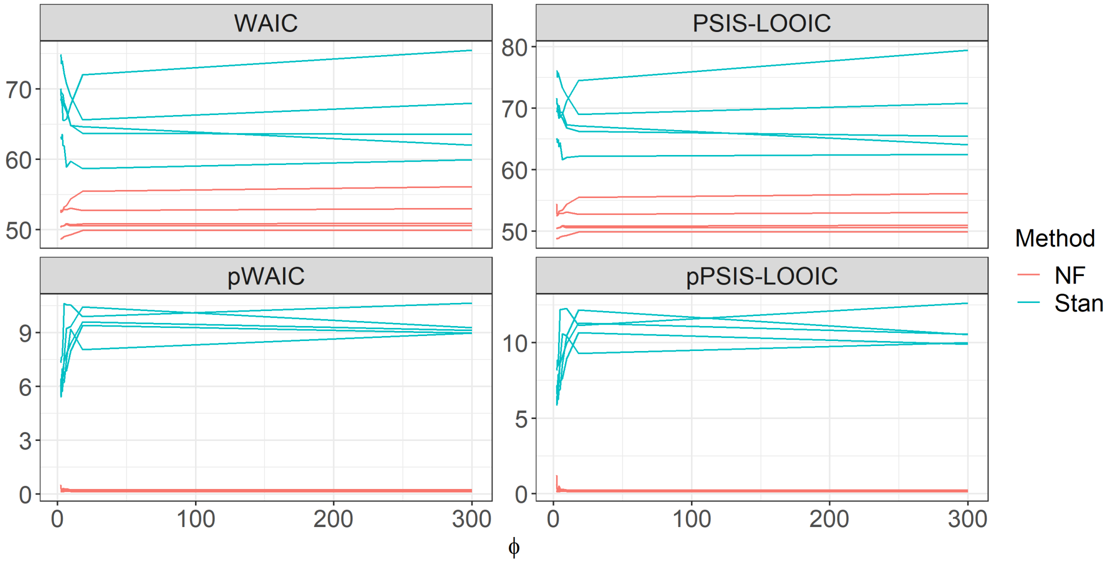

Figure F.1. WAIC and PSIS-LOOIC computed using non-factorisable model likelihood and Stan-extracted likelihood, for ϕ ranging from 3/0.01 to 3/1.41.

The experiment is also conducted for a range of σ^2_ω and τ^2 following the procedure described above. When we fit a range of σ^2_ω in the model, we have a prior on ϕ and τ^2. We investigated the 10 values from 1 to 5 from the `seq()` function for σ^2_ω. Likewise, when we fit a range of τ^2 in the model, we have a prior on ϕ and σ^2_ω, and we investigated the 10 values from 1 to 5 from the `seq()` function for τ^2. Figures F.2 and F.3 show the results of these experiments.

The procedure described above was repeated but with the models fitted with INLA. The results for these experiments are shown in Figures F.4, F.5 and F.6.

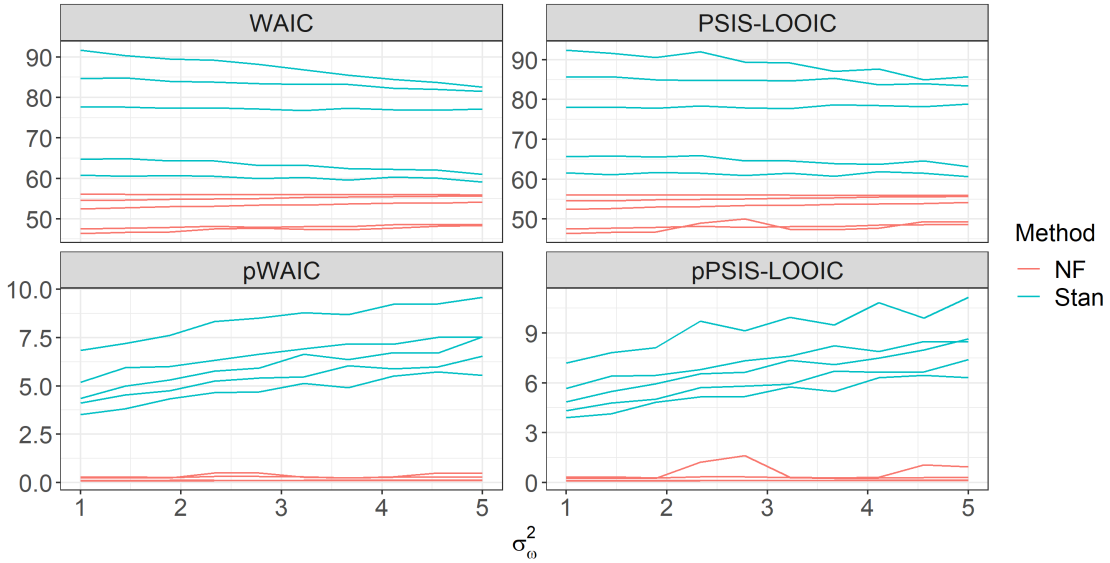

Figure F.2. WAIC and PSIS-LOOIC computed using non-factorisable model likelihood and Stan-extracted likelihood, for $\sigma_{\boldsymbol{\omega}}^2$ ranging from 1 to 5.

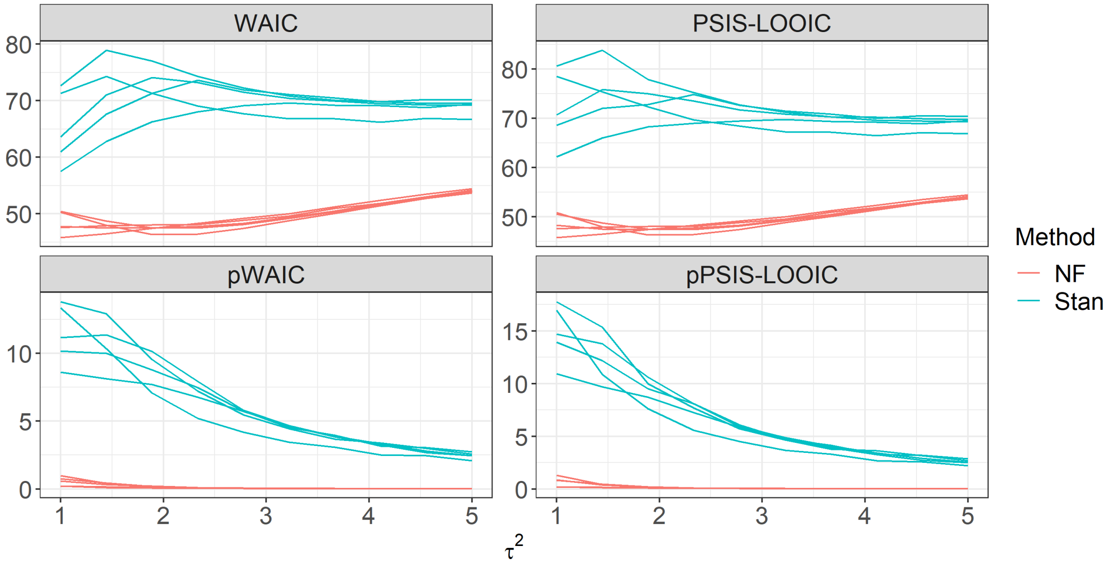

Figure F.3. WAIC and PSIS-LOOIC computed using non-factorisable model likelihood and Stan-extracted likelihood, for τ^2 ranging from 1 to 5.

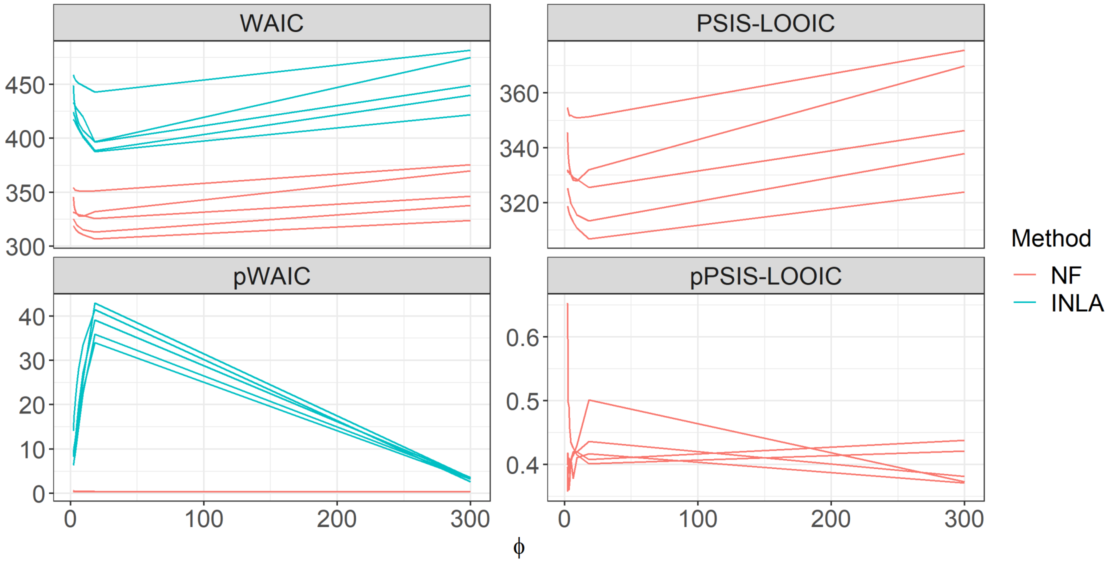

Figure F.4. WAIC and PSIS-LOOIC computed using non-factorisable model likelihood, along with INLA-derived WAIC, for ϕ ranging from 3/0.01 to 3/1.41.

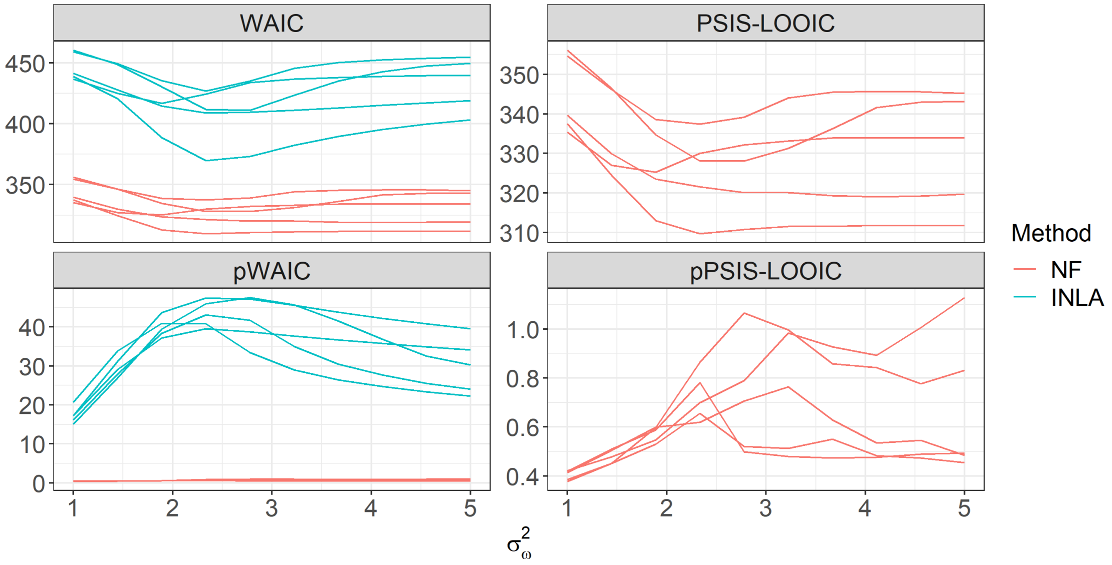

Figure F.5. WAIC and PSIS-LOOIC computed using non-factorisable model likelihood, along with INLA-derived WAIC, for σ^2_ω ranging from 1 to 5.

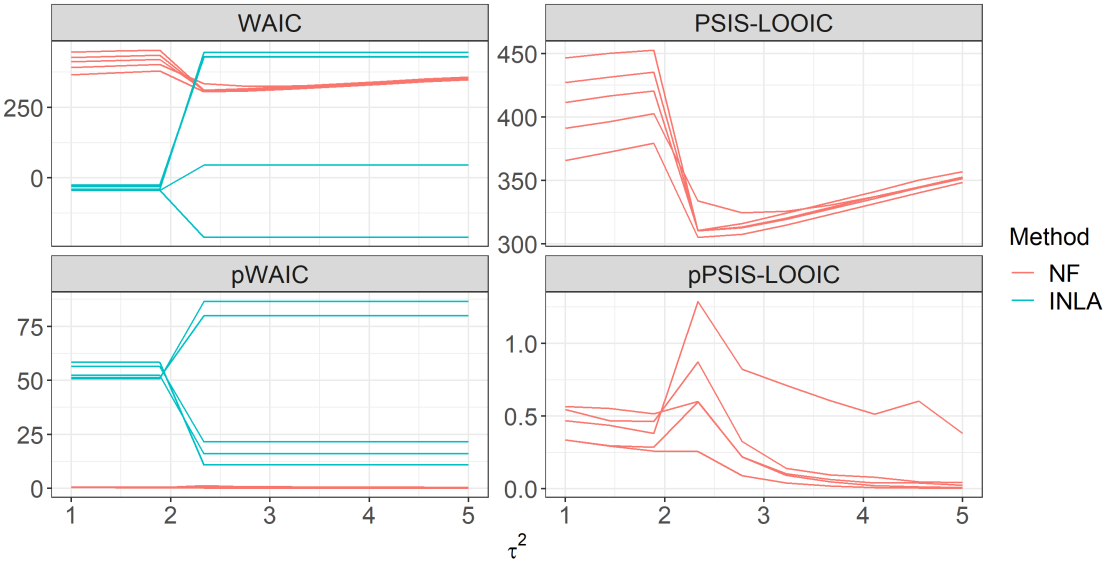

Figure F.6. WAIC and PSIS-LOOIC computed using non-factorisable model likelihood, along with INLA-derived WAIC, for τ^2 ranging from 1 to 5.

We see from Figure F.1 that the WAIC and PSIS-LOOIC values vary between the simulated datasets when ϕ is large. When ϕ is small, the WAIC and PSIS-LOOIC values for all the simulated datasets show a small initial change. The WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ remain unchanged as ϕ increases. For the penalty components, the p_{WAIC} and $p_{\text{PSIS}-\text{LOOIC}}$ values show an initial increase as ϕ increases whereas the penalty components for WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ remain unchanged as ϕ increases.

Figure F.2 shows that increasing the σ^2_{ω} parameter slightly decreases the WAIC and PSIS-LOOIC values but does change the WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ values. Increasing σ^2_{ω} also increase the calculated p_{WAIC} and $p_{\text{PSIS}-\text{LOOIC}}$ values, but does not change the penalty components of WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$. The WAIC and PSIS-LOOIC varies between simulated datasets although the simulation set up is the same.

Figure F.3 shows clear and consistent patterns in the WAIC, WAIC_{NF}, PSIS-LOOIC and $\text{PSIS-LOOIC}_{\text{NF}}$ values between the simulated datasets. Interestingly, opposite trends emerge as τ^2 increases. The calculated WAIC and PSIS-LOOIC values decrease as τ^2 increases, whereas the calculated WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$ values increase as τ^2 increases. While increasing τ^2 does not change the penalty component of WAIC_{NF} and $\text{PSIS-LOOIC}_{\text{NF}}$, it decreases the p_{WAIC} and $p_{\text{PSIS}-\text{LOOIC}}$

values.

For the results of the experiments where we fit the model in INLA, we see that increasing ϕ increases both the INLA WAIC and WAIC_{NF}. The results are consistent between the simulated datasets. Furthermore, we see that while the penalty component of WAIC_{NF} is unaffected by changes in ϕ, the INLA p_{WAIC} values show huge changes. Specifically, the INLA p_{WAIC} values initially increase, then steadily decrease as ϕ increases (Figure F.4).

We observe a similar pattern for both INLA WAIC values and WAIC_{NF} values as we increase $\sigma^2_{\boldsymbol{\omega}}$. The pattern exhibited is an initial decrease, then a gradual increase in the INLA WAIC and WAIC_{NF} values as $\sigma^2_{\boldsymbol{\omega}}$ increases. Increasing $\sigma^2_{\boldsymbol{\omega}}$ does not change the penalty component of the WAIC_{NF}, but causes the INLA p_{WAIC} to slightly increases until it flattens (Figure F.5).

Changes in τ^2 show unusual patterns in the INLA WAIC and WAIC_{NF} values. For the WAIC_{NF}, we see that the values slightly decrease when $\tau^2 \approx 2$ and flattens afterwards. On the other hand, there is are clear patterns with the INLA WAIC values. Instead, we observe that when $\tau^2 \approx 2$, the WAIC values either increase by a large amount or decrease by a large amount. We see the same pattern for the INLA p_{WAIC} as well while the penalty component of the WAIC_{NF} remains unchanged (Figure F.6).

Appendix G

Stan function block for latent NNGP

The subsequent code snippet represents the latent nearest-neighbour Gaussian process (NNGP) function encapsulated within a function block of a Stan file. This was developed by Zhang (2018) and is reliant on the contributions Finley et al. (2020). The latent NNGP serves as a pivotal component in the analytical in Chapter 6.3.

```
functions {

    real nngp_w_lpdf(
        vector w, real sigmasq, real phi, matrix NN_dist,
        matrix NN_distM, int[,] NN_ind, int N, int M) {

    vector[N] V;
    vector[N] I_Aw = w;
    int dim;
    int h;

    for (i in 2:N) {
        matrix[ i < (M + 1)? (i - 1) : M, i < (M + 1)? (i - 1): M]
        iNNdistM;
        matrix[ i < (M + 1)? (i - 1) : M, i < (M + 1)? (i - 1): M]
        iNNCholL;
        vector[ i < (M + 1)? (i - 1) : M] iNNcorr;
        vector[ i < (M + 1)? (i - 1) : M] v;
        row_vector[i < (M + 1)? (i - 1) : M] v2;
        dim = (i < (M + 1))? (i - 1) : M;

        if(dim == 1) {iNNdistM[1, 1] = 1;}
        else {
            h = 0;
            for (j in 1:(dim - 1)) {
                for (k in (j + 1):dim) {
                    h = h + 1;
                    iNNdistM[j, k] = exp(- phi * NN_distM[(i - 1), h]);
                    iNNdistM[k, j] = iNNdistM[j, k];
                }
            }
            for(j in 1:dim) {
                iNNdistM[j, j] = 1;
            }
        }

        iNNCholL = cholesky_decompose(iNNdistM);
        iNNcorr = to_vector(exp(- phi * NN_dist[(i - 1), 1:dim]));
        v = mdivide_left_tri_low(iNNCholL, iNNcorr);
        V[i] = 1 - dot_self(v);
        v2 = mdivide_right_tri_low(v', iNNCholL);
        I_Aw[i] = I_Aw[i] - v2 * w[NN_ind[(i - 1), 1:dim]];
      }

      V[1] = 1;
      return - 0.5 * ( 1 / sigmasq * dot_product(I_Aw, (I_Aw ./ V)) +
                      sum(log(V)) + N * log(sigmasq));
    }
}
```